Radicalization and ERG22 in Social Media

Yazid BOUNAB [1] and Mourad Oussalah [2]

[1]University of Oulu
[2]Affiliation not available

October 30, 2023

Abstract

social media became a fertile soil for various threats, extremism, and radicalization. This challenged policy-makers, researchers and practitioners. Preventing such extreme activities from happening becomes an ultimate priority at local and global scale. This paper introduces a new intertwine between radicalization and natural language processing capable of estimating the risk score of individuals based on their social media activities. The system uses a hybridized ERG22+ and VERA-ER model, which classifies individuals as high or low risk radicalization profile. The developed system was tested and validated on the Video Comments Threat Corpus dataset and Twitter pro-ISIS fanboys datasets where it achieves 95.1% and 64.9% accuracy, respectively.

Radicalization and ERG22 in Social Media

Yazid Bounab
Faculty of ITEE, CMVS
University of Oulu
Oulu, Finland
yazid.bounab@oulu.fi

Mourad Oussalah
Faculty of ITEE, CMVS
University of Oulu
Oulu, Finland
mourad.oussalah@oulu.fi

***Abstract*—social media became a fertile soil for various threats, extremism, and radicalization. This challenged policy-makers, researchers and practitioners. Preventing such extreme activities from happening becomes an ultimate priority at local and global scale. This paper introduces a new intertwine between radicalization and natural language processing capable of estimating the risk score of individuals based on their social media activities. The system uses a hybridized ERG22+ and VERA-ER model, which classifies individuals as high or low risk radicalization profile. The developed system was tested and validated on the Video Comments Threat Corpus dataset and Twitter pro-ISIS fanboys datasets where it achieves 95.1% and 64.9% accuracy, respectively.**

***Index Terms*—Radicalization, ERG22+, VERA-ER, risk perception, social media.**

1.. INTRODUCTION

The meteoric rise of social media activities together with the easiness, anonymization and popularity of open access social media platforms in the advent of Web 2.0 have substantially increased the size and scope of user generated content to reach astonishing level. Social media becomes a key forum where individuals can freely express their opinions, thoughts and establish their identities through posting, sharing, and liking [22]. This trend has unfortunately been accompanied by a malicious use of these open platforms to gain support of extremism groups and agenda where malignant activities like hate propaganda, brainwashing and fundraising were promoted. Malicious communication takes place through various mediums and forms, e.g., live-stream video, image, audio, online games, chatroom, textual description, links, likes / emojis, among others. Often, a given social media post may include a mixture of these forms, which can offer a tailored virtual space that accommodates individual desires, tendencies, emotions and illicit intentions. Indeed, malicious users exploit social media platforms to communicate or diffuse their messages and recruits in many parts of the world [47]. Furthermore, with the increased amount of radicalization content and extremism, social media platforms have become a fertilized ground for terrorists and self-radicalized individuals. Therefore, the need for building a risk assessment tools that detect individuals with extremist beliefs is of paramount importance to prevent and anticipate the occurrence of any potential harmful event [62]. A such tool, if any, should employ users' online posts and activities to predict radicalization risk [35], which provides useful inputs to national security intelligence services to act prior occurrence of harmful events and incidents caused by radicalized individuals. Research into online radicalization detection becomes very sparse and multidisciplinary where law enforcement agencies, social researchers, computer scientists and volunteers are actively working to tackle this problem, [3]. For instance, the voluntary organization Ctr-sec [1] claimed that volunteers report on ISIS propaganda in social media enabled them to close more than 200,000 Twitter accounts belonging to suspected individuals/organizations. Furthermore, the unstructured and informal nature of content with the increased use of abbreviations, colloquialism, and transliterations yield an extra layer of difficulties to the problem. Various projects such Dark Web project [56] funded by National Science Foundation of USA, Princip project of Safer Internet Plan in EU, together with various national, industrial efforts emerged in the last two decades for the purpose of achieving Safe Internet. Nevertheless, the challenges are still far to be overcome due to the dynamic nature of web, the complexity of regulation based solutions and the inherent limitations of algorithmic solutions promoted by research communities, which call for further research on the issue. Indeed, existing methods to automatically identify radical content online mainly rely on the use of glossaries such as aggregating lists of terms associated with religion, threat, offensive language, among others. The effectiveness of such an approach is often questioned because, e.g., the occurrence of hate speech terms makes no distinction between users who promote hate speech and those who combat it; the association of these terms with radicalization is very much context dependent and would require complex subsequent discourse analysis for disambiguation; the harmonization of the boundary of radicalization definition and its various ramifications is often open to debate, especially in the case of online radicalization; the scope and scale of the ground truth data employed in the testing and the evaluation tasks of the developed approaches is another striking limitation to the development of this field. Looked from another perspective, the above challenges can be cast into the difficulty of translating user's textual content into a reliable risk index associated with extremism / radicalization [28]. Strictly speaking, psychologists, sociologists and criminal justice lawyers developed numerous risk assessment frameworks that are used to evaluate a set of risk factors,

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

which enable us to predict whether an individual is likely to be radicalized or not. Several tools have been developed for the purpose of assessing whether an individual will engage in violent extremism or not. These instruments are implemented either in pre-trial, detention, or post-detention settings [55]. Typical models include RADAR, VAF (Vulnerability Assessment Framework), SQAT (Significance Quest Assessment Test), RRAP (Radicalization Risk Assessment in Prison), ERG22+ [46]. For each model, risk factors are associated with attributes such as belief, support to radicalized organizations, number of radicalized activities involved, among others, which provide a basis for the likelihood estimation. This opens up new horizon to study online radicalization from such well-established risk assessment instruments. This paper focuses on the study of online radicalization using ERG22 and VERA-ER risk assessment models. Nevertheless, there is a structural difference in the sense that ERG22-VERA-ER risk assessments are primarily designed for prisoners where officers can observe their behavior and interrogate them whenever needed. Therefore, the extension of this scheme to a virtual environment of blogosphere, despite being scientifically appealing, also bears inherent limitations due the absence of physical interactions and the complexity of natural language processing tasks involved. Although, the impact and interest of such analysis are well acknowledged given the role played by online radicalization into violent extremism and societal fragmentation as indicated by recent news stories. For example, police investigation revealed that individuals behind Paris 2015 bombing have been driven by motives gained through their online activities where interaction with radicalized groups was identified [18]. Although, the debate about the reasons behind terrorist attacks is widely open, where the leading causes are, in overall, rooted back to political, religious, and psychological motives [1], the impact of online behavior is well accredited by counter terrorism experts [21]. Therefore, any automated approach that would help law enforcement agencies to gain insights in terms of radicalization likelihood would provide a basis for subsequent monitoring tasks and planning. In overall, this research employs unstructured textual data from social media (posts, comments, and replies) to estimate the radicalization risk of an individuals by mapping the posts to ERG22-VERA-ER categorization and assessment framework. This research has three-fold objectives:

1) O_1: To transform ERG22- VERA-ER into an ontology that can be queried using natural language processing tools.
2) O_2: To build a monitoring system that assesses the radicalization risk using a hybrid ERG22-VERRA model.
3) O_3: To validate the model using two datasets: Video Comments Threat Corpus and Twitter Pro-ISIS Fanboys.

To achieve the above research objectives (O_1, O_2, and O_3), First, we performed devised a multi-step processing pipeline that includes building a hybrid ontology from ERG22+ and VERA-ER taxonomies, data preprocessing and feature extraction, radicalization risk score estimation and evaluation.

Contributions

This paper advocates essentially four-fold contributions.

- A comprehensive review of existing models of extremism/radicalization estimation is provided in the background section of this paper.
- A novel hybrid framework that uses ERG22+ and VERA-ER models is put forward, contributing towards objective O_2.
- A novel model that enables an estimation of individual's radicalization/extremism score according to the textual content of his post (s) is devised and implemented (contributing to O_2). The model evaluates the content of a user's post content with respect to ERG22+-VERA-ER ontology to distinguish high/low profile according to the estimate risk score.
- For testing and validation purpose, a novel annotation technique for labeling twitter Pro-ISIS fanboys dataset is devised and implemented, contributing to O_3.

Section II of this paper presents the background of the different risk assessment tools. Section III describes the datasets employed in this study. Section IV details the method and the data pipeline used to answer the aforementioned research questions. Results and discussions are reported in Section V.Finally, conclusive statements and perspective works are stated in Section VI.

