Language-Independent Approaches to Detect Extremism and Collective Radicalisation Online

Sebastião Pais
Computer Science Department
NOVA LINCS and UBI
Covilhã, Portugal
Email:sebastiao@di.ubi.pt

Irfan Khan Tanoli
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:irfan.khan.tanoli@ubi.pt

Miguel Albardeiro
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:miguel.albardeiro@ubi.pt

João Cordeiro
Computer Science Department
University of Beira Interior
Covilhã, Portugal
Email:jpaulo@di.ubi.pt

Abstract—**Due to lack of regulation, a lot of user-generated content reflects more closely the offline world than official news sources. Social media have become attractive platforms for anyone seeking independent information. Text mining and knowledge extraction are also crucial issues, in particular, directed toward social media and micro-blogging. The automatic identification of extremism and collective radicalisation require sophisticated Natural Language Processing (NLP) methods, text mining techniques, and resources, especially those dealing with opinions, emotions, or sentiment analysis. The area of understanding and detecting extremism and collective radicalism on social media has a connection with sentiment analysis and opinion mining. The main focus of this work is to provide the state-or-art to identify extremism and collective radicalisation on social networks based on user's sentiment analysis, and to develop an unsupervised and language-independent approach by relying on statistical and probabilistic methods. This paper discusses few important case studies related to the roots of radicalism, extremism detection, and terrorism detection using sentiment analysis and present machine learning models, and how these methodologies can be exploited to develop our desire system.**

Keywords–Natural Language Processing; Social Media; Extremism; Collective Radicalisation; Sentiment Analysis

I. INTRODUCTION

In the last few years, the advent of micro-blogging services has been impacting people's mind, communication, behavior, and activities conduct. It is due to several factors, including the use of convenience, and the lack of regulation, and the vast amounts of user-generated contents that reflect more closely the offline world than the official news source. Social media and network have become an attractive platform for anyone seeking independent information and eventually, more authentic news. Recently, we have assisted the news about the 'Yellow vests' or in French 'Gilets Jaunes' [1]. It began as a pacific manifestation, but later few extremist groups have joined the manifestations that made it a violent protest as was in the news: 'Absence of the progress of the movement, inexperience of the demonstrators, the action of extremist groups, the forces of higher duties' [2]. In Portugal, we have witnessed some radical events as few people were protesting against the actions taken by the police on a tough neighborhood referred as 'Bairro da Jamaica'; there were few from an extremist group protesting against the politicians, violently [3].

In social networks such as Facebook, Twitter, and Youtube, each cluster of posts, videos or tweets focus on a burst topic that may constitute a potential threat. However, the majority of clusters are harmless and represent casual, conventional or expressive crowds as well as noisy data [4]. To identify acting or protesting crowds on social networks, it is necessary to understand the tone of language usage, e.g., slang, abusive, jargon, formal, respectful etc., present in each cluster as well as its network activity. Ultimately, a crowd is characterized by its dominant emotions; it is the level of interaction and shared contents. The work in [5] discussed the technique that can be used to analyze the tweeter contents and detect the event related to the contents.

Users use social networks for various purposes. Unfortunately, few use it to spread distorted beliefs, negative opinion about things like spreading terrorism, extremism, and radicalism [6]. Since mid-2015 Twitter has already deleted more than 125,000 accounts that were somehow linked to terrorism [7]. Researchers focus Twitter for sentiment analysis due to few particular reasons: Twitter's popularity as enormous numbers of people continuously tweet on Twitter related to various topics. These topics could be political, about sports, religious, marketing, people's opinions or friend's conversations. Being an updated huge repository of facts, opinions banter and other minutiae, Twitter has received significant attention from business leaders, decision-makers, and politicians.

In this study, we aim to provide a theoretical review related to extremism and collective radicalisation detection. Extremism is a vague term that can be undermined in three different contexts [8]: Taking a political idea to its limits, regardless of unfortunate repercussions, impracticalities, arguments, and feelings to the contrary, and with the intention not only to confront but also to eliminate opposition; intolerance towards all views other than individual own; adoption of means to political ends which disregard accepted standards of conduct, in particular, which show disregard for the life, liberty and human rights of others. radicalisation is a process by which an individual or group comes to adopt increasingly extreme political, social, or religious ideals and aspirations.

