Detection of Radicalisation and Extremism Online: A Survey

Irfan Tanoli

SZABIST: Shaheed Zulfikar Ali Bhutto Institute of Science and Technology

Sebastião Pais (✉ sebastiao@di.ubi.pt)

Universidade da Beira Interior <https://orcid.org/0000-0003-2337-0779>

João Cordeiro

University of Beira Interior: Universidade da Beira Interior

Muhammad Luqman Jamil

University of Beira Interior: Universidade da Beira Interior

Research Article

Keywords: Natural Language Processing, Language Dependent and Independent, Social Media, Extremism, Radicalisation, Sentiment Analysis

Posted Date: January 5th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1185415/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Detection of Radicalisation and Extremism Online: A Survey

Irfan Tanoli^{4†}, Sebastião Pais^{1,2,3*†}, João Cordeiro^{1†}
and Muhammad Luqman Jamil^{1†}

^{1*}Department of Computer Science, University of Beira Interior,
Covilhã, Portugal.

²NOVA LINCS, New University of Lisboa, Lisboa, Portugal.

³GREYC, Groupe de Recherche en Informatique, Image et
Instrumentation de University of Caen Normandie, Caen, France.

⁴Szabist University, Karachi, Pakistan.

*Corresponding author(s). E-mail(s): sebastiao@di.ubi.pt;
Contributing authors: dr.irfankhan@szabist.pk; jpaulo@di.ubi.pt;
luqman.jamil@ubi.pt;

[†]These authors contributed equally to this work.

Abstract

Introduction: Due to the lack of regulation, the large volume of user-generated online content reflects more closely the offline world than official news sources. Therefore, social media platforms have become an attractive space for anyone seeking independent information. One of the main goals of this work is to clarify concepts such as *Extremism and Collective Radicalisation, Social Media, Sentiments/Emotions/Opinions Analysis*, as well as the combinations of all of them.

Methods: The automatic identification of extremism and collective radicalisation requires sophisticated Natural Language Processing (NLP) methods and resources, especially those dealing with opinions, emotions or sentiment analysis. Text mining and knowledge extraction are also crucial, in particular, directed toward social media and micro-blogging.

Results: The present document comprehends a study on theoretical material, focusing on the main concepts of the subject, including the main problems and challenges, from the different areas that compose online radicalisation research. Understanding and detecting extremism and collective radicalism online has a connection to sentiment analysis

and opinion mining. There are many barriers to understanding extremism and collective radicalisation; one is to differentiate between who is really engaged in the process and who is just eventually talking about it.

Conclusions: The other focus of this work is to find the best ways to identify extremism and collective radicalisation on the internet, using sentiment analysis and focusing on probabilistic methods to create an unsupervised and language-independent approach.

Keywords: Natural Language Processing, Language Dependent and Independent, Social Media, Extremism, Radicalisation, Sentiment Analysis

1 Introduction

In the last few years, the advent of micro-blogging services has heavily impacted how people think, communicate, behave, learn, and conduct their activities. Online social networks, such as Facebook ¹, Twitter ², and Youtube ³, have become a de-facto platform for millions of Internet users for several purposes like sharing photos, videos, ideas and blogging [1]. These popular social platforms make it easier for people to make connections worldwide. Writing posts, sharing articles, videos, links, or tweeting messages enables people to understand each other's constructive views, ideas, and thoughts. One recent event that falls into the scope of our study is the French protest named the "Yellow Vests" ("Gilets Jaunes" in French) [2]. It began as a pacific protest, but later some extremist groups have joined the protests and turned it into a violent protest as was seen in the news: 'Absence of the progress of the movement, inexperience of the demonstrators, the action of extremist groups, the forces of higher duties' [3]. In Portugal, we have witnessed some extreme events as few people were protesting against the actions taken by the police in a tough neighbourhood referred to as 'Bairro da Jamaica'; there were few from an extremist group violently protesting against the politicians [4].

Furthermore, each cluster of tweet messages or posts focusing on a burst topic may constitute a potential threat to society and individuals. However, the overwhelming majority of information posted on social media is harmless and represents casual, conventional, or expressive crowds, as well as noisy data [5]. Researchers and policymakers keep focusing on uncovering the increase in violent extremism among people that lead to extreme events, trying to adopt proper measures to stop it. The work in [6] discussed the use of specific radicalized language within acting and protest crowds on social media that can enhance violent extremism. Subsequently, online terrorist groups study human sentiments by accessing uncensored content(s). These groups gather and analyze the data on public views from social networks and use the data to classify the polarity of public opinions due to the use of concise language in

¹<https://facebook.com/>

²<https://twitter.com/>

³<https://www.youtube.com/>

posts or/and tweet messages [7]. Therefore, online communities enable violent extremists to quickly spread their ideas and so increase recruitment by allowing them to build personal relationships with a worldwide audience for their specific agenda [7].

Due to numerous factors, including the convenience of use and the lack of regulation(s), the vast amounts of user-generated content reflect the offline world more closely than official news sources. Therefore, online social media have become an attractive platform for anyone seeking independent information and eventually, more real news. We have recently been hearing news headlines like the following three:

"Yellow Vests Demonstrate Again in France";
 "Intractable Political Conflict Over the Wall in America";
 "Brexit in Britain, Catalanian Independence in Spain".

More recently, the killing of George Floyd as he was being taken under police custody led to massive protests not only in America but all over the world. These protests are heading towards violent protests and damaging the world badly. More recently, in Portugal, we have witnessed some events that can be categorised as radical. While some people were protesting against the actions taken by the police in a tough neighbourhood known as 'Bairro da Jamaica' ('Jamaica Neighborhood'), there were others, from an extremist group, protesting against the politicians, violently, who defended the people from the first protest [4]. Hence radicalisation in the form of violent extremism is a severe threat to individuals and society.

Generally, there are two broad NLP approaches for online radicalisation detection, i.e., Language Dependent and Independent. Here, by language-dependent, we mean using different NLP techniques, e.g., Rule-based matching, Named Entity Recognition, Aspect Mining, POS-tagging, etc., on a particular language, e.g., English, Chinese and German, etc. While in Language-Independent techniques, language knowledge is avoided [8] and the methods used are essentially based on probability, e.g., machine learning methods like Naïve Bayes, KNN, SVM, or even more recently, the use of Deep Learning.

1.1 Our Contributions

This paper falls under the domain of NLP techniques containing both language dependent and independent approaches. We have conducted an in-depth and rigorous literature review on a sub-topic within NLP techniques for online extremism and radicalisation detection; our review provides insights valuable to the research community that studies radicalism and extremism. To some extent, this work is the first such survey or study of existing literature on social networks in these domains. Although a similar effort is made in the work [9], the focus of the work is more on Intelligence and Security Information (ISI) to detect automatic radicalisation and predict the rise of civil unrest related

events. Our work focuses more on comparing language-dependent versus independent NLP techniques for extremism and radicalisation detection on social media.

This study is valuable for numerous reasons. Firstly, it covers two important research areas, i.e., extremism and radicalisation, and provides a better understanding of these areas; secondly, instead of just providing brief details of different related work based on language-dependent and independent NLP approaches (extremism and radicalism). Having a clear idea about these approaches, we present three different works for each class. We provide and analyse the case studies for each class in-depth to help readers understand the real difference between language dependent and independent techniques. This angle could also help new researchers unfamiliar with specific techniques dedicated to extremism and collective radicalisation. Further, we also highlight challenges and issues with both NLP approaches.

Road Map of Literature Survey

The approaches discussed in this paper are distributed and organised in multiple sections. Section 2 describes some background information regarding Extremism, Violent Extremism, radicalisation, and Collective Radicalisation, the scope and focus of our research problem. Section 3 describes the previous literature work. In Section 4, we present different language-dependent, and independent based works developed explicitly for online radicalisation and extremism detection. Section 5 discusses the issues and challenges with both these approaches in a broad scope. Finally, Section 6 concludes this study and provides future work directions.

2 Background and Research Focus

This section presents the essential elements underlying our study and research proposal to detect *Extremism and Collective Radicalization* in online text. In such a study, one shall start by identifying the main disciplines and research areas that might be involved, and at the same time, provide a precise definition of these critical elements. Therefore, we will start by defining *Extremism, Radicalization*, including *Collective Radicalization*, and then describe the main research areas engaged in studying these topics, particularly those more concerned with their online identification, like *Sentiment Analysis*. Since this work is thought to be more directed toward social networks, we try to maintain this perspective in our analysis.

2.1 Extremism

Extremism is a vague term that can be explained as (i) Taking a political idea to its limits, regardless of unfortunate repercussions, impracticality's, arguments, and feelings to the opponent, and with the intention not only to

confront but also to eliminate opposition [10]; (ii) Adoption of means to political ends which disregard accepted standards of conduct, in particular, those showing disregard for life, freedom and human rights of others [10]. Alternatively, extremism can also be defined as the statistical sense of extreme views concerning the frequent behavioural patterns observed within some aggregate or community. With such a perspective, extreme behaviour(s) infrequently occur(s) within a collectivity; it occupies a tail on the distribution of events. For example, practising intense exercises, following extreme diets, and/or perpetrating extreme violence is rare and unusual in societies and are behaviours that regular people ignore or avoid following. This notion of infrequent deviation qualifies as statistically extreme [11]. Our work considers extremism based on this notion of statistical deviation from what is expected, thus being more likely to be subject to quantitative and objective measurements.

2.2 Violent Extremism

Violent extremism (VE) refers essentially to violence and intolerance. It is a leap further into a deep and dark realm of human feelings and behaviour, generally with significant harmful and destructive consequences, for individuals and societies. Even when defined ahead with radicalism, VE moves a step further and completely undermines pluralism without allowing any different view or thinking. It places a strong emphasis on (dogmatic) ideologies and uses violent and oppressive methods to achieve its political and ideological objectives [12]. As stated in Shmid [12]: “*extremists strive to create a homogeneous society based on rigid, dogmatic ideological tenets*”. Thus defined, VE leaves no room for diversity or engagement. Violence is always accepted as a legitimate means to obtain and maintain political power, which manifests itself in violent attitudes, actions, or both. In [13] VE is defined as follows:

”An ideology that accepts the use of violence for the pursuit of goals that are generally social, racial, religious and/or political in nature”.

So, extremism become violent (VE) whenever subjects see widespread violence and acts of terror as a valid tool to state their ideological views. They may not directly engage in these acts but believe, endorse, and eventually support these acts. There are a number of different possible causes for violent extremism, but they share common patterns. According to Bartlett et Al. [14]: “*VE are underpinned by a multitude of different belief systems and ideological dynamics of religion, socio-economic and personal tribulations; and are dependent on the regional location individuals come from*”. Thus, different driving motivations, can start as extremism and might develop to VE through a process that normally is described as radicalization.

2.3 Radicalism

Radicalism implies a movement in the direction of supporting or enacting radical behaviour [15]. A radical behaviour can be considered: “when serving a given end, people undermine other people’s goals that matter to them

mostly” [15]. There is no universal definition for radicalisation in academia or government [12]. The concept of radicalisation is not as solid and transparent as many seem to take for granted. The Expert Group on Violent Radicalisation, established by the European Commission in 2006, tasked to analyse the state of academic research on radicalisation to violence, in particular, terrorism, stated that “*radicalisation is a context-bound phenomenon par excellence. Global, sociological and political drivers matter as much as ideological and psychological ones*” [12]. From this group, we find a short definition of violent radicalisation: ‘socialisation to extremism which manifests itself in terrorism’ [12].