II. Radicalization risk assessment tools

Assessment of radicalization risk differs according to the risk perception and attributes judged more important for the assessment task. For instance, some tools consider that the risk may refer to the chance of socializing with extremist networks, while others focus on risk of using violence in future acts or performing terrorist acts [59]. In overall, four methods can be distinguished in individual (radicalization) risk assessment tasks: unstructured clinical judgment, actuarial methods, structured professional judgment (SPJ), and self-assessment methods. In SPJ methods, decisions are based on guidelines, structured questions or a set of indicators issued from empirical evidence or professional practice. Such approach has gained an edge with practitioner community due to its demonstrated reliability and validity. In this respect, Lloyd [40] reviewed six commonly used tools for anti-terrorism risk assessment, which are summarized below.

1) Islamic Radicalisation (IR-46): IR-46 is an SPJ tool created in 2016 by the Dutch Police department in the Netherlands as a successor to the Kennis in Modellen (KIM) tool [64]. It delivers a framework for analyzing an individual's risk of violent extremism across two domains: social context and ideological factors. It includes 46 indicators to assess individuals involved in terrorist acts and violence driven by religion and/or social ideologies. However, the IR-46 is unsuitable for other ideological groups since it is originally designed to be used to assess Islamic radicalization only [40].

2) Multi-Level Guidelines (MLG): MLG is an SJP tool developed in 2013 by Cook, Hart and Kropp [13], widely used in North America and Europe. The tool's main target is the assessment of group-based violence (GBV), particularly with respect to terrorist activities [65]. GBV targets a set of threats, attempts, or actual violent activities which cause injuries, committed by either a single individual or a group, often brainwashed by their belonging mentor (s) [12]. MLG includes 20 systematic review-based risk factors across four domains: individual factors, individual-group, group, and group-societal factors [65]. Especially, MLG is used for reassessment purpose, to monitor any change due to the dynamic nature within on year time period [65]. MLG utilizes the entire SPJ strategy via scenario planning emphasizing an individual in his social and broader societal and political context, which provides an edge when dealing with gangs, terrorists, and those involved in organised crime [40]. Nevertheless, practitioners must be skilled risk assessors to analyse the flow of information adequately because the elements in the individual domain are generic [40] and lack the specificity required to perform a full terrorism assessment.

3) Extremism Risk Guide (ERG 22+): The ERG22+ is an SPJ tool created by the United Kingdom's Prison and Probation Service (UKPPS) in 2011 based on the literature on terrorists, casework of individuals convicted of terrorism offences, and a comparative analysis of the criminogenic profiles of individuals convicted of extremist offences. ERG22+ provides a guided framework for risk assessment according to threat severity as compiled by the National Offender Management Service (NOMS) [41]. This tool provides a way to determine an individual's risk level of involvement with an extremist group, share its cause or ideology as well as the individual's willingness to offend (UKPPS, 2019). Therefore, the ERG22+ is used not only on people convicted of extremist offences in England and Wales but also on individuals with no previous convictions (UKPPS, 2019). It includes three categories (engagement, intent, and capacity) with 22 risk indicators. The users of the ERG22+ are generally registered psychologists or experienced probation Officers. Despite its popularity in UK and elsewhere, the information on reliability and validity is still to be demonstrated, and it remains to be established whether the factors of the ERG22+ are correlates or predictors of risk [40].

4) Violent Extremism Risk Assessment-2 (VERA-2): VERA-2R is another SPJ tool created by the Netherlands Institute for Forensic Psychiatry and Psychology (Pressman *et al.* 2019) [53], [54] developed by academia and mental health experts. VERA-2R provides a framework for analyzing individual's risk of violent extremism across eight domains: Beliefs & Attitude, Social Context, History & Capacity, Motivators, Risk mitigating indicators, Personal history, Criminal history, and Psychopathology. VERA-2R holds 45 indicators used to assess individuals involved in violent extremism, terrorism, violence driven by religious, political or social ideologies. This, in principle, makes VERA-2R suitable for all types of extremism regardless the age and gender [40]. In addition, VERA-2R can inform about assessment, risk management, and decision-making through pre-crime or post-crime across any judicial setting. In addition, due to the emphasis on feeling alienated and needing social support, [7], hypothesized that the VERA would be simpler to apply to people who work in groups. However, VERA suffers from the small sample size that makes it not easy to generalize beyond Netherlands case study [40].

5) Terrorist Radicalization Assessment Protocol (TRAP-18): TRAP-18 is another SJP tool developed in 2018 by Meloy [42] as an investigative template. The tool assists in prioritising cases depending on the severity of the danger to overcome the challenges faced in counter-terrorism [44]. The tool focuses on preventing lone terrorist behaviour instead of predicting it. TRAP-18 targets individuals who attracted the attention of law enforcement due to concerns regarding engagement in an ideologically motivated violence. TRAP-18 includes two sets of indicators: 8 warning behaviours and 10 distal features. The warning behaviours were designed as a way to detect the relative risk of targeted or intentional violence [57]. The warning signs might suggest an increased danger of targeted violence [45]. Several distal traits, such as a history of criminal violence, remain static despite being drawn from the psychological study of lone-actor terrorism. The distal features and proximal warning behaviours can also be distinguished accordingly [42]. Although TRAP-18 can distinguish between empty threats and actual dangers [40], this tool focuses on lone-actors limiting its pertinence with group actors and the challenges of assessing the information needed to complete the assessment in a pre-crime scenario.

6) Vulnerability Assessment Framework (VAF): Developed by UK government, VAF consists of 22 factors -across three dimensions: engagement, intent and capability– that may cause an individual to (a) engage with a terrorist group; (b) develop the intent to cause harm, and; (c) develop the capability to cause harm. It is primarily used to assess whether individuals need support to safeguard them from the risk of being targeted by terrorists and radicalizers [2].

7) Non SPJ models: In addition to the aforementioned SPJ models, we shall also mention the existence of a set of non-SPJ models, which are less popular with practitioners. This includes the following, see [38] for details:

- Identifying Vulnerable People (IVP). The Guidance for IVP model [19] rather describes some risk behaviour but does not provide any risk assessment like-approach. Therefore it does not fit with the current purpose of study.
- Significance Quest Assessment Test (SQAT). SQAT model [38] is developed to measure detainee's degree of radicalization using a 66 item questionnaire over three categorization: 'needs'; 'narrative'; and 'network' (the 3N-approach).
- RADAR is a protocol designed to identify individuals that could benefit from early interventions, focusing on observable behavioural indicators (social context, ideology and criminal action orientation) and their potential for coping. So the tool rather acts as an aid to decision-making process for policy officers and municipalities.

[2]"Channel Vulnerability Assessment," HM Government, 2012, https://www.gov.uk/government/publications/channel-vulnerability-assessment.

Table I summarizes the key characteristics and our appreciation on the pros and cons of each method. Especially, our review of radicalization tools revealed the following. First, from a methodological perspective, the SPJ class of methods has an edge over other methods, due to the presence of clearly identified indicators and risk factors, which explain the high interest of research community. Second, some tools (e.g., IR-46, up to some extent ERG22 -while ERG22+ is meant to be applied to all extremism ideologies) are specifically tailored to one ideology (Islamic ideology for IR-46), which restricts their application to other ideologies. Third, there is an inherent difference when looking at radicalization event as an individual act or organization act. Similarly, the methods differ according to the level of expertise required by the officers who apply the protocol on the individuals. Fourth, among the SPJ methods, ERG22 and VERA-ER are by far the most popular with practitioner and scientific community due to their well structured risk indicators and boost from UK and USA jurisdiction organizations. Fifth, another critical issue, which is often not elucidated in the risk documents, concerns the aggregation of the various risk indicators. In this regards, very often, the experts conducting the interview /protocol are responsible for deciding on the way and type of such an aggregation.

III. Methodology

A. Background

The starting point in our methodology is to acknowledge the risk factor / indicators developed in ERG22+ and VERA-ER as key pillars in the development of an online risk assessment score. For this purpose, we hypothesize that

- H_1: the textual description of these indicators can be translated into a simple ontology used for text matching and retrieval task;
- H_2: the extent of textual matching can be used as a risk assessment pertaining to the corresponding indicator;
- H_3: the use of the state-of-the-art BERT model or external lexical database would enable us to account for various context in the text matching quantification task;
- H_4: In line with some expert-based aggregation of the various risk indicator employed in SPJ risk aggregation [40], we assume no preference among the risk indicators, and therefore a max combination rule will be used to aggregate the risk scores of the various indicators.
- H_5: Individuals can be classified into either high risk profile or low risk profile in terms of radicalization risk.

For H_1, it should be noted that since ERG22+ and VERA-2G were developed based on empirical research and interviews with terrorist offenders, this makes them an ideal starting point to identify online radicalization [38]. Indeed, both VERA-ER and ERG22+ have proven to be well suitable for identifying high risk individuals, not only for those who have already committed crimes, but also for suspected individuals. We therefore adopted a hybrid ERG22-VERA-ER solution by combining their associated factors, although many features are found to be overlapping. This hybridization also enables us to compensate for inherent limitations due to the lack of exemplification in the definition of some factors. Whereas H_2-H_4 provide a basis for quantifying individual risk score according to the extent of matching of user's input to Indicator's definitions. Especially, the use of BERT model enables us to represent textual description of both indicator textual description and user's textual input as numerical vectors, so that the matching can be evaluated using standard metrics like cosine similarity measure. Likewise, the use of the external lexical databases, e.g., WordNet, permits data augmentation of initial data that enable the system to go beyond standard string matching process in accounting for semantic aspect. H_5 attempts to accommodate the nature of the dataset employed in our study where both Youtube dataset and, up to some extent, ISIS dataset, provide insight to distinguish high risk profile and low risk profile. Therefore, risk evaluation score should be converted into a binary classification (low and high risk) problem to fit this purpose. On the other hand, since ISIS dataset lacks ground truth, a novel approach has been devised to use Youtube dataset as a guiding tool to annotate the dataset. Figure 1 provides a generic pipeline describing the overall architecture with different steps for building our risk assessment tool whose individual components are detailed in the next subsections.

B. Hybrid ERG22+ -VERA-ER ontology

The construction of the hybrid model involves merging the different factors definitions of both risk assessment tools in ERG22+ and VERA-ER. This step consists of building a set of vocabulary associated with each factor of the hybrid tool by extracting and normalizing the relevant tokens contained in the definition statements. Table II presents the factors' definitions used for building the hybrid model ERG22+ and VERA-ER. We then create an expanded keyword list linked with each factor definition statement(s), say i^{th} factor Hf_i. For this purpose, we utilize a three-stage process. First, we extract words associated with each ontology from the hybrid factor definitions $Hf = \{Hf_1, Hf_2, ...Hf_{23}\}$, followed by vocabulary augmentation using the lexical database WordNet for synonymy relation extraction. Finally, a refinement using an old-fashioned manual checking stage is performed for possible inconsistency detection.

C. DataSets

This paper uses two datasets involving violence and threats to test our online risk radicalization model.

Video Comments Threat Corpus (VCTC): This dataset was collected in 2013 from 19 different YouTube videos related to various topics (religious beliefs and political conflicts) that trigger anger and hatred emotions. The dataset consists of 9.845 comments with 28.643 sentences written by 5.484 different users. Its annotation uses a binary format indicating whether it corresponds to a threat or not. In total, 993 users wrote 1.287 comments where 1.387 sentences annotated as violent threats. In addition, some of the content of the

TABLE I
REVIEW OF EXISTING RISK RADICALIZATION TOOLS

Tool	Summary	Category Names	Aadventages	Disadventages
Extremist Risk Guidance (ERG22+) (M. Lloyd & C. Dean (NOMS))	SPJ tool developed in the UK, It has 22 Factors. Targeting extremist prisoners in England and Wales	Engagement, Intent & Capability	Provide sentence planning, intervention and release planning. Developed by international experts and advisory group.	Unknown Reliability and validity. Developed on Al-Qaeda extremists. No consideration of other factors. Questionable when apply to different types of extremism and different populations.
Violent Extremism Risk Assessment (VERA-2R) D.E. Pressman, N. Duits, T. Rinne & J. Flockton	SPJ tool developed in Canada /USA.It has 45 Factors, Targeting all types of violent extremists, offenders, and terrorists driven by religious, political, or social ideologies. Pre/Post crime usage	Beliefs & Attitude; Social Context; History & Capacity; Motivators; Risk mitigating indicators. Personal history; Criminal history; Psychopathology	Revised version Flexibility to add new factors. Applicability to all ideological types.	No access for assessors to classified information. Long time in rating quantitative and qualitative information
Terrorist Radicalization Assessment Protocol (TRAP-18) J. Reid Meloy	SPJ tool developed in Netherlands. It has 18 Factors. targeting lone-actor intended to commit terrorism driven by ideologically	Proximal warning behaviours distal characteristics	Pre-crime screening and informs if monitoring is needed. Several studies proved the utility of the framework.	Limited to individual assessment. Lack of information in pre-crime scenarios.
Multi-Level Guidelines (MLG) A. Cook, S.D. Hart & P.R. Kropp	SPJ tool developed in Canada, It has 20 Factors. Can be used pre/post crime with member of a group	Individual risk factors, individual group factors, group factors group societal factors	Usability with terrorists and organised crime.	The individual domain lacks detail in assessing violence as a backgrounds key of individuals involved in terrorism.
Islamic Radicalization (IR-46) Dutch Police Forceq	SPJ tool developed in Netherlands. It has 46 Factors To be used pre-crime with individuals displaying signs of Islamic radicalization	Social context & ideological factors	Easy to use, Ability of structuring the management of risk. Widly used by police	Limited to Islamist extremism. No individual assessment.
Structured Assessment of Violent Extremism (SAVE) G. Dean & G. Pettet	Self-report tool developed in Australia. It has 30 Factors. To be used for pre-crime	Cognitive risk factors. Terrorism, militant, shooter.	Ability to capture the subjectivity in decision making.	Little research on SAVE
Vulnerability Assessment Framework (VAF) NOMS/Channel Program	Self-report tool developed in the UK. It has 22 Factors. Targeting Individuals considered vulnerable to radicalization	Engagement, Intent & Capability	Flexibility of usage on individuals work in education, local authorities, youth services and the health sector.	Little research on the VAF
Identifying Vulnerable People (IVP) J. Cole, B. Cole, L. Allison & E. Allison	SPJ tool developed in the UK. It has 16 Factors. Targeting Individuals considered vulnerable to radicalization	Generic risk indicators & Red flag indicators	Accessible online. Easy to administer. No required training or licensing. Ability to structure concerns and inform post-assessment actions.	Inspired by AL Qaeda extremism. No protective factors or risk management.
RADAR K. Barelle & S. Harris-Hogan	SPJ tool developed in Australia. It has 27 Factors. Targeting radicalized individuals in/out of prison	Social Relations, Coping, Identity, Ideology & Action Orientation	RADAR can be used in and outside the prison context	Little research on RADAR
Significant Quest Assessment Test (SQAT) A.W. Kruglanski	Self-report tool developed in The USA. It has 66 Factors. Targeting radicalized prisoners	Needs, Narrative & Network	As it is completed by the individual, there is no need to obtain information.	Individuals may provide socially desirable answers
Radicalization Risk Assessment (RRAP) P. das Neves	Self-report tool developed in Protugal. It has 39 Factors. Targeting prisoners thought to be vulnerable or in process of radicalization	Emotional uncertainty, self-esteem, radicalism, distance, and societal disconnection, need to belong, legitimization of terrorism, perceived in group superiority, identity fusion, and identification, and activism	Designed specifically for use in prisons and probation settings	Little research on the RRAP

comments were quoted as originated from either the Quoran or the Bible [25].

Twitter Pro-ISIS fanboys: This contains Twitter discussion around the November 2015 Paris attacks where over

TABLE II
MAPPING BETWEEN ERG22+ AND VERA-ER

Hybrid Factors (ERG22)	Definitions (VERA-ER)	Hybrid Factors (ERG22)	Definitions (VERA-ER)
Need to redress grievance	Victim of justice Rejection of democratic values Hostility to collective national identity Feelings of hate, frustration, persecution and alienation anger	Evaluated psychopathology	Evaluated psychopathology
Need to defend against threat	Feelings of hate and persecution	Over-identification	Over-identification
Need for identity, meaning & belonging, and comradeship	Need for identity Driven by comradeship, group belonging, status in the group	Us and them thinking	Us and them thinking Hostility to national collective identity/identity conflict
Need for significance & status	Need for significance and status Driven by status in group, acquisition of status Search for significance, meaning in life	Dehumanisation of the enemy	Dehumanisation of the enemy Dehumanisation/demonisation of target group
Desire for excitement & adventure	Desire for excitement & adventure Driven by excitement & adventure	Attitudes that justify offending	Attitudes that justify offending Commitment to ideology justifying violence Glorification of violent action
Need to dominate others	Need to dominate others	Harmful means to an end	Harmful means to an end Willingness to die for cause
Susceptibility to indoctrination	Susceptibility to influence and indoctrination	Harmful end objectives	Harmful end objectives Expressed intent to plan violent action Expressed intent to act violently & to plan & prepare action Identification of a target Lack of empathy for outgroups Seeker/consumer/developer violent materials
Political, moral motivation	Political, moral motivation Driven by moral imperative and superiority by religion or noble cause	Individual knowledge, skills & competencies	Individual knowledge, Skills and competencies Tactical paramilitary explosives training
Opportunistic involvement	Opportunistic involvement Criminal Opportunism	Access to networks, funding & equipment	Access to networks, Funding & equipment Personal contact with extremists Funds, resources & organisational skills
Family/friends support extremism	Family/friends support extremism Network (family/friends) involved in violent action	Criminal history	Criminal history Prior criminal history of violence Personal history: early exposure to violent extremism and ideology
Transitional periods	Transitional periods	Other factor	lack of resilience Relational problems Lack of healthy father role model Desire to be a hero Hedonistic guilt Employment problems Previous trauma Failure to meet cultural or family expectations
Group influence and control	Group influence and control Forced, coerced to participate susceptible to influence		