From a broader perspective, we can see radicalism as a complex process that turns extremists into *violent extremists*, through ideological indoctrination, as noted in [14]: “*Radicalisation is a process that leads an individual down a path to extremism*”. This affirmation means that before an *action phase*, where violence emerges and manifests in society, there is usually an *indoctrination phase* where subjects are targeted by individuals that espouse their extreme ideology [13]. Nowadays, this *indoctrination phase* may use every online resource available, including social networks. Therefore, automatic methods able to identify radical movements and actions in online platforms are of paramount importance.

2.4 Collective Radicalism

Collective Radicalisation (CR) is a broad-spectrum radicalisation process on a crowd, e.g., online communities, aiming to recruit and indoctrinate individuals prone to extremism. It is a kind of amplified radicalism targeting as many subjects as possible. This amplification process has a modern online platform, from the *dark web* to conventional blogs and *social networks*. It represents a severe threat to societies and an increasing challenge to intelligence institutions and law enforcement services [16, 17]. In CR, complex sociological dynamics are observed, among the subjects involved, by using modern technological devices and platforms. Therefore, sophisticated and intelligent surveillance systems are crucial tactical elements to fight and prevent CR from escalating to dramatic outcomes for citizens and society in general.

The previously described phenomena of extremism and radicalisation are the subject of study of different disciplines, like *politics* [16], *psychology* [17], and *sociology* [18]. The latter has a major role and engagement in these studies by focusing on the social patterns and dynamics involved, how people interact, change, and reinforce each other’s narratives and beliefs. It is known that these interaction dynamics contain certain patterns with features that might allow one to identify what is going on. For instance, the kind of connections and messages recurrently widespread and exchanged between community members can reveal that a radicalisation process is on the move. We hypothesise that these messages might be modelled using certain features, involving sentiment lexicon, cultural and ideological expressions, and psychological personality models (e.g., *the Big Five* or the *LIWC*).

2.5 Sentiment Analysis

The sentiment, polarity, and opinion mining or sentiment analysis deal with direction-based text analysis, i.e., text with opinions and emotions [19]. 'Sentiment Analysis' or 'Opinion Mining' is the computational study of people's opinions, attitudes, and emotions toward an entity. The entity can represent an individual, an event, or even a topic [20]. Opinion Mining (OM) is different from Sentiment Analysis (SA), where the former focus on the extraction and analysis of the opinion about something, and the latter (SA) is more about the sentiment and emotional impact that something can cause on an individual, usually expressed through likes or shares of posts (e.g., tweets or Facebook posts) [20]. SA can also be observed as a type of text classification, but dealing with subjective statements that are harder to classify [21].

Looking at SA as a classification process, it can be divided into three main levels: *document-level*, *sentence-level*, and *aspect-level*. At the *document-level*, SA classifies an opinion document according to its polarity (negative or positive). The entire document should be considered as the primary unit of information. At the *sentence-level*, SA is an expressed feeling classified in each sentence. Usually, it starts by assessing whether the sentence is subjective or objective. If the sentence is subjective, SA determines the polarity of its opinion (positive or negative) [20]. However, using these two levels does not provide the necessary details on all aspects of the required entities in many applications. The *aspect-level* classifies a text segment, taking into account the specific aspects of entities. It is necessary to identify the entities and their aspects. The opinion holders can give different opinions on different aspects of the same entity [20].

2.6 Social Media and Radicalisation

Social media is a communication medium that increases people's ability to share, cooperate, make social connections, and take collective action(s) by not following traditional approaches [22]. With time, social media has become an essential tool to urge crowds to gather or reach one common agenda, objective, or view without world boundaries.

Fig. 1 Number of social network users worldwide from 2013 to 2025 [23, 24].

According to statisticians [25], the use of social media and its messaging applications grew 115% year-to-year in 2013. This growth meant that 1.61 billion people were active in social media around the world at that time. The number has advanced according to expectations to more than 2 billion users in 2016 [23]. From 2017 to the present (2021), this number increased at a rate of 7.24% per year, adding 230 million new users each year [24]. Moreover, the number of people worldwide using the internet has risen to 4.54 billion, 7% (298 million new users) since January 2019. In January 2020, there were 3.80 billion social media users worldwide, and this number increased by more than 9 per cent (321 million new users) last year. More than 5.19 billion people around the world now use mobile phones, which are a standard gateway to access social media platforms; the number of users increased by 124 million (2.4 per cent) last year [26].

The use of social networks also has a powerful impact on public opinion(s). Using social media, people can organise mobs and riots, and it is straightforward to recruit people with extremist ideologies for extremist groups [27]. For example, ISIS publishes a monthly online magazine through social media to make people travel to Syria [28]. Hence, social media is an essential factor to promote radicalisation and violent extremism. The policymakers and researchers are proposing and developing methodologies to identify or detect radical elements and to adopt preventive measures to counter them.

2.7 Research Focus

The focus of our research is at the intersection of the three following fields:

- Extremism and Radicalisation;
- Online Social Media Platform(s);
- Language-Dependent versus Language-Independent Approaches for Extremism and Radicalisation (or Collective Radicalisation) Detection;

- Multimodal propose for Extremism and Radicalisation (or Collective Radicalisation) Detection.

We restrict our work to such approaches only, either language-dependent or independent and developed explicitly for online radicalisation and/or extremism detection. The focus is more on analysing these two categories to compare which approach is more effective than the other. The purpose is not to claim one category is better than the other but instead to analyse which approach is more effective and efficient. We also restrict our analysis to studies that experimented with the approach based on user-generated content on social media platforms like Twitter (micro-blogging website), YouTube (video-sharing website), blogs, and discussion forums.

2.8 Literature Review

We followed a straightforward way to collect the papers for conducting our literature review. The collection of research papers is based upon language dependent and independent works specifically developed or proposed for online radicalisation and/or extremism detection. First, we peer review several NLP papers based upon language dependent and independent approaches from 2014 to 2020, presented or generally proposed for extremism, violent extremism, and radicalisation. We further perform a meta-analysis on articles and determine their relevance to the research area we are focused on here. Subsequently, we have decided to select six different articles that seemed to represent the best research dedicated explicitly to online radicalisation detection and comprise language dependent and independent approaches.

3 Natural Language Processing classification

NLP approaches generally rely upon machine learning algorithms (generally based on statistical ML). To perform linguistic processing mainly was dependent upon rules, and these rules are synthesised to process different NLP tasks [8]. Contrary, the ML approaches automatically learns the rules from the training dataset(s). The Rule-based approaches were developed manually by linguistic experts for a particular task. Later, NLP approaches adopted ML algorithms to handle challenges with NLP [8].

Fig. 2 Natural Language Processing models.

NLP tasks based on ML algorithms are mainly carried out with statistical approaches that include stochastic and probabilistic techniques [8]. As shown in Figure 2, current NLP approaches, which are often used to carry out the most crucial NLP tasks, can be divided into two broad categories: language-dependent and language-independent. These two approaches are further categorised into supervised, unsupervised, and rule-based learning.

The currently prevailing technique for solving problems in NLP is supervised learning (statistical learning). The basic idea behind language-independent supervised NLP learning models is that they automatically induce the rules from the training data. Supervised learning can be (a) sequential and (b) non-sequential. The most common sequential supervised machine learning techniques used are *Hidden Markov Model* (HMM), *Conditional Random Fields* (CRF), *Maximum Entropy* (MaxEnt), and *Deep Learning* (DL), while non-sequential supervised machine learning techniques include *Support Vector Machines* (SVM), *Decision Trees* (DT) and *Naïve Bayes* [8]. Unsupervised machine learning can be language dependent and independent, and these approaches do not require training the model. The task is accomplished by finding intra-similarity and inter-similarity between objects. The most common approaches to the unsupervised category are *clustering*, *concept matrix*, and *matrix factorisation* [8].

Rule-based techniques are grounded on sets of rules or patterns that are defined to perform various NLP tasks [29]. The developer defines these rules, and the system follows the procedure accordingly. The most common approaches are POS (Part Of Speech) tagging, dependency parsing, tokenization, stemming/lemmatization, and sentence segmentation. Also, language-independent approaches are mainly supervised, e.g. Narr et al. [30]. However, language-independent approaches can also be unsupervised. For example, the work in [31] presents an unsupervised approach that uses matrix factorisation to extract latent (or hidden) topics from the text. This approach is unsupervised as there is no train/test pre-labelled dataset.

Researchers can also combine specific techniques as needed to perform NLP tasks. The authors in [32] have experimentally shown that the results of their proposed NER technique consisting of Maximum Entropy (ME) and HMM are better than using a single statistical model (supervised). This type of technique is known as a hybrid technique. Hybrid techniques are usually a combination of statistical and rule-based techniques. These techniques use predefined rules that have been hand-built for many NLP tasks and use ML models that automatically induce rules from the training data set. Hybrid techniques generally outperform statistical and rule-based techniques related to data scarcity control to some extent [29]. Rehman and Anwar [33] presented a hybrid technique for disambiguating sentence boundaries in Urdu, which consists of a statistical unigram model and a rule-based algorithm. Therefore, any approach applied or developed for NLP tasks might be dependent or independent from language, as well as a hybrid combination of both.

4 Approaches

This section presents both language dependent and independent NLP approaches that are applied for detecting online radicalisation. Our attention is explicitly devoted to NLP approaches for the automatic detection of radicalisation in online platforms. To be more precise, to satisfy the objective of a language-independent system, we consider probabilistic approaches that can be applied to any language or even in a mix of languages. While for language-dependent, we consider approaches that use POS tagging, dependency parsing, n-gram techniques, name entity recognition, etc. Below, we present three approaches for each category, such that the reader can easily understand the difference between the two. The purpose is not to criticise or compare the case studies but rather to understand the difference between language-dependent and language-independent techniques and how they are designed and developed.

4.1 Language Dependent Approaches

Any NLP approach specifically designed for a particular Natural Language (NL), e.g., English, Chinese, German, or other languages, uses specific linguistic resources from that language, and so it is usually referred to as being ‘language dependent’. Using NLP techniques such as part of speech (POS) tagging, stemming and lemmatisation, dependency analysis, named entity recognition (NER), or using the Wordnet¹ thesaurus, for developing a system to solve a specific research problem can be considered as a language-dependent approach. A simple example of a subtle language dependency happens with n-gram models, which work better for languages that share important typological properties with English. At first glance, n-grams perform independently from any language knowledge since they treat NL-text as a simple sequence of symbols, automatically reflecting the ‘hidden’ structure by inferring the

¹<https://wordnet.princeton.edu/>

distribution of words in different contexts (flat, unstructured) [34]. However, language typological elements, like the alphabet and word formation, play a role in the induction and use of these models. To better understand this, we present and discuss three language-dependent sample approaches specially developed for online radicalisation detection.

4.1.1 Approach 1:

The first study related to radicalisation detection is presented in [35]. The authors propose a set of radicalisation indicators and a model to assess them using social network data published by several Islamic State of Iraq and Sham (ISIS) sympathisers. According to the authors, it is easy to measure radicalisation vulnerability factors in their study, like socioeconomic and demographic conditions that make jihadist militants suitable targets (e.g., extremist people). Moreover, radicalisation is also triggered by feelings, basic needs, personal life incidents, emotions, and experiences. People usually start their radicalisation by reaching out to radical individuals or groups and digging into extremist ideas when they are seeking to fulfil the needs above. These groups provide social recognition and a sense of belonging, which promotes the ingress into extremist networks and active membership. In the end, anger, humiliation, frustration, and hatred in people towards others lead to violent extremism and radicalisation [35]. Since Social Networks enable radicalised people to reach vulnerable individuals and trigger their radicalisation process, it is necessary to counter such radicalisation risks before it harms people and society [35].