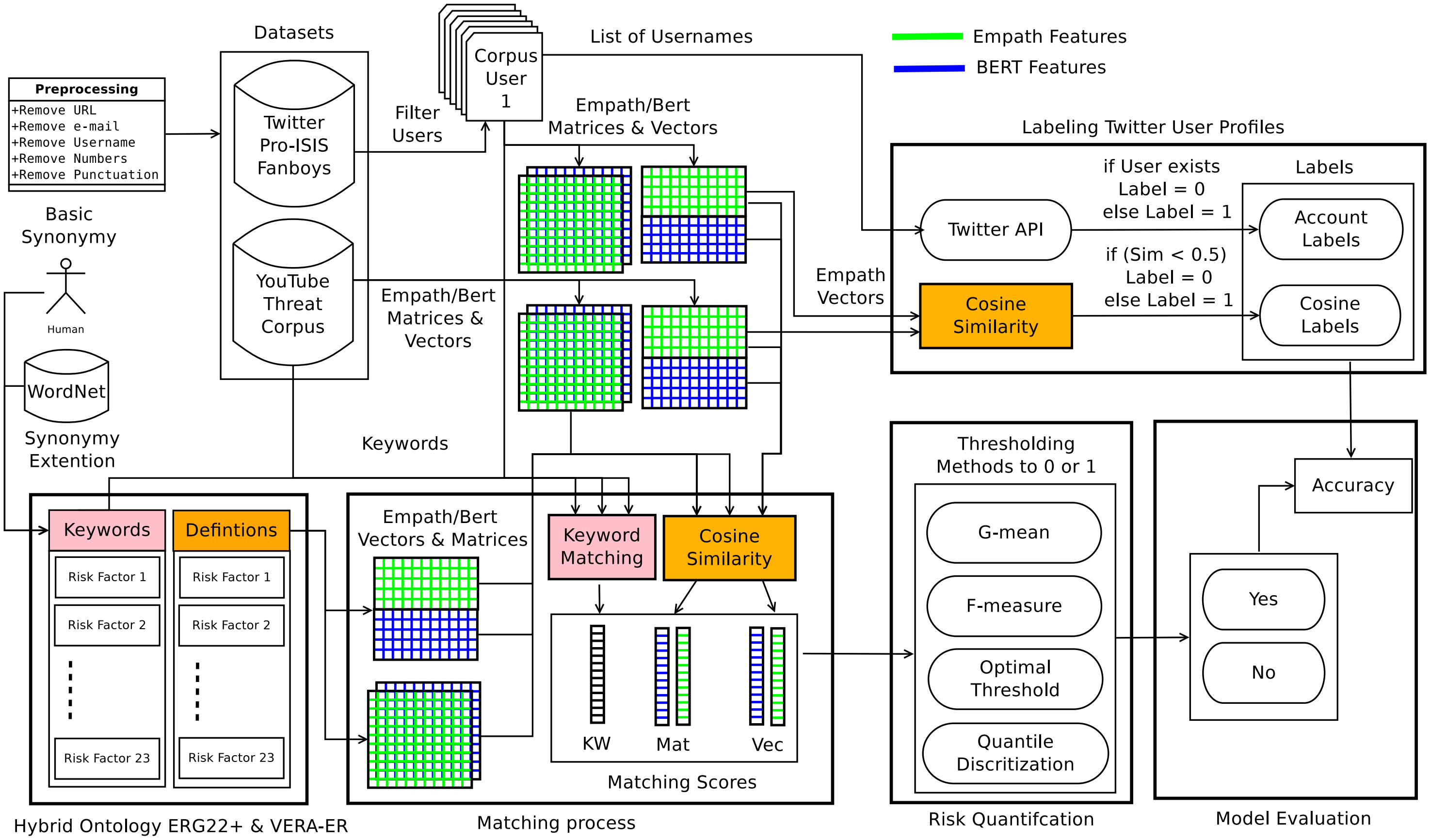

Fig. 1. Generique pipeline.

17,000 tweets from 100+ pro-ISIS [3] supporters worldwide have been reported. The dataset includes attributes: name, username, location, number of followers, number of statuses, timestamp, and the tweet in different languages. The tweets are dominantly written in English, although, we may notice some Arabic tweets as well. The content of a tweet might be connected to a propaganda video link or promoting anti-US and anti-western countries slogans using various hashtags. However, unlike Youtube dataset, the processing of this dataset is challenged by the lack of formal ground truth, which motivated the development of automated data annotation approaches as highlighted in the generic pipeline illustration.

D. Data preprocessing

Standardized text preprocessing techniques have been performed to eliminate any noise and inconsistencies from the gathered text that will influence the matching process. The preprocessing is slightly polished to accommodate the nature of source data employed (Youtube data and Twitter) where Twitter dataset is usually highly noisy and ignoring some relevant characters (e.g., #,) can yield significant gap). In overall, the preprocessing includes the following functions:

- Remove emails and URLs.
- Replace combined tokens by separate ones, e.g., *"hasn't"* becomes *"has not"*.
- Remove Stopwords.
- Remove distracting single quotes.
- Remove punctuation, extra spaces, Numbers, user mentions, Emojis, reserved words (RT, FAV), hahstags.

E. Matching user textual input to hybrid ontology

The process of matching individual post content to the risk indicator ontology has been considered from two perspectives. The first one performs this matching process at each post of an individual user and then aggregates all all posts of the same to user to yield an overall assessment with respect to each risk factor. The second one concatenates all posts of an individual user as a single document that is then matched to each risk factor to yield a single individual assessment score. The first approach yields a matrix evaluation score with respect to number of posts of the user and number of risk factor ontologies in hybrid eERG22++ -VERA-ERA, while the second approach yields a vector representation corresponding to the matching score for each risk factor, see Fig. 2. Intuitively, the matrix and vector-based approaches correspond to two decision strategies where in the former we tolerate to judge about individual's radicalization on flight according to his current statement, which sometimes does make a sense too, for example when the

[3] https://www.kaggle.com/fifthtribe/how-isis-uses-twitter?select=tweets.csv

user stated his willingness to perform a violent act. While, in the vector-based approach a more cautious attitude towards risk assessment is judged necessarily to take into account the context user's statement and possibly any psychological, amusement, rumour impact.

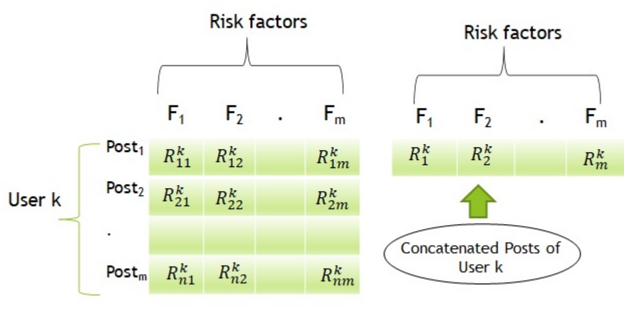

Fig. 2. Matrix versus vector risk assessment

Therefore, let F_i, i=1 to m, be the risk factors in ERG22+-VERA-ER ontology, and let R_i^j be the risk assessment of the i^{th} user with respect to j^{th} risk factor F_j when considering all his/her posts (in case of vector-based approach), then the overall risk assessment of User i is provided by (1), as per hypothesis $H-4$:

$$Risk_i = \underset{j=1,m}{max} R_i^j \tag{1}$$

Similarly, in the case of matrix-based approach, a counterpart of (1) is the following:

$$Risk_i = \underset{j=1,m}{max} \Phi(R_{i,1}^j, R_{i,2}^j, ..., R_{i,n}^j) \tag{2}$$

where $\Phi(.)$ stands for some aggregation function of the risk assessment of individual posts of the user. Especially, in a prudent attitude, the risk factor can be dedicated by the most risky post in terms of the underlined risk factor content, which is translated into a *max* combination operator where $\Phi(.) = max(.)$, while an incautious attitude can be translated into a *min* combination operator ($\Phi(.) = min(.)$).

On the other hand, the quantification of individual risk assessment score R_i^j of User i with respect to F_j risk factor in the hybrid ERG22+ & VERA-ERA model is performed solely on the basis of the textual matching in according to H_2. For this purpose two competing approaches that use embedding and deep-learning models are developed for this purpose: Empath [20] feature matching and BERT [14] model matching in line with H_3. While a third approach that uses standard string matching taking into account keyword augmentation is employed as baseline model. Below these three approaches are detailed.

1) Embedding-based approach: Both Empath and BERT embedding are detailed in this subsection. Empath [20] imitates the concept of LIWC (Linguistic Inquiry and Word Count) [51] and yields a set of categories with associated weights for which the input word or sentence likely matches. The model uses a neural embedding model trained on more than 1.8 billion words of modern fiction and using 194 built-in, pre-validated categories. For example, the text (*bleed* and *kill*) will be categorized as *violence = 1.0, crime = 0.12, prison = 0.12, pain = 0.37* and zeros for the other categories that are not triggered by these terms. In overall any textual input yields an embedding vector of 194 components indicating the level of matching to each predefined categories.

Similarly, the Bidirectional Encoder Representations from Transformers (BERT) architecture [14] released by the Google research group in 2018 becomes nowadays the state-of-the-art in many NLP applications. Unlike other word embedding techniques such as Glove or Word2Vec, which provide a feature vector for each word of the text sequence, BERT delivers a way to encode the entire text sequence into a single feature vector taking into account the word order and context. For each textual input, it generates a 768 size encoding vector. Therefore, for a given risk factor, say, F_i and k^{th} post L_k^j of User j, the associated individual risk assessment score $R_{i,k}^j$ is determine as a cosine similarity of the embedding vectors generated by empath categorization on statement (s) associated to F_i and k^{th} post L_k^j of User j:

$$R_{i,k}^j = \frac{Empath(F_i) \bullet Empath(L_k^j)}{\|Empath(F_i)\| \,.\, \left\|Empath(L_k^j)\right\|} \tag{3}$$

The counterpart of (3) in case of use BERT embedding is provided by (4):

$$R_{i,k}^j = \frac{BERT(F_i) \bullet BERT(L_k^j)}{\|BERT(F_i)\| \,.\, \left\|BERT(L_k^j)\right\|} \tag{4}$$

(3) and (4) apply in case of matrix-based methodolgy, when the risk assessment is performed is performed at each post of the user. In this case, the aggregation of risk score across all posts is performed using mean operator; namely, for a User j who has n posts, the overall risk score with respect to j^{th} risk factor is:

$$R_i^j = (1/n) \sum_{k=1}^{n} R_{i,k}^j \tag{5}$$

Alternatively, if all posts, say L^j for User j, are concatenated together (yielding a vector-like representation as in Fig. 2), the risk score are calculated:

$$R_{i,k}^j = \frac{Empath(F_i) \bullet Empath(L^j)}{\|Empath(F_i)\| \,.\, \|Empath(L^j)\|} \tag{6}$$

And

$$R_{i,k}^j = \frac{BERT(F_i) \bullet BERT(L^j)}{\|BERT(F_i)\| \,.\, \|BERT(L^j)\|} \tag{7}$$

Finally, from the risk score associated to each risk factor, the overall risk score of a given is calculated as:

$$R^j = \underset{k}{max} R_k^j \tag{8}$$

2) String matching based approach: The basis of string-matching is to use the expanded list of keywords generated by the use of WordNet lexical database for synonymy relation on tokens of the risk factor definition statements as pointed in the generic pipeline illustration of Fig. 1. Then a modified Jaccard similarity like measure is used to quantify the amount of overlapping between an individual post k of a user j, represented by a bag-of-words $Post_k^j$ and a risk factor F_i, represented by the bag-of-word $VocF_i$ of its expanded tokens, as in Eq.(9).

$$R_{i,k}^j = \frac{\left\| Post_k^j \bigcap VocF_i \right\|}{\left\| Post_{i,k}^j \right\|} \tag{9}$$

Similarly to embedding case, the risk score of individual with respect to a given risk factor is calculated as the average over all the risk score of all its individual posts. Whereas, in case all posts of a given individual are concatenated, the $Post_k^j$ is substituted by the concatenated input $Post^j$. Finally, the overall individual risk assessment is computed as in (8) by maximizing over all risk factor results.

F. Risk quantification

The previous two subsection provide a basis for quantifying the individual radicalization risk R^j of User j as a numerical score in the unit interval. In order to accommodate the context of our study and the annotated dataset, a binarization is required to transform individual score into high risk or low risk quantification. For this purpose, we adopted the following thresholding strategies:

- *Geometric mean*. The Geometric Mean or G-Mean is a metric for imbalanced classification that seeks to optimize the balance between the sensitivity and the specificity. G-Mean uses all the thresholds from Receiver Operating Characteristic (ROC) Curve, where the optimal threshold would produce the most significant G-Mean value [58].
- *F-measure*.In this case, the threshold is chosen so that the F-measure on the training dataset is maximized.
- *Optimal Threshold Tuning*. This approach is similar to the grid-search method, selecting the optimal threshold among others with the largest F-Measure. The evaluation involves applying a single threshold on the predicted probabilities and mapping all values equal to or greater than the selected threshold to 1 and all values less than the threshold to 0.
- *Quantile-based discretization*. The automatic thresholding uses the Quantile-based discretization function to select the best threshold that maximizes the accuracy of the training set to apply it on the test set to measure the total accuracy of the system eventually. Quantile-based discretization is one of the approaches used in the discretization process [32]. This process is used to transform continuous variables, models or functions into a discrete form by creating a set of contiguous intervals (bins) that go across the range of the desired variable, model, or function [31].

IV. Experimental results

A. Labeling Twitter Pro-ISIS fanboys dataset:

In contrast to Youtube dataset, Twitter Pro-ISIS fanboys dataset is not annotated. Therefore, a labelling process needs to be performed. Strictly speaking, text labelling is a complex and tedious process involving human judgment and sometimes crowd-sourcing and/or automatic techniques depending on the nature and structure of dataset. For instance, studies in [2], [6] advocated the use of sentiment analysis as a labelling technique to discriminate between threat and non threat. Others studies, e.g., [5] considered the state of the Twitter account of the user, speculating that a twitter user who shares inappropriate language is likely to be deleted or suspended by Twitter. In our study, two distinct approaches are pursued.

Approach 1. The first approach follows the Twitter account activity assuming that the user is considered a high risk profile if his Twitter account is banned.

Learning from Youtube annotation. In this original automated procedure, the goal is to learn from the annotation made by Youtube dataset. Formally, we take the embedding vector (calculated using either BERT or Empath features) of every threat user in Youtube dataset. Similarly, for a given Twitter user dataset, we compute the corresponding embedding of its concatenated posts and then calculate the cosine similarity with every (high risk profile user) vector embedding in Youtube dataset. If there exists at least one similarity score whose value is beyond some predefined threshold, then the corresponding Twitter user is judged high risk profile, otherwise, it is annotated as low risk profile. Algorithm 1 shows the labelling processes for the witter Pro-ISIS fanboys dataset. See also Fig. 4 and Fig. 3 for an illustration of the annotation results when using Youtube dataset and Twitter account status, respectively. A quick reading of these illustrations reveals that the use of Twitter account status method leads to a classification of almost all users as threat (high risk profile), which may render the evaluation of the developed method non-effective due to strong class balance. We therefore adopted the YouTube-based labelling strategy only.