This work presents some results related to a project called Risk-Track 1 [36], whose main objective is to develop a monitoring tool based on social networks to assess the risk of radicalisation. This project focuses on the extraction of radicalisation factors in social networks and developing a detection tool. The work corresponds to the first step in the development of a risk assessment tool in social networks. The objective is to define (and validate) the different indicators that will later be used to identify those members of a social network at high risk of radicalisation.

The focus of the work is the use of social media by only taking into account indicators that can be measured by the activity of the target users on social networks. The authors classify five indicators (Fig 3) and merge them into two categories: attitudes and beliefs towards the Muslim religion and Western society and personality and interpersonal relationships, as summarised by an expert on radicalisation. The former are indicators that are measured by the content of the tweets, while the latter contains indicators related to the specific writing style of each user.

The first two indicators are personality-oriented, i.e., the individual is frustrated and introverted. At the same time, the last three are attributes and beliefs-oriented indicators, i.e., discrimination perception for being Muslim, showing negative ideas about Western society, and positive ideas regarding Jihadism. It was observed that an introverted and frustrated person has a

high chance of becoming a target of radicalisation whenever sharing some common ground of beliefs with existing radicalised groups and networks [35]. The other three indicators are related to religion and consist of a Muslim person, someone's hate towards Western society, and/or expressing positive thoughts regarding Jihadism.

Fig. 3 Data preprocessing, cleaning, and analysis to compute indicators.

The evaluation is carried out as shown in Figure 3 following the series of steps collecting and processing the text (data) obtained from social networks (i.e. the posts, tweets, and/or status updates). Next, the messages are divided into sentences using a sentence tokenizer because some indicators are based on this language unit during the data representation phase. As data is processed and transformed, the indicators can be computed as follows: Frustration (indicator 1) indicates the use of swear words and a sentence with a negative connotation. The methodology proposes the frequency count of swear and negative words for computation and normalises them like other indicators. Introversion (indicator 2) refers more to the length of sentences and the use of ellipses. The model counts average sentence length for all users and the number of ellipses in users tweets to compute this indicator. The value is then normalised by dividing it by the maximum value obtained and obtaining an indicator in the $[0, 1]$ interval. The last three indicators are related to keywords that show religious views, negative thoughts about western society, and positivity towards Jihadism. Sets of keywords are defined, and the model counts how many times at least two keywords occur in a sentence, with subsequent normalisation of these counting [35].

For experimentation, a dataset from Kaggle, with over 17k tweets from numerous Twitter account of pro-ISIS, since the Paris attack in November 2015. Out of 112 user accounts, only 78 unique descriptions are identified (More details in Section V of Lara-Cabrera et al. [35]). Next, the dataset is processed to obtain the indicators and analyse standard features of the user, similarities in indicators values, and the correlation between metrics.

In the results presented in Lara-Cabrera et al. [35] (Section VI), the first impression of the indicators reveals that values are distributed on the lower zone of the range of possible values $[0, 1]$, with some outliers in the mid-range as well as high values in the case of *swearing* and *negative words*. This highlights the fact that there are many users with similar values of the indicators and a few users whose indicators are far away from the former. A noteworthy result from this analysis is the lack of correlation between the average sentence length and the rest of the variables, even when this correlation is generally high. The authors found strong correlations between expressing positive ideas about Jihadism and both the perception of discrimination ($\rho = 0.831$; p-value $\downarrow 2.2\text{e-}16$) and the use of swear words ($\rho = 0.857$; p-value $\downarrow 2.2\text{e-}16$) as well as the use of negative words ($\rho = 0.813$; p-value $\downarrow 2.2\text{e-}16$)

To conclude this work, the contributions of this paper are the following: it proposes five indicators and their corresponding metrics that can be used to measure the online radicalisation assessment of a given individual using public data obtained from social networks. An experimental evaluation for these indicators was carried out using a public dataset of tweets obtained from Kaggle. The work analysed the pairwise correlation between the different indicators. In general, there are strong correlations between the majority of the indicators defined in this work: expressing positive ideas about Jihadism, the perception of discrimination, swearing, and negative words. This work also proposed an appropriate step towards the fight against radicalisation. As stated by the authors, once an individual has been identified as radical, several actions can be performed to stop this individual from becoming a Jihadist [35].

4.1.2 Approach 2:

In Hung et al. [37], a supervised and language-dependent approach is presented to indicate radicalisation behavioural indicators. The authors use NLP techniques and supervised machine learning models to classify textual data in notes and reports from analysts and researchers to radicalisation behavioural pointers. These efforts to generate structured knowledge will result in an operational capability to help analysts quickly search for clues and indicators of risk in law enforcement and intelligence databases.

This research aims to help law enforcement and intelligence agencies promote state of art in identifying domestic radicalisation to violent extremism and preventing future extremist attacks. According to the authors, the new approach to threat assessment, signals a shift in the understanding of violent extremism, builds on evidence of increasing extremism behaviour, and

tracks clues about changes in an individual's behaviour that indicate increased concern about progression toward violence—violent extremism.

The dataset used in this work [37] is part of the *Klausen Western Jihadism Database* (WJDB) [37], a collection containing information on approximately 6,600 individual jihadists of the western origin or residents who have been involved in criminal terrorist acts. The authors first read various publicly available documentation to corroborate the 24 different behavioural indicators believed to be related to radicalisation and record the dates when such behaviours were publicly observed. The authors manually extracted a core set of sentences and sentence fragments that were used to create a tagged dataset for training machine learning models [37]. To reduce the initial complexity and problem space, the team reduced the number of indicators (or variables) from 24 to a *critical* subset of 10 (see Figure 1 in [37]). Then 1273 sentences from various sources (e.g., *The Western Jihadism Project* and terrorist profiles) were extracted [37]. Since many sentences refer to more than two indicators, a training dataset was created that consists of more than 1619 marked sentences or paragraphs. Figure 2 in [37] shows the number of occurrences for each of the ten radicalisation indicators in the training dataset. The proposed approach was based upon three different techniques: radical behaviour extraction from the text (Named Entity Recognition (NER)), rule-based matching, and multi-label text classification using SpaCy, Stanford NLP and Prodigy [37]:

NER Classifiers: are applied with keywords and keyword phrases, serving as training data on all radicalisation objects and generic radicalisation indicators to develop custom NER classifiers by highlighting phrases.

Rule-based Matching is an annotation method for finding specific patterns of tokens in text. It requires the manual production of an extensive set of rules based upon keyword phrases from the training dataset. SpaCy² provides a robust rule matching engine while allowing the implementation of any number of rules. The authors used SpaCy for rule-based matching.

Multi-Label Text Classification is a popular approach for the information extraction problem in which text documents are assigned one or more categories or labels [37]. A workflow is implemented using Prodigy, an text annotation tool that supports multi-label categorisation of training data and convolutional neural network (CNN) models for text classification using SpaCy.

Fig. 4 Processing pipeline of the two-stage classification model. The first model is a binary classifier for relevant sentences to radicalisation indicators in general, agnostic to any particular indicator. The second model is a multi-label classifier to identify the specific probabilities for each of the ten indicators [37].

²<https://spacy.io>

For evaluation, the SpaCy multi-label text classification model for neural networks, which was trained in more than 1273 marked sentences, and a division of the data into training and test sets using 90% and 10% of the data, respectively resulted in a model with 80% of precision, 71% recall, 75% F-score, and 95% accuracy. Figure 5 shows a sample of the output of the model to evaluate sentences with suitable indicators of radicalisation. It is also clear that the model can detect sentences related to the behaviour of individuals who try to join a foreign terrorist organisation (sentences 4, 18) and even the possible presence of peer immersion if the perpetrator asks for help from others). However, with limited data, the model still lacks the desired generalisation, revealed by subsequent tests in new unseen documents.

According to the authors, there are false positives, mainly appearing in radicalisation indicator classification sets. Therefore, another model was implemented with a two-phase processing pipeline, as shown in Figure 4. The model checks the sentences for radicalisation indicators in the first phase. The second model then classifies those sentences that cleared the first phase for the actual type of indicators.

[4] A Virginia man who was allegedly attempting to travel to Syria to join Islamic State and a man accused of helping him have been arrested.		
JOINS_FTO 0.996	INFO_SEEK 0.070	DES_ACT 0.060
[7] .Officials also arrested 25-year-old Mahmoud Amin Mohamed Elhassan, who they say drove Farrokh to Richmond.		
PEER_IM 0.880	JOINS_FTO 0.055	DES_ACT 0.040
[15] .He later met with two other FBI informants he believed were people who could help him join Isis.		
PEER_IM 0.887	JOINS_FTO 0.181	INFO_SEEK 0.046
[18] .He also allegedly said that he wanted to die a martyr but did ask if his wife and family could eventually join him in Syria.		
JOINS_FTO 0.974	DES_ACT 0.940	LIFE_CHG 0.024
[21] .He asked the opinion of one of the FBI informants of his plan to buy a round-trip plane ticket and reserve a hotel room in Jordan, to minimize suspicion.		
STEPS_VIOL 0.925	PEER_IM 0.090	SEEK_REL_AUTH 0.062
[26] When questioned by FBI agents after he dropped Farrokh off, Elhassan said Farrokh was traveling to California to attend a funeral and would be back in two weeks, court documents said.		
STEPS_VIOL 0.669	JOINS_FTO 0.591	CONVERT 0.099

Fig. 5 Sample output of multi-label classification model [37].

The *CNN screening model* was trained with a dataset containing 1273 sentences plus 575 negative sentences that were collected from a different source and were not relevant for radicalisation detection. Thus, SpaCy's *CNN screening model* achieved an F-score of 99%, with 99% precision, 99% recovery). To elaborate on the two-phase process, Figure 6 shows seven sentences of a Department of Justice statement on public affairs announcing the transfer of charges

against a suspected radicalised person, as explained in [37]. Figure 6.a shows the result of the sentence indicator classification model. Figure 6.b shows the results of the screening model ('Y' means relevant to radicalisation behaviour, and 'N' means not relevant). A more complete and detailed description can be observed in Section VI³ from [37].

To conclude the presentation of this work, the approach demonstrates progress in the application of NLP and machine learning techniques to address the classification of textual data in online documents and analyst and researcher reports for indications of radicalising behaviour. According to the authors, these efforts fulfil the ability of law enforcement and intelligence agencies to investigate and intercept those who are taking suspicious radicalisation routes into violent extremism.