Algorithm 1 Labeling_Twitter_Pro-ISIS(Threshold = 0.5)

```
1: Twiiter_Labels ← []
2: Threat_Labels ← Threat_Corpus['labels']
3: for User_Empath in Tweets_Empath do
4:     Sims ← []
5:     for Threat_Empath in Threats_Empath do
6:         Sims.append(Cos(User_Empath, Threat_Empath))
7:     end for
8:     if max(Sims) < Threshold then
9:         Tweets_Labels.append(0)
10:    else
11:        Tweets_Labels.append(Threat_Labels[index(max(Sims))])
12:    end if
13: end for
14: return Tweets_Labels
```

B. Results and discussions

1) Exploratory analysis: We initially performed an exploratory analysis to apprehend the scope of the two datasets using WordCloud visualisation. This visualisation provides

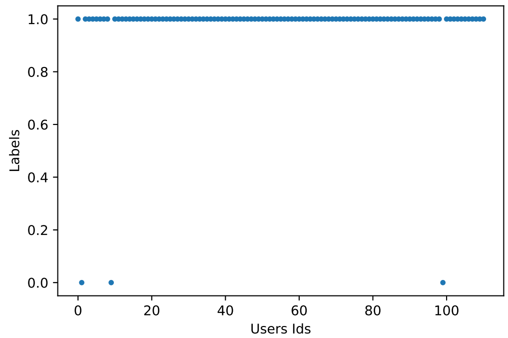

Fig. 3. Twitter Pro-ISIS fanboys dataset using users account status

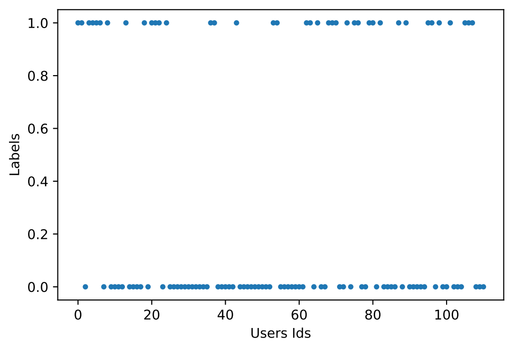

Fig. 4. Twitter Pro-ISIS fanboys dataset Labeling using YouTube

general insights about the most frequent words used in the case of extremism and no extremism, along with the most discussed topics of the two datasets. Figures 5 and **??** show the WordCloud representation of YouTube Threat corpus and the Twitter Pro-ISIS fanboys dataset, respectively.

Figure 5 pictures the frequency distribution of the comments related to threats which are mainly about Islam and killing Muslims in different forms, exemplified using words like 'die', 'death', 'kill', 'shoot', 'booming', 'nuke' and 'burn'. This part of the Threat corpus also shows some racism manifest in the words of 'racist', 'deported', 'white', and 'people', besides cyberbullying Muslims using different cursing words such as 'scum', 'bastard' and 'pigs'. On the other hand, religious conflict and hatred were clearly between religions, such as 'Christianity' and 'Judaism'.

Figure 6 shows the frequent words of the Twitter Pro-ISIS fanboys dataset, where the highlight of the ISIS organization, attacks committed in Irak and Syria can be noticed. It also includes some tragic incidents and reports about attacks in Turkey, Yemen, Burma, where many civilians/children were victims of such terror as well as special operations performed by Turkey, Russia, and the USA. We also notice the mentioning of political and religious conflicts between Muslims and non-Muslims as well as racism. The importance of internet channels in their propaganda is highlighted.

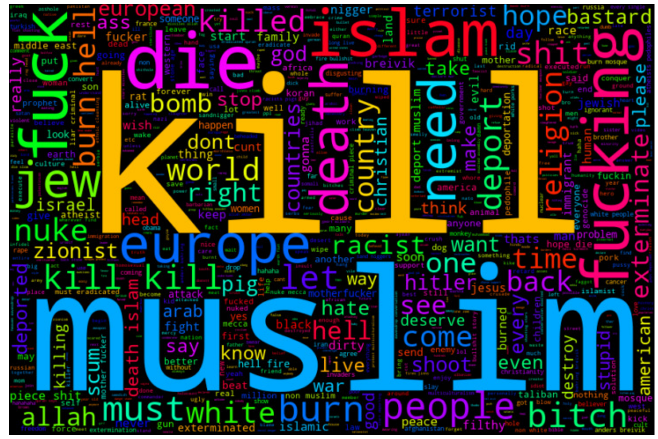

Fig. 5. WordCloud of YouTube Threat corpus

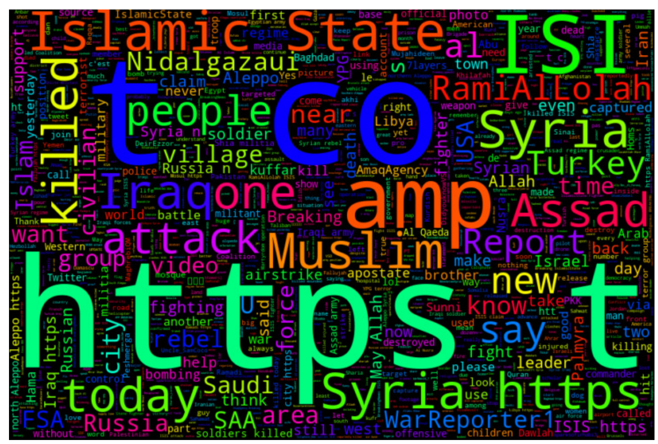

Fig. 6. WordCloud illustration of Twitter Pro-ISIS fanboy dataset

2) Comparative analysis: In this subsection, we evaluate the performance of the various of approaches (string matching, Empath embedding, BERT embedding considering either vector or matrix-based representation) and using various thresholding techniques. In order to find the optimal threshold, the two datasets were split into 80% train and 20% test. The results for Twitter Pro-ISIS Fanboy and YouTube Threat Corpus datasets are summarized in tables III and IV, respectively. In the same table, the optimal threshold value generated by the use of the corresponding thresholding technique is also displayed.

Tables III and IV reveal that the use of BERT embedding at post level (matrix-based approach) yields the best accuracy of

60.9% and 95% for Twitter and YouTube dataset, respectively. The former is obtained using G-mean thresholding with a threshold of 0.01, while F-measure thresholding techniques (with a threshold of 0.04) was used in case of YouTube dataset. Furthermore, the result showed that in the case of Youtube dataset, where the textual inputs are slightly more structured as compared to twitter dataset, the keyword matching can lead to relatively good result as the accuracy achieved 86.4% in case of F-measure thresholding tuning technique with an optimal threshold of 0.154. The same accuracy level is also reached using optimal threshold tuning technique.

The results also show that quantile-based discretization technique gives the best accuracies equal to 73.9% and 64.9%, for YouTube and Twitter dataset, respectively, regardless of the feature representations.

TABLE III
ACCURACY SCORES USING DIFFERENT THRESHOLDS OF TWITTER PRO-ISIS FANBOYS USERS USING YOUTUBE THREAT CORPUS LABELING

Twitter Pro-ISIS Fanboys Users YouTube Labeling		keyword Matching	Empath Vector	Empath Matrix	Bert Vector	Bert Matrix
G-mean	Thr.	0.004	0.480	0.409	0.990	0.010
	Acc	26.1	47.8	39.1	34.8	60.9
F-measure	Thr.	0.020	0.630	0.631	0.980	0.000
	Acc	56.5	60.9	60.9	30.4	39.1
Optimal Threshold Tuning	Thr.	0.001	0.181	0.407	0.000	0.005
	Acc	34.8	30.4	34.8	30.4	34.8
Quantile discreti-zation	Thr.	0.011	0.52	0.523	0.99	0.011
	Acc	60.9	47.84	47.8	34.8	56.5

TABLE IV
ACCURACY SCORES USING DIFFERENT THRESHOLDS OF YOUTUBE THREAT CORPUS

YouTube Threat Corpus		keyword Matching	Empath Vector	Empath Matrix	Bert Vector	Bert Matrix
G-mean	Thr.	0.017	0.340	0.356	0.980	0.010
	Acc	72.0	59.3	49.6	35.3	50.2
F-measure	Thr.	0.154	0.390	0.325	0.980	0.040
	Acc	86.4	65.6	43.9	35.3	95.0
Optimal Threshold Tuning	Thr.	0.143	0.381	0.327	0.971	0.013
	Acc	86.4	65.6	44.0	35.3	62.7
Quantile discreti-zation	Thr.	0.0625	0.440	0.50	0.99	0.0171
	Acc	76.2	73.9	72.0	71.0	72.5

Besides, to comprehend the distribution of the risk assessment scores prior to thresholding step, we present in Fig. 7 the risk assessment scores of all 111 distinct Twitter users when the embedding method is employed either using Empath or BERT model applied to vector or matrix-based representation.

The illustration provides a basis to understand the threshold score generated by the various thresholding techniques provided earlier. We may observe for instance that the vector based BERT embedding yields less variability of risk assessment scores, where the quasi majority tends towards 0.99 value!