Output File	Corrected Output File
<p>[1] FOR IMMEDIATE RELEASE:Friday, August 30, 2019.Individual Who Planned Attack in Queens Charged with Attempting to Provide Material Support to ISIS.Awais Chudhary Recorded Videos and Took Photos of the Flushing Bay Promenade And the Worlds Fair Marina in Preparation for an Attack. STEPS_VIOL.0.826 NVIOL_SUP.0.204 INFO_SEEK.0.113</p> <p>[2] A criminal complaint filed today in federal court in Brooklyn charged Awais Chudhary, 19, a naturalized U.S. citizen born in Pakistan, with attempting to provide material support to the Islamic State of Iraq and al-Sham (ISIS), a designated foreign terrorist organization. NVIOL_SUP.0.732 JOINS_FTO.0.005 INFO_SEEK.0.042</p> <p>[3] Chudhary was arrested yesterday, and made his initial appearance this afternoon before United States Magistrate Judge James Orenstein. UFE_CHG.0.519 PEER_IM.0.103 INFO_SEEK.0.081</p> <p>[4] The defendant allegedly planned to conduct a deadly attack in New York on behalf of ISIS, stated Assistant Attorney General for National Security John C. Demers. STEPS_VIOL.0.986 DES_ACT.0.036 PEER_IM.0.027</p> <p>[5] The National Security Division, working with our partners, will remain vigilant in our efforts to identify, disrupt, and hold accountable those who would conduct a terrorist attack on our soil. STEPS_VIOL.0.874 NVIOL_SUP.0.135 CONVERT.0.041</p> <p>[6] I want to thank the agents, analysts, and prosecutors who are responsible for this case and prevented this defendant from carrying out his deadly plans. STEPS_VIOL.0.774 JOINS_FTO.0.278 DES_ACT.0.124</p> <p>[7] As alleged, Awais Chudhary planned to kill innocent civilians on behalf of ISIS and record the bloodshed in the hope of inspiring others to commit attacks, stated United States Attorney Donoghue. DES_ACT.0.493 THREAT.0.413 STEPS_VIOL.0.063</p>	<p>[1] FOR IMMEDIATE RELEASE:Friday, August 30, 2019.Individual Who Planned Attack in Queens Charged with Attempting to Provide Material Support to ISIS.Awais Chudhary Recorded Videos and Took Photos of the Flushing Bay Promenade And the Worlds Fair Marina in Preparation for an Attack. Y.0.999 N.4.54e-05 STEPS_VIOL.0.826 NVIOL_SUP.0.204 INFO_SEEK.0.113</p> <p>[2] A criminal complaint filed today in federal court in Brooklyn charged Awais Chudhary, 19, a naturalized U.S. citizen born in Pakistan, with attempting to provide material support to the Islamic State of Iraq and al-Sham (ISIS), a designated foreign terrorist organization. N.0.999 Y.0.001</p> <p>[3] Chudhary was arrested yesterday, and made his initial appearance this afternoon before United States Magistrate Judge James Orenstein. N.0.929 Y.0.078</p> <p>[4] The defendant allegedly planned to conduct a deadly attack in New York on behalf of ISIS, stated Assistant Attorney General for National Security John C. Demers. N.0.790 Y.0.210</p> <p>[5] The National Security Division, working with our partners, will remain vigilant in our efforts to identify, disrupt, and hold accountable those who would conduct a terrorist attack on our soil. N.0.696 Y.0.315</p> <p>[6] I want to thank the agents, analysts, and prosecutors who are responsible for this case and prevented this defendant from carrying out his deadly plans. N.0.900 Y.0.096</p> <p>[7] As alleged, Awais Chudhary planned to kill innocent civilians on behalf of ISIS and record the bloodshed in the hope of inspiring others to commit attacks, stated United States Attorney Donoghue. Y.0.999 N.3.84e-04 DES_ACT.0.493 THREAT.0.413 STEPS_VIOL.0.063</p>
(a)	(b)

Fig. 6 Output file of sample sentences scored using the multi-label classification model before the screening, and (b) a corrected output file from the two-stage pipeline when the screening model filters for relevance, and then returns only the classification scores for those sentences which provide radicalisation indicator information [37].

4.1.3 Approach 3:

It is about a supervised, language-dependent approach to extract and analyse violent vocabulary shared on social media and to be able to detect the emergence of radicalism [38]. First, the approach relies on a series of collected profiles from social networks interpreted by a domain expert as both *extremist* and *non-extremist* users. Later, the authors focus on their textual content to extract the specific vocabulary from radical and non-radical contexts. This analysed content is generally shared in Arabic, which raises additional requirements that must be considered in data analysis. Ultimately, the methodology attempts to extract a violent vocabulary, weighted according to the degree of violence, using a variety of NLP and data mining techniques [38]. This

³Section VI - Results and Human-in-the-log (HTTL) Multi-Disciplinary Collaboration in Model Development.

study also helps analyse standard malicious content and extract specific violent vocabulary from radical discourses, leading to the discovery of various extremist profiles and, therefore, detecting radicalism in social networks.

The main focus of the approach is the analysis of the Arabic language to extract the violent vocabulary of radical communities. The methodology involves the three phases that will be described next.

Fig. 7 The overall process of the proposed methodology in [38].

Data Collection and Pre-processing: The data collection step involves the extraction of textual content shared by radical and non-radical communities on social media. The authors collect several extremist and non-extremist users from Twitter and YouTube. The process shows the construction of two communities to analyse their textual content(s): (1) the radical community with extremist users and (2) the non-radical communities with non-extremist users. Then, text data is extracted for each Twitter user that belongs to each community. For YouTube users, the extracted data consists of comments about malicious videos, their titles, and description that users liked and shared. The extracted data is further preprocessed by removing diacritics, punctuation marks, numbers, and stop-words from the Arabic text data to obtain the final two datasets for radical and non-radical textual content [38].

N-Gram and Itemsets Mining: This phase focuses on extracting the common n -grams and *itemsets* from the two collected preprocessed datasets. The use of vocabulary by radical organisations always differs from one social network to another. Initially, all the data shared by the radical and non-radical community on Twitter and YouTube are given in n -grams with $n \geq 3$. A few examples of the used n -grams in [38] are represented as follows:

Unigram ($n = 1$): Example: ‘(day) ‘,(state) ’

Unigram ($n = 2$): Example: ‘(Islam state)’, ‘(Bardo museum)’

Unigram ($n = 3$): Example: ‘(United Arab Emirates)’

The next step is the extraction of common n -grams and element sets. An *itemset* is an allocation of n -grams that appear together in the collected data, ignoring the order. For the calculation, each itemset has the support that corresponds to the frequency of simultaneous occurrence of n -grams contained in the dataset using the following formula:

$$Support(itemSet) = \frac{NumberDataContaining(itemSet)}{TotalNumberOfData}$$

Where $NumberDataContaining(itemSet)$ refers to the number of data composing the *itemSet*, and $TotalNumberOfData$ is the size of the overall dataset. For complete details, refer to Subsection 3.2, *N-grams and Itemsets Mining*, from Rekik et al. [38].

The final support for each obtained n -grams or itemsets, both on Twitter or/and Youtube is calculated as follows:

Twitter:

$$SupportItemset(itemSet) = Support(itemSet)_{Twitter}$$

Youtube:

$$SupportItemset(itemSet) = Support(itemSet)_{Youtube}$$

Twitter and Youtube:

$$Support(itemSet) = \frac{Support(itemSet)_{Youtube} + Support(itemSet)_{Twitter}}{2}$$

Next, a weight is assigned to each item and itemset referring to its importance in the dataset. This weight can be calculated as follow:

$$Weight(itemSet) = \frac{(N + 1) - Rank(itemSet)}{(N + 1)}$$

Where N is the number of the extracted frequent n -grams or itemsets and $Rank$ is the order of the n -gram or itemset according its support⁴. Hence, two types of n -grams and itemsets are obtained at this stage, i.e., a set of n -grams and itemsets representing the radical and non-radical context.

Violent Vocabulary Extraction: This phase includes the mining of violent vocabulary. The aim is to examine the collected frequent n -grams and itemsets of both radical and non-radical contexts shared by the extremist and extract the violent vocabulary from the data [38]. Due to the number of frequent n -grams and itemsets used by radical and non-radical organizations, it

⁴Note: Examples of both radical and non-radical itemsets, their supports, and their weights are presented in Table 1, Subsection 3.2, *N-Gram and Itemsets Mining* in [38].

is impossible to classify these common n -grams/itemsets as radical or non-radical. Therefore, a degree of violence refers to the degree of danger and is calculated using the following formula:

$$ViolenceDegree(itemSet) = weight(itemSet)_{Radical} - weight(itemSet)_{NonRadical}$$

Where $weight(itemSet)_{Radical}$ and $weight(itemSet)_{NonRadical}$ are self explanatory. An example of the final obtained violent vocabulary is shown in Figure 8. It can be observed that n -grams and itemsets with a positive degree of danger are annotated as radicals while with a negative degree as non-radicals. Following this approach, the final violent vocabulary collected is used for the detection of extremist users who are involved in spreading radicalisation on social networks.

Itemsets and their English translation	Weight		Violence Degree	Class
	Radical	NonRadical		
{باقية تتمدد قتل} {Residual stretch, kill}	0,83	0	0,83	Radical
{جيش إسلام} {Army of Islam}	0,66	0	0,66	Radical
{قصف، مجاهد} {Shell, mujahid}	0,5	0	0,5	Radical
{جيش عزة قتل} {Army of Azza, kill}	0,33	0	0,33	Non-Radical
{دولة} {Country}	0,16	0,71	-0,55	Non-Radical
{خير} {benevolent}	0	0,86	-0,86	Non-Radical
{سوريا، سلام} {Syria, peace}	0	0,57	-0,57	Non-Radical
{سعادة} {happiness}	0	0,43	-0,43	Non-Radical
{مكافحة، إرهاب} {Counter, terrorism}	0	0,29	-0,29	Non-Radical
{عالم، حب} {world, love}	0	0,14	-0,14	Non-Radical

Fig. 8 Example of itemsets composing the final vocabulary [39].

For evaluation and results, the authors applied different libraries and APIs on the *RStudio* platform that require a list of development tools to perform the main task of automatically collecting violent vocabulary. These libraries include *Twitter API*, *Youtube API*, *arabicStemR*, and rules thoroughly described in the Subsection *Implementation details* in Rekiki et al. [38]. The collected data consists of 8301 n -grams and itemsets annotated as radical and non-radical. The left-hand side in Figure 9 shows the statistics about the radical and non-radical portion of the analysed profile, while the right-hand side shows a portion of analysed data. The statistics about the proportion of radical and non-radical extracted vocabulary is presented in Figure 10.

They resorted to a domain expert who confirmed for each collected n -gram and itemset whether it was a radical or non-radical element for further performance evaluation. Also, accuracy, recall, precision, and f-measure were calculated to ensure effectiveness. The obtained value for accuracy is 0.945, for recall 0.976, for precision 0.951, which results in an f-measure of 0.960. This verified that the proposed approach could extract violent vocabulary from the social networks, according to the authors. Moreover, to validate the expert's

annotation, a sociologist was involved who plays the role of a second expert to estimate the *inter-annotation agreement* between this two experts [38].

In conclusion, the approach in [38] shows a new methodology for violent vocabulary extraction from social networks. The approach collected a list of radical and non-radical users who shared violent extremism and radicalisation on social media. The focus was on users and their written content by extracting relevant n -grams and itemsets from each community. The authors claimed the expert's evaluation evidence that the approach can efficiently extract radical vocabulary, frequently issued by extremists in social media.

Fig. 9 Statistics about the analyzed profiles and their collected textual data [38].

Fig. 10 Statistics about the itemsets composing the extracted vocabulary [38].

4.2 Language Independent Approaches

Language independence is another approach to modern ML-based NLP, which is scalable and effective [40]. Anytime an algorithm is developed for a specific language, the same question arises: can it be trivially extended to another language? All we need is an adequate amount of training data for the new language. However, the usual method for developing a language-independent system is to avoid using any particular linguistic knowledge for its development [34]. As shown in Section 3, the current prevailing techniques for addressing various NLP tasks, through supervised learning, are *Hidden Markov Models* (HMM), *Conditional Random Field* (CRF), *Maximum Entropy* models (MaxEnt), *Support Vector Machines* (SVM), *Naïve Bayes*, and *Deep Learning* (DL).

The current subsection reviews three language-independent approaches developed explicitly for radicalisation and extremism detection on social platforms. We will not discuss or present complete technical details or experimental results and instead give a critical review of the approach.

4.2.1 Approach 1:

An interesting language-independent approach is presented in [41] for addressing affect signs in social media and networks. Internet is now a common and popular medium for terrorist organisations to spread their propaganda and recruitment strategy, which increases radicalisation. Hence online radicalisation detection is a significant concern of counter-extremist authorities [42]. The focus of work [41] is on detecting emotions expressed in online radicalisation publications. To do so, the authors investigate three research questions:

1. Can emotion information be used for radicalisation detection? If so, how?
2. How can radical vocabularies be obtained and exploited?
3. Can semantic similarity-based features be used effectively for radicalisation detection?

With these research questions, this approach depends upon NLP to detect radical text in two domains: online press and Twitter. The proposed approach generates distributed representations of the text that are fed into an ML classifier. These representations are generated by computing the similarity between the analysed text and a particular lexicon. The similarity measure is obtained through a pre-trained word embedding model. In this way, the approach intends to exploit the knowledge contained in word embedding models and a lexical resource.

Fig. 11 General Architecture representation of the proposed model in [41].

The system shown in the Figure 11 is composed of two basic sub-modules: the system is composed of two sub-modules: emotion-based and embedding word similarity, and the general model is shown in Figure 11. The two modules process the input text yielding as output a feature vector that represents the input. These feature vectors are concatenated and fed to a machine learning classifier, which outputs a prediction based on the information given by the features. As classifiers, in this work, both Logistic Regression and Linear SVM are considered.

Emotion Based Features: This module follows the use of an emotion lexicon referred to as EmoFeat (Emotion Features) to extract emotion-driven features that are fed to an ML algorithm. The goal is to investigate whether this kind of information is relevant for radicalisation detection and to which extent. So, an emotion lexicon-based representation is proposed that makes use of statistical measures to encode the emotion of the text [41].

This feature extraction method can be expressed as an algorithm, as shown in Algorithm 12. The function *emotionAnnotation* extracts the emotion annotation vector corresponding to the word w_k in the array L . The *startMeasure* function also calculates the corresponding statistical measure in column j of the matrix E . The index i indicates which statistical measure is used (for example, average if $i = 1$, maximum if $i = 2$, etc.) [41].

Algorithm 1 EmoFeat

Require: Emotion lexicon composed by a vocabulary $W^{(l)}$ and annotations L

Ensure: $v \in \mathbb{R}^{n \cdot m}$, the final feature vector

```

1: for all  $w_k \in W^{(l)} \cap W^{(i)}$  do
2:    $E_{k,:} \leftarrow \text{emotionAnnotation}(w_k, L)$ 
3: end for
4: for  $i \leftarrow 1, n$  do
5:   for  $j \in 1, m$  do
6:      $index \leftarrow i + n(j - 1)$ 
7:      $v_{index} \leftarrow \text{statMeasure}(i, E_{:,j})$ 
8:   end for
9: end for
```

Fig. 12 Algorithm for feature extraction [41].

Embedding Based Semantic Similarity: The *SIMILARITY-based Sentiment Project* (SIMON) module is suggested as a feature extractor for the detection of radicalisation. This technique uses a word embedding model and aligns the extracted features to a particular domain using a domain-centric lexicon. The SIMON method is adapted to extract radicalisation detection characteristics using radical-oriented lexicons. The gist behind SIMON is that

in a domain lexicon, the input text is measured against it and a vector is calculated that encodes the similarity between input text and the lexicon. With such a model, it can use the knowledge contained in the word embedding, as well as the domain information that the lexicon provides. Furthermore, this method does not require large training corpora and can, therefore, be used in problems where the data recorded is scarce [41].

The evaluation is carried out using a text categorisation task, in which the motive for a given text is to determine whether it contains evidence of radicalisation. The evaluation takes place by performing a binary classification task and following a $k=10$ fold cross-validation method. The weighted average of the F1-Scores is used for the performance metric. The performance of different features is analysed in three ways: (i) EmoFeat, (ii) SIMON, and (iii) EmoFeat combined with SIMON.

The result obtained for EmoFeat, the authors, set different parameters for emotion, i.e., several emotions refer as parameter m and several statistical measures considered as parameter n and use of RFE method [41]. With different values for m and n , the performance of EmoFeat is quite reasonable. Next, for the SIMON approach, three variations are evaluated, i.e., word vector, collection of the word, and percentage of filtering over this collection. The results obtained are quite mix for all three variations. Thus the performance of this module is still satisfactory.

Finally, the combination between the emotion-based and SIMON features has been evaluated. The performance of this combination appeared to be not as good compared to the performance when both modules were evaluated separately. However, there are still many cases where this combination improves over the particular method, which shows the combination can lead to an improvement in the performance according.

In short, according to the authors, the contributions of this work are the following: a new dataset for use in radicalisation detection work; a method to use a dictionary of emotions to identify radicalisation; an application to the radical detection domain of an embedding-based semantic similarity model. The results show that emotions can be reliable indicators of radicalisation and that the proposed feature extraction methods can lead to high-performance scores. Furthermore, according to the authors, this work [41] offers a novel approach in which a statistical summary of the emotions present in the analysed text is calculated using a dictionary of emotions. The authors aim to use the existing idea in the literature that emotions can play a role in triggering radicalism. To expand the current scope of this research and future ones, the authors present a new data set, i.e., a collection domain word (FreqSelect). Such data have been collected from radical and neutral online press sources dealing with ISIS-related issues. The original domain differs from the data typically used in previous work, potentially allowing for more extensive studies on radicalisation in texts.

4.2.2 Approach 2:

In [43], a language-independent model, is proposed to identify measures for the automatic classification of radical content on social media. Online social media has changed the dynamics with which terrorist and extremist groups can influence and radicalise people. The dissemination of extremist material online to a wide audience facilitates the increase of radicalisation among people. This radical content needs to be identified online before broadcast. The approach presented in the work [43] identifies several signals, including textual, psychological, and behavioural, which together allow radical messages to be identified. These signals are developed based on knowledge gained from analysing propaganda material called ‘Dabiq’, published by known extremist groups, and use data extraction techniques to computationally discover the contextual text and psychological properties of these groups. The authors mainly focus on the ISIS group, as it is one of the main terrorist groups that use social media to share their propaganda and recruit people. This enables them to create a general radical profile that will be used as a signal to identify ISIS supporters on Twitter. The results show that these identified signals are indeed crucial to improving existing efforts to detect radicalisation online [43].

The approach shown in Figures 13 and 14 is based on two steps: Phase 1: Radical Properties Extraction, in which articles from extremist Dabiq magazines are input to perform two parallel tasks. In the first task, a language model is built using (i) TF-IDF⁵ scores from unigrams, bigrams, and trigrams, and (ii) generated word embedding from a word2vec model. The result of this task is a radical corpus of higher k-grams and a word embedding model that provides a vector representation for each word in the corpus. The second task creates a psychological profile based on the language used in extremist propaganda articles and composed of a series of emotional and topical categories using the LIWC dictionary. Phase 2: The classification of tweets involves using the models generated in Phase 1 to develop functions related to radical activities. The authors classify three groups of characteristics and then train a binary classifier to recognize radical tweets [43].

Feature engineering examines large spaces of heterogeneous features to identify significant features that can be useful in modelling the problem at hand. Three categories of information to identify relevant traits for radical content were examined. Some functions are user-based, while others are message-based. The three categories are 1) Radical language (Textual Feature F_T); 2) Psychological Signals (F_P); and 3) Behavioral features (B_F) [43].

⁵TF-IDF: Term Frequency Inverse Document Frequency

Fig. 13 Approach overview - Phase 1: Radical Properties Extraction

For evaluation, authors acquire a publicly available dataset of tweets posted by known pro-ISIS Twitter accounts published during the 2015 Paris attacks by the Kaggle data science community. The dataset consists of around 17,000 tweets posted by more than 100 users. These tweets were labelled as being pro-ISIS by looking at specific indicators, such as a set of keywords used (in the user's name, description, tweet text), their network of follower/following of other known radical accounts, and sharing of images of the ISIS flag or some radical leaders. Only legitimate accounts are picked because most of the malicious are suspended by Twitter. The authors filter out any tweets posted by existing active users and label this dataset as *known-bad*.

Afterwards, to model normal behaviour, a set of random sample top-trending tweets is collected and filtered with keywords that may be interlinked with extremist views. This dataset contains 8000 tweets by around 1000 users. Two hundred random tweets were manually verified, and it was confirmed that they did not contain any radical views. This dataset is categorized as *random-good* data. A third dataset is selected from Kaggle community [43] with 122k tweets from 95k users. After verification, a subset of 24k users was suspended and are removed from the data sets, and the active users were kept. This data set is refer as *counterpoise* data.

Fig. 14 Approach overview - Phase 2: Tweets Classification

Two different experiments are carried out: in experiment 1, the first two datasets, i.e., *known-bad* and the *random-good* datasets, are used to classify tweets as radical or standard. Classification results are given in Table 1 of [43] using the known bad and randomly valuable datasets. It shows the average accuracy, precision, recovery, and F-measure obtained from each distinct category (F_T , F_P , F_B) and their combination (F_{All}). The authors compared the two text models and found that word insertion results outperformed using n-gram TF-IDF scores. This also confirms that contextual information is essential to identify radicalisation activities.

In Experiment 2, the authors investigated whether the classifier can also differentiate between tweets discussing similar topics (related to ISIS) by using the *known-bad* and the *counterpoise* datasets. Table II in [43] shows the various metrics obtained from each category of characteristics. The F_T feature group achieves an accuracy of 80% and respectively 91% and 100% for the F_B and F_P feature groups. The results are consistent with those of the first experiment, and the characteristics of the group F_P contribute to the high precision of the model [43].

To sum it up, the system presented in [43] offers three main contributions: (1) Analysis of propaganda material published by extremist groups and creation of a contextual model of radical text-based content; (2) A model of psychological properties derived from this material; (3) Evaluation of these models on Twitter to determine the extent to which radical tweets can be automatically identified online. The results show that radical users show distinguishable textual, psychological, and behavioural characteristics, and psychological properties are among the most relevant features. Furthermore, the results show that text models using vector embedding functions significantly improve recognition over the TF-IDF representations. The authors validated the system in two experiments achieving high accuracy. According

to the authors, these results can be used as signals to detect radicalisation activity online.

Fig. 15 Approach overview [44].

4.2.3 Approach 3:

Emotions are another way of expressing viewpoints, thoughts, and positions on certain things. In recent years, people who express emotions on social networks through messages or their posts have become a subject of research due to their influence on the spread of misinformation and radicalisation over social networks. A language-independent approach is demonstrated in the work [44] to identify emotions relevant for inferences from social media messages. The approach is based on NLP techniques and specific heuristics that describe how humans naturally assess emotions in written text.

The proposed approach [44], as shown in Figure 15, comprises two phases: *Preparing the ground truth* and *Emotion extraction*. The authors compare three lexicons: *NRC*, *EmosenticNet3*, and *DepecheMood* available for research works. These lexicons are exploited for performance checking using NLP methods composed of different linguistic features over 7,691 social media messages datasets. To determine the ground truth, the authors used the *ISEAR* dataset along with the results of a questionnaire-based survey in which they asked participants to manually assign emotion scores to 25 Facebook posts and Twitter tweets.

Preparing the ground truth: To evaluate emotion extraction, two online questionnaires were designed for conducting the survey. The first questionnaire aims to measure the impact that the message context has on the human perception of emotions. The first questionnaire contains 15 real-word Facebook comments and ten tweets. The second questionnaire intends to compare the annotator’s tagging at the sentence level (less information about the message context). The first survey was answered by 38 people, while 23 responded to the second survey. In total, 61 people responded to the survey, i.e., 32 male, 29 female; mean age=30.24, sd=8.05). After obtaining the data, the authors computed Spearman’s rank correlation (r_s) to test a correspondence between

human-rated sentences and automated ratings of emotions (based on the NRC lexicon). The calculated values for four categories, i.e., joy, anger, sadness, and fear, are reported in Table 1 of [44]. The notable sample results in that table show that automated ‘joy’ ratings correlated firmly with human rating on the comment level, i.e., $r_s = 0.97$, $p. = 0.01$. Also, the results for correlation between human and automated ratings upon removing context of a sentence dropped to a weak positive correlation i.e., joy $r_s = 0.39$, $p. = 0.01$ and fear $r_s = 0.38$, $p. = 0.01$.