3) Discussions:

- The results highlighted in previous subsection where relatively high accuracy rate were obtained (60.9% for Twitter ISIS dataset and 95% for YouTube data) demonstrate the feasibility of our processing data pipeline for assessing the radicalization risk from online content.
- Comparing the vector and matrix representation reveals the superiority of the latter. In other words, calculating the risk level at each of post of the user and then aggregate the risk according to max rule is much more efficient than concatenating all user's posts as a single textual input, which is then used to calculate the risk score.
- The relatively low accuracy obtained for Twitter dataset as compared to YouTube dataset can be rooted back to the impact of the annotation method employed, which is also directly linked to the extent of overlapping with YouTube dataset and not to the explicit content of the Twitter dataset.
- The approach developed in this paper opens up new horizons for radicalization analysis using other ontologies, beyond the employed ERG22+-VERE-ER.

V. CONCLUSION

Terrorism and crime prevention becomes one of the top national priority concerns that helps to protect national assets from foreign and domestic threats. However, this faces complex challenges related to identifying relevant individuals and groups that are considered high risk profiles, especially with proliferation of extremism acts globally. This research uses online discussion data to build a system capable of identifying high risk individuals. For this purpose, the proposed model builds on the well-established radicalisation risk assessment ontologies of ERG22+ and VERA-ER risk assessment tools, where the associated risk indicators are expanded. Each indicator includes different definitions in the form of short text. This expansion creates a representative vocabulary for each risk indicator. The adopted approach assumes two key phases: matching the user's textual input to each risk indicator ontology where the individual risk indicators are aggregated using max-combination rule, and then followed by the binary risk assessment in terms of high- or low- risk profile. For the first phase, two methodologies are contrasted: Embedding-based approach where both BERT and Empath-category are evaluated, and string matching using Wordnet-based expansion vocabulary are employed. In the second phase, various thresholding techniques are compared and discussed. In both steps, we have also contrasted two views of looking into user's post (s) depending whether one wants to assess the individual's risk after scrutinizing all his/her posts or one wants to take a decision on spot (at post level) so that each time a radical and violent post is generated, an action should be taken. Both views are well founded in security studies and cannot be ignored. For the evaluation purpose, we have considered two publicly available datasets: Video Comment Threat Corpus and

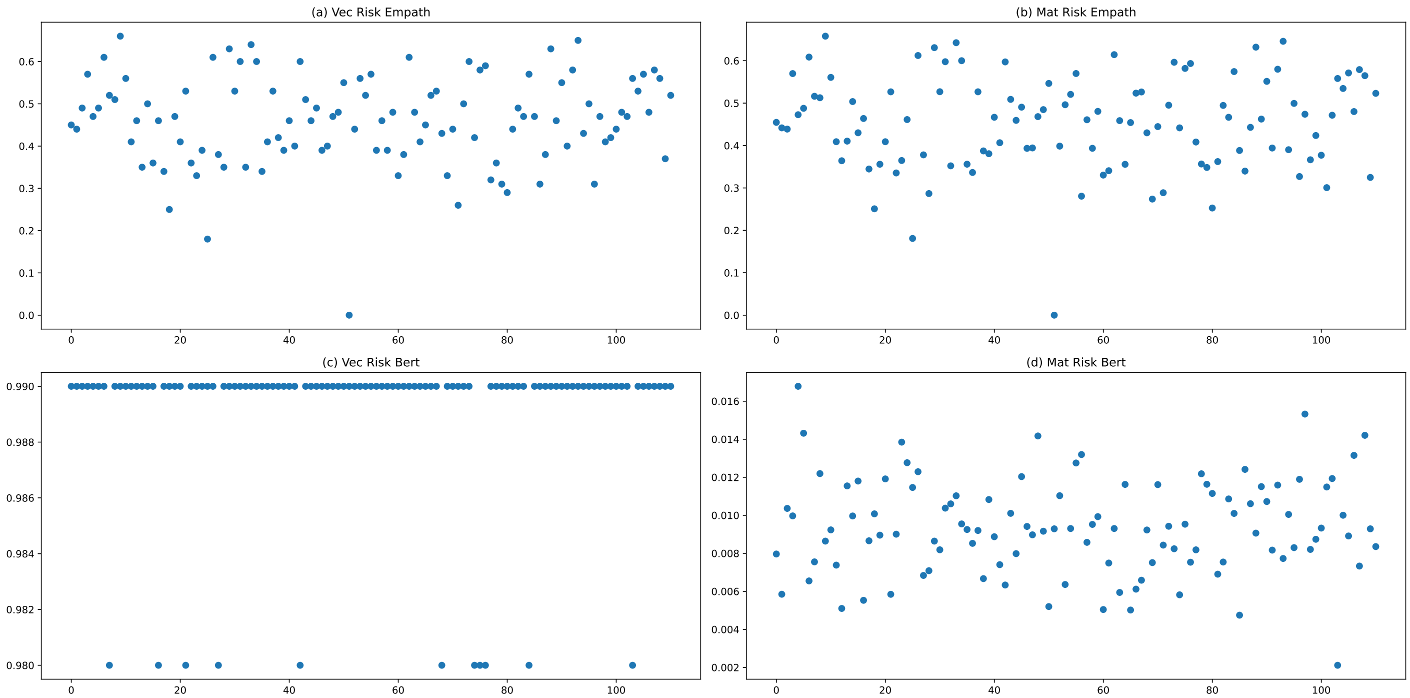

Fig. 7. Risk assessment scores of Twitter users using the hybrid model.

Twitter Pro-ISIS Fanboys dataset. Although the first dataset is well labelled according to the purpose of this study, new techniques have been suggested to automatically label Pro-ISIS dataset. Especially, one approach advocates the view that radicalized Twitter users should have been reported to Twitter, which will then suspend their accounts. The second one uses the knowledge about Facebook labelling as a guideline to label Twitter dataset as well, so that a mapping strategy employed embedding representation was devised and successfully tested. The experimental results in terms of high accuracy rate achieve 95% and 60.9% for Youtube and ISIS dataset, respectively, confirming the technical soundness of the developed approach and its prospects to lead new horizons in tackling radicalization online.

ACKNOWLEDGMENT

This work is partly supported by the European YoungRes project (Ref. 823701) and EU CBC Karelia on IoT and Business Creation which are gratefully acknowledged.

REFERENCES

[1] Max Abrahms. What terrorists really want: Terrorist motives and counterterrorism strategy. *International Security*, 32(4):78–105, 2008.

[2] Shakeel Ahmad, Muhammad Zubair Asghar, Fahad M Alotaibi, and Irfanullah Awan. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 9(1):1–23, 2019.

[3] I. Ajala, S. Feroze, M. El-Barachi, F. Oroumchian, S. Mathew, R. Yasin, and S. Lutfi. Combining artificial intelligence and expert content analysis to explore radical views on twitter: Case study on far-right discourse. *Journal of Cleaner Production*, 362:132263, 2022.

[4] Chris Angus. Radicalisation and violent extremism: Causes and responses. 2016.

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

[6] Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, Hanan Aljuaid, and Jalal Shah. Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, 48:101345, 2020.

[7] Nicola L Beardsley and Anthony R Beech. Applying the violent extremist risk assessment (vera) to a sample of terrorist case studies. *Journal of Aggression, Conflict and Peace Research*, 2013.

[8] Tina Besley and Michael A Peters. Terrorism, trauma, tolerance: Bearing witness to white supremacist attack on muslims in christchurch, new zealand, 2020.

[9] Sarah Brown, Erica Bowen, and David S Prescott. *The Forensic Psychologist's Report Writing Guide*. Routledge New York, 2017.

[10] Michele Burman, Sarah Armstrong, Susan Batchelor, Fergus McNeil, Jan Nicholson, et al. Research and practice in risk assessment and risk management of children and young people engaging in offending behaviour. 2007.

[11] Daniel Byman. How to hunt a lone wolf: Countering terrorists who act on their own. *Foreign Aff.*, 96:96, 2017.

[12] Alana Nicole Cook. Risk assessment and management of group-based violence (doctoral dissertation). *Simon Fraser University, Burnaby, British Columbia, Canada. Retrieved from summit. sfu. ca/system/files/iritems1/14289/etd8437_ACook. pdf*, 2014.

[13] AN Cook, SD Hart, and PR Kropp. Multi-level guidelines for the assessment and management of group-based violence. *Mental Health, Law, and Policy Institute, Simon Fraser University*, 2013.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Kevin S Douglas and P Randall Kropp. A prevention-based paradigm for violence risk assessment: Clinical and research applications. *Criminal Justice and Behavior*, 29(5):617–658, 2002.

[16] Kevin S Douglas, James RP Ogloff, and Stephen D Hart. Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services*, 54(10):1372–1379, 2003.