Emotion Extraction: The authors further used the ISEAR dataset, including 7,666 human-annotated entries concerning seven emotions (fear, shame, anger, disgust, guilt, joy, and sadness). However, this data does not include online social network emoticons, e.g., LOL. Therefore, ISEAR is extended for 25 real-world Online Social Networks (OSN) messages and their corresponding annotation collected through the survey. The final established ground truth is composed of 7691 annotated texts resembling people expressed the emotion in the written text explicitly using words like ‘happy, sad’ etc. or using phrases, e.g., ‘I broke my toy’ → sadness. NRC, DepecheMood, and EmosenticNet word-emotion lexicons extract emotion scores. However, a few issues can come up if one only depends on searching words in the word-emotion lexicon, e.g., ‘I am not happy. → joy = +1 (no negation)’, ‘Snakes!!!’ → fear = 0 (no direct lexicon match for words in their plural form).

To counter these issues, an algorithm was designed using NLP-methods, i.e., lemmatization, part-of-speech (POS) tagging, and a number of heuristics (e.g., adverbs of degree and negation) that reflect the way humans determine emotions in texts as shown in Figure 16. In the algorithm, a list of OSN texts $c_1, c_2, \dots, c_N$ is represented by C . Each comment $c_i \in C$ is composed of sentences, $c_i = s_{i1}, s_{i2}, \dots, s_{im}$, and each sentence s_{ij} is composed of words, $s_{ij} = w_{ij1}, w_{ij2}, \dots, w_{ijk}$. There are certain functions implemented in the pseudocode, like *CodeSmiley(i)*, that detect emoticons. The *Lemmatizer(i)* function lemmatizes the words in a message, and *FindMatch(i)* identifies lexicon matches and assigns an emotion score e_i to each comment c_i . A dictionary of emotions, *Secondary dictionary*, and an *emoticons dictionary* were employed, as well as the AFINN lexicon that provides intensities against a word’s effective valence [44].

PSEUDOCODE 1: Emotion extraction.

Data: C , Lexicon_dict , Secondary_dicts
Result: Emotions for each $c \in C$

```

1  $C\_clean = []$ 
2  $Emoticons = []$ 
3  $Lemmatized = []$ 
4 foreach  $c_i \in C$  do
    |  $/*$  clean unnecessary HTML, tags  $*/$ 
5 |  $C\_clean.append(\text{RemoveCode}(c_i))$ 
6 foreach  $c_i \in C\_clean$  do
    |  $/*$  Extract emoticons  $*/$ 
7 |  $Emoticons.append(\text{CodeSmiley}(c_i))$ 
8 foreach  $c_i \in Emoticons$  do
    |  $/*$  Lemmatize the short texts  $*/$ 
9 |  $Lemmatized.append(\text{Lemmatizer}(c_i))$ 
10  $dict\_emot\_v = \{ \}$ 
11 foreach  $c_i \in Lemmatized$  do
12 | foreach  $sent_q \in sentiments$  do
13 | |  $dict\_emot\_v[i][sent_q] = 0$ 
14 | |  $/*$  Split text to sentences  $*/$ 
14 | |  $dict\_emot\_v[i] = \{ \}$ 
15 | |  $sentences = \text{SplitToSentence}(c_i)$ 
16 | | foreach  $s_j \in sentences$  do
17 | | |  $/*$  Split sentence to words  $*/$ 
17 | | |  $words = \text{SplitToWord}(s_j)$ 
18 | | | foreach  $w_k \in words$  do
19 | | | |  $/*$  Find lexicon match  $*/$ 
19 | | | | if  $w_k \in \text{Lexicon\_dict}$  then
20 | | | | |  $/*$  Update emotion score corresponding to  $w_k$   $*/$ 
20 | | | | |  $dict\_emot\_v[i][\text{Lexicon\_dict}[w_k][0]] += [\text{Lexicon\_dict}[w_k][1]$ 
21 | | | | |  $/*$  Find negation and intensifiers  $*/$ 
21 | | | | | foreach  $w_r \in [w_{k-3} : w_{k-1}, w_{k+1} : w_{k+3}]$  do
22 | | | | | | foreach  $d_t \in \text{Secondary\_dicts}$  do
23 | | | | | | |  $/*$  Updating emotions related to  $w_k$   $*/$ 
23 | | | | | | |  $/*$  using negation and intensifiers  $*/$ 
23 | | | | | | | if  $w_r \in d_t$  then
24 | | | | | | | |  $dict\_emot\_v[i][\text{Lexicon\_dict}[w_k]] += d_t[w_r]$ 
25 return  $dict\_emot\_v$ 

```

Fig. 16 Pseudocode Emotion Extraction [44].

In terms of evaluation, the authors had to deal with three different word-emotion lexicon, with different numbers of basic emotions, relying on different psychological models (*Plutchik*, *Ekman*, *Rappler*): NRC (8), EmosenticNet (6), and DepecheMood (8). The intersection of these three models contains four

basic emotions: anger, sadness, fear, and joy, used for inter-lexicon agreement comparison. As the same word might have different values, in each one of these three lexicons, a rescaling to the $[0, 1]$ interval took place, according to the following equation:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Afterwards, each lexicon was represented by a vector of their corresponding scores and the similarities between each pair of the three lexicons were computed through the cosine similarity. These results were reported in Table 2, Section *Result* of [44], and they indicate high similarity between the NRC and EmosenticNet for the emotions, anger, fear, and sadness. However, there is a low similarity when comparing NRC and EmosenticNet with DepecheMood (the results vary between $[0.81, 0.91]$).

The three lexicons were compared against the ground truth obtained from the questionnaires for performance evaluation. The authors used the convolutional precision, recall, and F-measure to assess the classification for the four intersecting emotions: anger, sadness, fear, and joy.

	NRC			DepecheMood			EmosenticNet		
	P	R	F	P	R	F	P	R	F
Negative	84%	86%	0.85	83%	70%	0.76	88%	20%	0.32
Anger	37%	68%	0.48	50%	28%	0.36	29%	7%	0.11
Fear	48%	34%	0.40	41%	12%	0.18	56%	1%	0.03
Sadness	53%	27%	0.36	34%	56%	0.42	56%	23%	0.33
Positive (joy)	56%	52%	0.54	39%	57%	0.46	28%	92%	0.42

Fig. 17 Precision (P), recall (R), and F-score (F) for each emotion and class of emotions with respect to the ground truth [44].

In Figure 17, high precisions of 84%, 83%, and 88% can be observed for negative emotions. However, there is a significant degree of incompleteness for EmosenticNet, with only 20% recall, while NRC has the highest recall, 86%. The overall results also show that the three lexicons are generally able to detect the emotional value. However, the detection of specific dominant emotion is still not precise. For complete details check Section *Results of the Comparison* in [44]).

In summary, the approach in [44] compares the performance of three-word emotion lexicons (NRC, DepecheMood, EmoSenticNet) by applying an algorithm composed of different NLP techniques and few heuristics. Two online questionnaires are used first to determine the ground truth and assess the appropriateness of the found emotional values. The performance is assessed using accuracy, recall, and F-measure. Results show that NRC performs better than the other two in identifying anger, fear, and joy, while DepecheMood

scores better identify sadness. Also, the result shows a high number of false-positive results in the assessment of the correctness of the lexicon along with certain emotions [44].

5 Discussion

In Section 4, we present three approaches as case studies for each category, i.e., language-dependent and language-independent, that are specifically developed for online extremism and radicalisation detection. This section generally discusses critical challenges and issues with both dependents and independent NLP approaches.

5.1 Challenges and Issues with Language Dependent NLP

Language-dependent approaches in NLP are mostly supervised and text analytic. These approaches involve a set of statistical techniques for identifying parts of speech, word/sentence tokenisation, word extraction, phrase extraction, lemmatisation/stemming, rule-based matching, and named entity recognition. These dependent approaches can also be unsupervised. However, the text classification task, which assigns labels to a text, extraction of words, and Named Entity Recognition (NER) is supervised. As described in Section 4.1, a certain technique starts from a set of classified samples, train a model, and use it to classify new samples. A simple example of subtle language dependency is how n -gram models work better for languages that share important typological properties with English 4.1.3. They deal with natural language text as simple sequences of symbols and automatically reflect the ‘hidden’ structure that affects the distributions of words in various contexts (flat, unstructured) contents. However, the efficiency of the n -gram models with natural languages, e.g., English (or other similar languages), is partially predicated on two properties: relatively low levels of inflectional morphology and relatively fixed word order [40]. An example is presented in Section 4.1.3.

Data sparsity is a real issue with language having different morphology, e.g., words with more morphemes and fewer uninflected words. Due to this, the n -grams model’s ability is limited to catch the dependencies between open-class morphemes but also for closed-class morphemes [40]. The information expressed by short functional words in English is typically expressed by inflexible morphology in languages with more sophisticated morphological systems. Word-based n -gram models cannot represent function morphemes in these languages. For n -gram models to catch dependencies between words, they must be displayed in the n -gram boundary. This happens more consistently in languages with a relatively fixed word order than in languages with a relatively free word order [40].

The author argues in [40] that the n -gram models can be developed without manually coding language skills; they are not language-independent. Perhaps, their success depends on the language typological properties for which they were first developed. A more linguistically informed (and therefore less

language-independent) approach for n-gram models is the factored language model presented by Bilmes et al. [45]. Factorized language models address data scarcity problems in morphologically complex languages by representing words as a set of features and thereby capturing dependencies between sub-word parts of neighbouring words. Moreover, to handle data sparsity, it is considered necessary to examine n -gram coverage for a significantly large corpus [46], which is not readily available in many of the non-English like languages. The work in [46] shows that more data is always better and discusses why so much of the language will not even be represented within even massive corpora. More extensive data is better due to a few factors, i.e., the source of these additional data and test documents and the language model, which are trimmed to account for sampling errors and make a meaningful calculation.

Language Specificity: A common issue for the language-dependent approach is its language specificity. For example, the approaches discussed in 4.1 clearly show their construction with language-oriented vocabulary, i.e., English and/or Arabic, and Evaluation of the approach is carried upon using English datasets. Thus, this limits these approaches to test the approach with other datasets, e.g., German, French, Chinese, etc. These approaches are developed or constructed for a particular scenario or context and are defined with language-oriented vocabulary (e.g., English, Chinese, German, etc.), with specific terms or indicators. Later the systems are trained on the vocabulary and apply it to get the desired results, e.g., [47]. Since the vocabulary is defined or designed for a specific context or data is collected for a particular language, the approach cannot efficiently be utilised in other languages.

Low language resource: Low language resource is a pretty general issue for language-dependent systems. While there is a massive amount of data for popular languages, English, Chinese or Spanish, thousands of other languages are spoken by a small number of people and receive much less attention. There are 1,250 to 2,100 languages in Africa alone, but data for these languages are scarce. Besides, the transfer of operation that requires a comprehensive understanding from high-resource to low-resource languages is still very challenging with these approaches [48]. The most promising approaches are the cross-lingual transformer language model ⁶ and the incorporation of cross-lingual sentences [49] that exploit the universal similarities between languages. However, these models are sample-efficient because they only require translation of word pairs or even just monolingual data. Developing cross-lingual datasets, such as XNLI, should make developing more robust cross-lingual models easier.

Evaluation: The Evaluation of the language-dependent approach is a big challenge. Usually, the Evaluation is carried out on small datasets, e.g., 12000 or 17000 tweets are collected for Evaluation, and based on the collected results, the appropriateness of the approach is determined (e.g., check subsection 4.1). In social media's context, getting more data leads to more variability (remember adding new documents to a dataset) or impossible (like getting more resources for low-resource languages). Even if there is the necessary data

⁶<https://github.com/google-research/bert/blob/master/multilingual.md>

to define a problem or task correctly, there is a need to build datasets and develop appropriate evaluation procedures to measure the progress towards concrete goals. However, as NLP is data-driven, it is not easy to understand what kind of data is needed to answer this question. Scarce, unbalanced, and overly heterogeneous data often affect the effectiveness of language-dependent approaches.

Named Entity Recognition: The language-dependent approach mainly relies upon the extraction of named entities (refer also Named Entity Recognition (NER)). The execution of NER is essential when training a system to differentiate between simple vocabulary and named entities. The focus of the NER task is to search a person, location, and brand names, abbreviations, designation, date, time, number, etc., and classify them into pre-defined different categories. For example, the most recent work of the NER task based upon English as a source language for social data is presented in [50]. In general, there have also been surveys focused on NER systems for specific domains and languages, including biomedical NER [51], Chinese clinical NER [52], Arabic NER [53, 54], and NER for Indian languages [55]. However, the approaches based on NER are specifically language-specific, that approaches are limited to a specific language.

Synonymy: People can express the same viewpoint using different terms and words depending on the context. For example, extremism, radicalization, violent extremism, boycott, etc., can be synonymous when talking about a specific topic or context. To perform NLP tasks, it is necessary to consider the knowledge of synonyms and different ways of naming the same object or phenomenon, especially in high-level assignments that mimic human dialogue. One problem with language-dependent approaches is that synonyms are defined for a target language, e.g., recognition of synonyms in Turkish [56], use of WordNet [57] to obtain a synonym for the extraction of event information from social text streams [58], and such approach cannot be easily applied to low resource languages.

Ambiguity in the text is the core challenge for NLP techniques. The hurdle for a language-dependent system is understanding and modelling elements in a variable context. Users write freely on social media, and words are ambiguous; text can have numerous meanings depending on the context, creating ambiguities at the lexical, syntactic, and semantic levels. Ambiguity has different types, such as structural, syntactic, form class, word sense, and local ambiguity [59]. The challenge for the POS-based language-dependent system is mainly the form class ambiguity, as a given the word can be parsed as more than one POS. For example, *conduct* can be a noun or a verb, *account* can be a verb or a noun, *radical* can be a noun and adjective, etc. The ambiguity of form classes inevitably leads to structural ambiguity, as in the famous ‘*Face-book saves your data*’ example. The words *your* and *data* are ambiguous. If *you are* is taken as a possessive pronoun and *data* as a noun, we get a structure of [noun phrase, verb, noun, phrase]. However, *your* as a personal and *data* as a verb, will yield a structure of [noun phrase, verb, [noun phrase, verb]] [59].

In a nutshell, considering these issues and challenges, language-dependent approaches are helpful for specific scenarios or contexts, but in general, these approaches may not be efficient enough to handle complex issues on a large scale with big datasets.

5.2 Challenges and Issues with Language-Independent NLP

There are specific challenges when developing language-independent systems. In fact, in a few cases, a specific language can present a more complex problem than other languages in an objective sense [40]. A clear example of this is converting letters-to-phoneme in English, which is a more complex problem than converting letters-to-phoneme in other languages due to the lack of transparency in English spelling. It is not surprising that those in, e.g., [60] and [61] perform worse with English test data than with German, Dutch or French. However, for a given NLP approach, performance differences cannot be wholly and permanently attributed to language specificities since other issues related to the problem at hand might be involved, being the main reason for such differences. However, it is clear that even language-independent approaches must consider inner linguistic typological features to handle the challenges at hand better.

Big or numerous documents: One of the systematic challenges of language-independent NLP approaches in handling large-scale or multiple/mixed documents, as supervision is scarce and expensive to collect. Current supervised models based on recurrent language-independent neural networks cannot represent a more extended context. Addressing broad contexts are related to *Natural Language Understanding*, and the current approaches/systems should be expanded until they can read process, and *understand* long documents as an entire book, journals, magazines, and movie scripts. However, this method may be too ineffective, and a more helpful direction seems to be multi-document synthesis and multi-documents question answering using unsupervised NLP systems.

Ambiguity: Ambiguity has been a critical issue for NLP researchers for decades. This concept is closely related to the semantic gap between the user's intentions and how people can convey, leading to more than one interpretation of the user's input. For both approaches, i.e., language-dependent and language-independent, the challenge is always to find the exact meaning of the text or context of discussion and elements modelling within a variable context. Although a few results have been obtained on resolving ambiguity issues, several important research issues still need to be resolved [62].

A vital challenge for the supervised language-independent system is *Word Sense Disambiguation* (WSD). WSD is the ability to determine the meaning of a word triggered by using a word in a particular context. Solutions for WSD are mainly classified into supervised and unsupervised (knowledge-based) approaches [63]. Furthermore, the support vector machine and memory-based learning are also the most successful supervised learning methods for WSD [64].

These methods rely on a substantial amount of manually tagged corpora by detection, which is very expensive to create. However, these systems may suffer from the problem of data sparseness, as it is unlikely that an adequately sized training set will be available for comprehensive coverage, especially in the supervision approach.

Since current language-independent approaches are primarily controlled and efficient, the unsupervised language-independent approach could help treat WSD more comprehensively. The neglected systems based their disambiguation decisions on knowledge sources [59]. A recent article posted in [63] proposed a knowledge-based (neglected) method that models the problem with semantic space and the semantic path hidden behind a particular sentence. Unsupervised approaches assume that similar senses occur in a similar context. For this reason, meanings can be induced from the text by grouping word occurrences using a measure of context similarity. This task is known as word sense induction or distinction. Unsupervised methods have great potential to overcome the knowledge bottleneck, as they do not rely on manual efforts [64].

Emotion, personality, and style: Understanding speaker tone and emotion is a crucial challenge for language-independent approaches from a social network’s perspective. Depending on the author’s personality or the speaker, their intention, and emotions, they might also use different styles to express the same view. Some of them (such as irony or sarcasm) may convey a meaning opposite to the literal one. Even though SA has seen significant progress in recent years, the correct understanding of the pragmatics of the text remains a genuine threat.

5.3 Challenges with Radicalisation Detection

Online radicalisation detection using either language-dependent or independent approaches is still challenging to perform. Most of the ‘ground truth’ datasets used in different works are not reliably verified from an accuracy perspective. Many of such datasets, e.g., [65], [66], [67] are collected using keyword sets, with users tweeting those words would be regarded as in the ‘radicalised’ set. It is also possible that users who use radicalisation terminology in their tweets may sometimes report on some event (e.g., ‘Islamic State is hacking a Swedish radio station’) or share harmless religious rhetoric (e.g., ‘If you want to talk to Allah, pray, if you want Allah to speak to you, read the Quran’).

A standard gold data set is required to train recognition models, and experts must manually check this dataset to ensure that the cases are real positives and/or real negatives. One source of manually identified radical accounts is *CtrlSec*⁷, where volunteers report ISIS propaganda on social media. This initiative claimed that they had closed more than 200,000 Twitter accounts in three years [68]. While these are critical mechanisms to encounter radicalisation online, the quick close of accounts once identified as radical means that the data cannot be further collected and analysed to train automated methods.

⁷<https://twitter.com/CtrlSec>

From a policy perspective, radicalisation is not a crime. Radicals of all religions and ideologies can freely express their beliefs and practice their freedom of expression in a democracy. However, adopting or preaching violent radicalisation is a crime [68]. Hence, considering the above-presented work, our finding is that online radicalisation detection needs a multi-pronged approach(es). Researchers need to focus on this research area and develop/propose more constructive approaches to develop the best and the most effective ways to prevent society from radicalisation.

To conclude, both language dependent and independent approaches have issues and challenges. However, language-independent approaches are still better than language dependent in a broad scope. Since language-independent systems are mainly supervised, there is a need for an unsupervised system to handle NLP issues more comprehensively for radicalisation, extremism, and collective radicalisation detection.

6 Conclusion

Social networks have a significant role in extreme ideas and radicalisation dissemination all over the world. People disseminate similar information, which can lead to radicalisation, collective radicalisation, and violent extremism. Our analysis shows that micro-blogging sites like Twitter and Tumblr are the two most common social media data sources for radicalism detection and forecasting applications. Twitter has played an essential role in facilitating political mobilisation compared to other social media platforms, as it is inherent to share short texts through direct messages, top trending, and relationships with followers. Interestingly, despite YouTube's immense popularity and proliferation as an online video-sharing website, none of the previous research was used to plan or predict protests using data from it. On the contrary, YouTube is the most widespread platform for online radicalisation, hate promotion, and extremism promotion, according to published research reports. Compared to Twitter, Tumblr, a popular micro-blogging website, has not been the main focus of research for online radicalisation detection applications.

Our analysis also reveals a variety of information retrieval and machine learning methods and techniques that researchers are using to explore solutions for detecting radicalisation online. We found that POS tagging, stemming/lemmatisation, Named Entity Recognition, and n -gram models are a few standard components in the process of several proposed approaches and techniques. Our study also shows that language-independent supervised approaches based mainly on KNN (K-Nearest Neighbor), Naive Bayes, Support Vector Machine, Rule-Based Classification, Decision Tree, Exploratory Data Analysis, and markup Keyword-Based, are the most widely used to detect online radicalisation on social media websites.

We try to establish a differentiation between two main NLP approaches, i.e., language-dependent and language-independent. This study aims to provide the state-of-art to construct an unsupervised system for collective

radicalisation and extremism detection. For this purpose, we presented an in-depth view of three different case studies for each category. This study covers two important research areas, i.e., extremism and radicalisation, to understand each of these areas better. Instead of just providing brief details of different related work based on language-dependent and independent NLP approaches dedicated to these areas (extremism and radicalism), we further discussed challenges and issues with these two approaches. We found that language-independent systems are still better than language-dependent systems.

Furthermore, this study presented a generic structure and guidelines for developing a new unsupervised language-independent system for addressing radicalism and collective radicalism issues. This study intended to cover language-dependent versus language-independent in the context of NLP tasks to develop an efficient and effective system. Hopefully, this study will also guide students and researchers with essential resources to learn foundational knowledge of the field and further integrate unsupervised and language-independent techniques with different machine learning models.

Acknowledgments. This work was supported by National Founding from the FCT-Fundação para a Ciência e a Tecnologia, through the MOVES Project - PTDC/EEI-AUT/28918/2017 and by operation Centro-01-0145-FEDER-000019-C4 - Centro de Competências em Cloud Computing, co-financed by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica.

Compliance with Ethical Standards

Funding

This work was supported by National Founding from the FCT Fundação para a Ciência e a Tecnologia, through the MOVES Project- PTDC/EEI-AUT/28918/2017, and by Operação Centro-01-0145-FEDER-000019 – C4 – Centro de Competências em Cloud Computing, co-financed by the Programa Operacional Regional do Centro (CENTRO 2020), through the Sistema de Apoio à Investigação Científica e Tecnológica – Programas Integrados de IC&DT.

Ethics approval and consent to participate

The authors declare that they have no known ethics issue that could have appeared to influence the work reported in this paper.

This article does not contain any studies with human participants or animals performed by any of the authors.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships or conflicts of interest that could have appeared to influence the work reported in this paper.