[17] Kevin S Douglas and Randy K Otto. *Handbook of violence risk assessment*. Routledge, 2021.

[18] Mario Arturo Ruiz Estrada and Evangelos Koutronas. Terrorist attack

assessment: Paris november 2015 and brussels march 2016. *Journal of Policy Modeling*, 38(3):553–571, 2016.

[19] Jon Cole et al. *Guidance for Identifying People Vulnerable to Recruitment in Violent Extremism.* niversity of Liverpool, 2010.

[20] 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.

[21] Brian Fishman. Crossroads: Counter-terrorism and the internet (february 2019). *Texas National Security Review*, 2019.

[22] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262. 2011.

[23] Alasdair Goodwill and J Reid Meloy. Visualizing the relationship among indicators for lone actor terrorist attacks: Multidimensional scaling and the trap-18. *Behavioral sciences & the law*, 37(5):522–539, 2019.

[24] Laura S Guy, Ira K Packer, and William Warnken. Assessing risk of violence using structured professional judgment guidelines. *Journal of Forensic Psychology Practice*, 12(3):270@articleangus2016radicalisation, title=Radicalisation and violent extremism: Causes and responses, author=Angus, Chris, year=2016, publisher=Parliamentary Research Service (NSW) –283, 2012.

[25] Hugo L. Hammer, Michael A. Riegler, Lilja Øvrelid, and Erik Velldal. Threat: A large annotated corpus for detection of violent threats. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–5, 2019.

[26] Stephen D Hart. The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and criminological psychology*, 3(1):121–137, 1998.

[27] Stephen D Hart, Christine Michie, and David J Cooke. Precision of actuarial risk assessment instruments: Evaluating the 'margins of error'of group v. individual predictions of violence. *The British Journal of Psychiatry*, 190(S49):s60–s65, 2007.

[28] Victoria Herrington, Karl Roberts, et al. Risk assessment in counterterrorism. *Countering terrorism: Psychosocial strategies*, pages 282–305, 2012.

[29] Stephen Holmes. Al-qaeda, september 11, 2001. *Making sense of suicide missions*, 164, 2005.

[30] Thomas Holtgraves. Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, 30(2):161–172, 2004.

[31] Yong Huang, Mingzhen Zhang, and Yue He. Research on improved rfm customer segmentation model based on k-means algorithm. In *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, pages 24–27. IEEE, 2020.

[32] Jean Jacod and Philip Protter. *Discretization of processes*, volume 67. Springer Science & Business Media, 2011.

[33] Wm. Robert Johnston. Johnston's archive terrorism, counterterrorism, and unconventional warfare. https://www.johnstonsarchive.net/terrorism/wrjp255us.html, 2021. Accessed: 2021.

[34] Roberts Karl and Horgan John. Risk assessment and the terrorist. *Perspectives on Terrorism*, 2(6), 2010.

[35] Majeed Khader. *Combating violent extremism and radicalization in the digital era*. IGI Global, 2016.

[36] Joshua Kilberg. A basic model explaining terrorist group organizational structure. *Studies in Conflict & Terrorism*, 35(11):810–830, 2012.

[37] Heflbrun Kirk. Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. In *Clinical Forensic Psychology and Law*, pages 347–359. Routledge, 2019.

[38] Liesbeth van der Heide, Marieke van der Zwan, and Maarten van Leyenhorst. The practitioners guide to the galaxy - a comparison of risk assessment tools for violent extremism. *International Center for Counter Terrorism*, 2019.

[39] Thomas R Litwack. Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, 7(2):409, 2001.

[40] Monica Lloyd. Extremism risk assessment: A directory. the centre for research and evidence on security threats (crest), 2019.

[41] Monica Lloyd and Christopher Dean. The development of structured guidelines for assessing risk in extremist offenders. *Journal of Threat Assessment and Management*, 2(1):40, 2015.

[42] J Reid Meloy. The operational development and empirical testing of the terrorist radicalization assessment protocol (trap–18). *Journal of personality assessment*, 100(5):483–492, 2018.

[43] J Reid Meloy and Paul Gill. The lone-actor terrorist and the trap-18. *Journal of Threat Assessment and Management*, 3(1):37, 2016.

[44] J Reid Meloy, Alasdair M Goodwill, MJ Meloy, Gwyn Amat, Maria Martinez, and Melinda Morgan. Some trap-18 indicators discriminate between terrorist attackers and other subjects of national security concern. *Journal of Threat Assessment and Management*, 6(2):93, 2019.

[45] J Reid Meloy, Karoline Roshdi, Justine Glaz-Ocik, and Jens Hoffmann. Investigating the individual terrorist in europe. *Journal of Threat Assessment and Management*, 2(3-4):140, 2015.

[46] John Monahan. *The Individual Risk Assessment of Terrorism: Recent Developments, in: The Handbook of the Criminology of Terrorism, eds. Gary LaFree and Joshua D. Freilich*. Wiley, New York, 2017.

[47] Loo Seng Neo. Detecting markers of radicalisation in social media posts: insights from modified delphi technique and literature review. *International Journal of Cyber Warfare and Terrorism (IJCWT)*, 11(2):12–28, 2021.

[48] Tomasz Pander. A new approach to adaptive threshold based method for qrs detection with fuzzy clustering. *Biocybernetics and Biomedical Engineering*, 42(1):404–425, 2022.

[49] RJ Parker. *Killing the rainbow: Violence against LGBT*. RJ PARKER PUBLISHING, INC., 2017.

[50] J. W. Pennebaker, C. K. Chung, M. I. Ireland, A. L. Gonzales, and R. J. Booth. *The development and psychometric properties of LIWC2007.* LIWC.net, Austin, TX, 2007.

[51] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[52] D Elaine Pressman. Risk assessment decisions for violent political extremism. 2009.

[53] D Elaine Pressman, Nils Duits, Thomas Rinne, and John Flockton. Vera-2r: Violent extremism risk assessment-version 2 revised. *Netherlands Ministry of Security and Justice, Netherlands Institute for Forensic Psychiatry and Psychology*, 2016.

[54] D Elaine Pressman and John Flockton. Calibrating risk for violent political extremists and terrorists: The vera 2 structured assessment. *The British Journal of Forensic Practice*, 2012.

[55] Elaine D Pressman. *Risk Assessment Decisions for Violent Political Extremism*, volume 67. Public Safety Canada, 2009.

[56] J. Qin, Y. Zhou, E. Lai, G.and Reid, M. Sageman, and H. Chen. The dark web portal project: Collecting and analyzing the presence of terrorist groups on the web. *Intelligence and Security Informatics. ISI 2005. Lecture Notes in Computer Science*, 3495.

[57] J Reid Meloy, Jens Hoffmann, Angela Guldimann, and David James. The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral sciences & the law*, 30(3):256–279, 2012.

[58] Somayeh Sadeghi, Davood Khalili, Azra Ramezankhani, Mohammad Ali Mansournia, and Mahboubeh Parsaeian. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. *BMC Medical Informatics and Decision Making*, 22(1):1–12, 2022.

[59] Kiran M Sarma. Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist*, 72(3):278, 2017.

[60] Akimi Scarcella, Ruairi Page, and Vivek Furtado. Terrorism, radicalisation, extremism, authoritarianism and fundamentalism: A systematic review of the quality and psychometric properties of assessments. *PloS one*, 11(12):e0166947, 2016.

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

[62] Allison G Smith. *Risk factors and indicators associated with radicalization to terrorism in the United States: What research sponsored by the National Institute of Justice tells us*. US Department Of Justice, Office of Justice Programs, National Institute of …, 2018.

[63] Sarah Teich. Trends and developments in lone wolf terrorism in the western world: An analysis of terrorist attacks and attempted attacks by islamic extremists. *International Institute for Counter-Terrorism*, 19, 2013.

[64] Liesbeth Van der Heide, Marieke van der Zwan, and Maarten van Leyenhorst. *The Practitioner's Guide to the Galaxy: A Comparison of Risk Assessment Tools for Violent Extremism*. JSTOR, 2019.

[65] Lee Vargen. *Risk domains and factors of the Multi-Level Guidelines: An updated examination of their support in the literature*. PhD thesis, Arts & Social Sciences: Department of Psychology, 2019.
[66] Herrington Victoria and Roberts Karl. Risk assessment in counterterrorism. *Psychosocial Strategies, eds. Updesh Kumar and Manas*, pages 282–305, 2012.
[67] B Yazid, O Mourad, and T Abdelmalik. Semantic similarity approach between two sentences. In *Proceedings of the 5th International Conference on the Image and Signal Processing and their Applications*, 2019.