References

- [1] Tanoli, I.K., Petrocchi, M., De Nicola, R.: Towards automatic translation of social network policies into controlled natural language. In: 2018 12th International Conference on Research Challenges in Information Science (RCIS), pp. 1–12 (2018). IEEE
- [2] Williamson, L.: Gilets Jaunes: Anger of yellow vests still grips France a year on. URL: <https://www.bbc.com/news/world-europe-50424469> [accessed: 2019-09-18] (2019)
- [3] Crettiez, X.: «gilets jaunes»: la violence, l’arme des bavards, est aussi celle des silencieux. *Le Monde* (2018)
- [4] Reis, C.: Polícia não vai deixar manifestação da extrema-direita chegar à sede do bloco de esquerda. *Diário de Notícias* (2019)
- [5] Becker, H., Naaman, M., Gravano, L., et al.: Selecting quality twitter content for events. *ICWSM* **11** (2011)
- [6] Krumm, J.S.: Influence of social media on crowd behavior and the operational environment. Technical report, ARMY COMMAND AND GENERAL STAFF COLLEGE FORT LEAVENWORTH KS SCHOOL OF ... (2013)
- [7] Scanlon, J.R., Gerber, M.S.: Automatic detection of cyber-recruitment by violent extremists. *Security Informatics* **3**(1), 5 (2014)
- [8] Khan, W., Daud, A., Nasir, J.A., Amjad, T.: A survey on the state-of-the-art machine learning models in the context of nlp. *Kuwait journal of Science* **43**(4) (2016)
- [9] Agarwal, S., Sureka, A.: Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. arXiv preprint arXiv:1511.06858 (2015)
- [10] Scruton, R.: *The Palgrave Macmillan Dictionary of Political Thought*. Springer, ??? (2007)
- [11] Kruglanski, A.W., Fernandez, J.R., Factor, A.R., Szumowska, E.: Cognitive mechanisms in violent extremism. *Cognition* **188**, 116–123 (2019)

- [12] Schmid, A.P.: Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper* **97**(1), 22 (2013)
- [13] Striegheer, J.-L.: *Violent-extremism: An examination of a definitional dilemma* (2015)
- [14] Bartlett, J., Birdwell, J., King, M.: *The edge of violence: A radical approach to extremism*. Demos, 5–75 (2010)
- [15] Kruglanski, A.W., Gelfand, M.J., Bélanger, J.J., Sheveland, A., Hetiarachchi, M., Gunaratna, R.: The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology* **35**, 69–93 (2014)
- [16] Steiner, K., Önnersfors, A.: *Expressions of radicalization. Global Politics, Processes and Practices* (2018)
- [17] Moghaddam, F.M.: *Mutual Radicalization: How Groups and Nations Drive Each Other to Extremes*. American Psychological Association, ??? (2018)
- [18] Malthaner, S.: Radicalization: The evolution of an analytical paradigm. *European Journal of Sociology* **58**(3), 369–401 (2017). <https://doi.org/10.1017/S0003975617000182>
- [19] Ismail, H., Harous, S., Belkhouche, B.: A comparative analysis of machine learning classifiers for twitter sentiment analysis. *Res. Comput. Sci.* **110**, 71–83 (2016)
- [20] Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* **5**(4), 1093–1113 (2014)
- [21] Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S., Khan, K.H.: Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling* **4**(1), 2 (2016)
- [22] Shirky, C.: *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin, ??? (2008)
- [23] Statista: Number of Social Media Users Worldwide from 2010 to 2021. eMarketer, (2017). eMarketer. URL: <https://web.archive.org/web/20170711060853/https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
- [24] Statista: Number of Social Media Users Worldwide from 2017 to 2025.

- eMarketer, (2021). eMarketer. URL: <http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users>
- [25] Khalaf, S.: Mobile Use Grows 115Tutorial Point (2014). <https://www.flurry.com/post/115191226770/mobile-use-grows-115-in-2013-propelled-by>
- [26] Kemp, S.: DIGITAL 2020: 3.8 BILLION PEOPLE USE SOCIAL MEDIA. DIGITAL 2020 (2020). <https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media>
- [27] NEWS, A.: A Look at How ISIS Is Recruiting Young Americans Through the Internet, (2017). <https://www.youtube.com/watch?v=4FZC0WWzHQs>
- [28] Observador: O Estado Islâmico e a Internet: Onde e Como Eles Recrutam, (2015). <https://observador.pt/2015/11/20/o-estado-islamico-e-a-internet-onde-e-como-recruta/>
- [29] Daud, A., Khan, W., Che, D.: Urdu language processing: a survey. *Artificial Intelligence Review* **47**(3), 279–311 (2017)
- [30] Narr, S., Hulfenhaus, M., Albayrak, S.: Language-independent twitter sentiment analysis. *Knowledge discovery and machine learning (KDML)*, LWA, 12–14 (2012)
- [31] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of machine Learning research* **3**(Jan), 993–1022 (2003)
- [32] Chiong, R., Wei, W.: Named entity recognition using hybrid machine learning approach. In: 2006 5th IEEE International Conference on Cognitive Informatics, vol. 1, pp. 578–583 (2006). IEEE
- [33] Rehman, Z., Anwar, W.: A hybrid approach for urdu sentence boundary disambiguation. *Int. Arab J. Inf. Technol.* **9**(3), 250–255 (2012)
- [34] Bender, E.M.: Linguistically naïve!= language independent: Why nlp needs linguistic typology. In: *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pp. 26–32 (2009)
- [35] Lara-Cabrera, R., Pardo, A.G., Benouaret, K., Faci, N., Benslimane, D., Camacho, D.: Measuring the radicalisation risk in social networks. *IEEE Access* **5**, 10892–10900 (2017)
- [36] Camacho, D., Gilpérez-López, I., Gonzalez-Pardo, A., Ortigosa, A., Urruela, C.: Risktrack: a new approach for risk assessment of radicalisation

- based on social media data. In: CEUR Workshop Proceedings (2016)
- [37] Hung, B.W., Muramudalige, S.R., Jayasumana, A.P., Klausen, J., Libretti, R., Moloney, E., Renugopalakrishnan, P.: Recognizing radicalization indicators in text documents using human-in-the-loop information extraction and nlp techniques. In: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–7 (2019). IEEE
 - [38] Rekik, A., Jamoussi, S., Hamadou, A.B.: Violent vocabulary extraction methodology: Application to the radicalism detection on social media. In: International Conference on Computational Collective Intelligence, pp. 97–109 (2019). Springer
 - [39] Oleksyk, P., Hernes, M., Nita, B., Dudycz, H., Kozina, A., Janus, J.: Supporting investment decisions based on cognitive technology. In: Towards Industry 4.0—Current Challenges in Information Systems, pp. 41–59. Springer, ??? (2020)
 - [40] Bender, E.M.: On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology* **6**(3), 1–26 (2011)
 - [41] Araque, O., Iglesias, C.A.: An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access* **8**, 17877–17891 (2020)
 - [42] Bermingham, A., Conway, M., McInerney, L., O’Hare, N., Smeaton, A.F.: Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In: 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 231–236 (2009). IEEE
 - [43] Nouh, M., Nurse, R.J., Goldsmith, M.: Understanding the radical mind: Identifying signals to detect extremist content on twitter. In: 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 98–103 (2019). IEEE
 - [44] Kušen, E., Cascavilla, G., Figl, K., Conti, M., Strembeck, M.: Identifying emotions in social media: comparison of word-emotion lexicons. In: 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 132–137 (2017). IEEE
 - [45] Bilmes, J., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers, pp. 4–6 (2003)
 - [46] Allison, B., Guthrie, D., Guthrie, L.: Another look at the data sparsity problem. In: International Conference on Text, Speech and Dialogue, pp.

327–334 (2006). Springer

- [47] Derbas, N., Segond, F., Padró, M., Dusserre, E., Dobre, T., Monaci, S., Mastrobuoni, G.: Semantic analysis supporting de-radicalisation. In: 2017 28th International Workshop on Database and Expert Systems Applications (DEXA), pp. 17–21 (2017). IEEE
- [48] Khalaf, S.: Open problems with NLP. Flurry (2014). <https://medium.com/sciforce/biggest-open-problems-in-natural-language-processing-7eb101ccfc9>
- [49] Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* **7**, 597–610 (2019)
- [50] Aguilar, G., Maharjan, S., López-Monroy, A.P., Solorio, T.: A multi-task approach for named entity recognition in social media data. *arXiv preprint arXiv:1906.04135* (2019)
- [51] Leaman, R., Gonzalez, G.: Banner: an executable survey of advances in biomedical named entity recognition. In: *Biocomputing 2008*, pp. 652–663. World Scientific, ??? (2008)
- [52] Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M., Xu, H.: A comprehensive study of named entity recognition in chinese clinical text. *Journal of the American Medical Informatics Association* **21**(5), 808–814 (2014)
- [53] Shaalan, K.: A survey of arabic named entity recognition and classification. *Computational Linguistics* **40**(2), 469–510 (2014)
- [54] Etaiwi, W., Awajan, A., Suleiman, D.: Statistical arabic name entity recognition approaches: A survey. *Procedia computer science* **113**, 57–64 (2017)
- [55] Patil, N., Patil, A.S., Pawar, B.: Survey of named entity recognition systems with respect to indian and foreign languages. *International Journal of Computer Applications* **134**(16) (2016)
- [56] Yıldız, T., Yıldırım, S., Diri, B.: An integrated approach to automatic synonym detection in turkish corpus. In: *International Conference on Natural Language Processing*, pp. 116–127 (2014). Springer
- [57] Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41 (1995)
- [58] Iqbal, K., Khan, M.Y., Wasi, S., Mahboob, S., Ahmed, T.: On extraction of event information from social text streams: An unpretentious nlp

- solution. *IJCSNS* **19**(9), 1 (2019)
- [59] Jusoh, S.: A study on nlp applications and ambiguity problems. *Journal of Theoretical and Applied Information Technology* **96**, 1486–1499 (2018)
 - [60] Jiampojarn, S., Cherry, C., Kondrak, G.: Joint processing and discriminative training for letter-to-phoneme conversion. In: *Proceedings of ACL-08: HLT*, pp. 905–913 (2008)
 - [61] Bartlett, S., Kondrak, G., Cherry, C.: Automatic syllabification with structured svms for letter-to-phoneme conversion. In: *Proceedings of ACL-08: HLT*, pp. 568–576 (2008)
 - [62] Alfawareh, H.M., Jusoh, S.: Resolving ambiguous entity through context knowledge and fuzzy approach. *International journal on computer science and engineering (IJCSE)* **3**(1), 410–422 (2011)
 - [63] Wang, Y., Wang, M., Fujita, H.: Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems* **190**, 105030 (2020)
 - [64] Point, T.: NLP - Word Sense Disambiguation. Tutorial Point (2020). https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_word_sense_disambiguation.htm
 - [65] Agarwal, S., Sureka, A.: Using knn and svm based one-class classifier for detecting online radicalization on twitter. In: *International Conference on Distributed Computing and Internet Technology*, pp. 431–442 (2015). Springer
 - [66] Rowe, M., Saif, H.: Mining pro-isis radicalisation signals from social media users. In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pp. 329–338 (2016)
 - [67] Ashcroft, M., Fisher, A., Kaati, L., Omer, E., Prucha, N.: Detecting jihadist messages on twitter. In: *2015 European Intelligence and Security Informatics Conference*, pp. 161–164 (2015). IEEE
 - [68] Fernandez, M., Asif, M., Alani, H.: Understanding the roots of radicalisation on twitter. In: *Proceedings of the 10th ACM Conference on Web Science*, pp. 1–10 (2018)