With our understanding, it is clear that extremism and collective radicalisation has a direct connection with people's sentiments and opinions. There are many barriers to understand extremism and collective radicalisation on the social network. Among these challenges, one big challenge is to differentiate between the users commanding this process and the users talking about it. Hence, the main goal of this study is to propose an effective system to detect extremism and collective radicalisation on social media based on sentiment analysis. To do so, our focus is on statistical and probabilistic methods that can be used to develop an unsupervised and language-independent system.

The main contribution of this study is significant for many reasons. First, it covers three different research areas, i.e., extremism, collective radicalisation, and sentiment analysis, and to provide a better understanding related to these areas. Second, instead of just providing brief details of different works for these areas, we analyzed three essential case studies in-depth to help readers understand different approaches that have been used for these fields. This angle could also help the researchers who are familiar with specific techniques dedicated to extremism and collective radicalisation, to exploit and choose the appropriate one for their work. Third, this study also present supervised, unsupervised, and language-independent approaches proposed for extremism, collective radicalisation, and sentiment analysis with brief details of the algorithms and their originating references. This can help us to develop an efficient unsupervised and language-independent system for extremism and collective radicalisation detection. Finally, the survey is enhanced with models related to everyday NLP tasks, and we discuss which one can be exploited for our desire system.

In the following sections, we deeply review three different works for extremism, radicalisation, and sentiment analysis detection. First, we present a work that discusses the roots of radicalism. Next, we analyze a work proposed for terrorism detection based on sentiment analysis. Finally, another work for sentiment detection on Twitter using hashtags. In section III, we present a few proposed methodology. Section IV overview few standard Machine Learning (ML) models. In the end, we provide a conclusion and a future direction for our ongoing work.

II. Machine Learning Classifiers for Extremism and Collective radicalisation

radicalisation involves a movement towards the support or representation of radical behavior(s). Radical behavior can be viewed as 'when it serves a specific purpose; it undermines other goals that are important to most people' [9]. At the same time, collective radicalisation is defined as a collective intergroup process. People are not radicalized on their own, but rather as part of a group and through the socially constructed reality of their group by gathering people on the streets to show their motive and protests against specific entity or entities. However, sometimes these protests become violent.

A. The roots of radicalism

Fernandez et Al [10] proposed an innovative NLP and Collaborative Filtering (CF) based approach for detecting radicalisation on social networks, The different roots of radicalisation, i.e., micro-roots, meso-roots, and macro-roots, are captured [11], and each user is represented through keyword-based vector description. The approach presented in [10], is sufficient enough to detect and predict radicalism. On social networks, the user either creates or posts the contents or shares other people's contents; the authors assumed that micro-roots or meso-roots are captured from the user's shared or created contents. While macro-roots are captured that are external to the given social network (links/URLs) and from other websites or other social networks, and videos, etc. [10].

In [10], the authors used keyword-based vectors that include the user's post(s). These vectors represent micro-roots and meso-roots influences over users, and they are transformed into n-grams (uni-grams, bi-grams, and tri-grams) [10]. Next, the value of each n-gram in the micro-root user's vector is computed as the frequency of the n-gram in the user's post, and normalized by the number of posts. In the macro-roots influences case, an automatic data scrapping over the URLs included on a macro-roots vector is performed by automatically parsing the HTML, and extracting the title and description of the websites. Giving the set of n-grams obtained after pre-processing, all the links defined the macro-roots of the user. The value of each word in a macro-roots of user's vector is computed as the frequency of the n-gram in all user's share URL entries and normalized the number of URL [10].

The authors in [10] further collected and integrated existing lexicons, i.e., ICT Glossary, Saffron Experts, Saffron Dabiq Magazines, Rowe, and Saif to create a single lexicon containing a more comprehensive set of terms and expression that shows radicalisation terminology. To mitigate lexicon merging issues, the authors first remove incorporated syntactic variances of each term, i.e., lowercase, uppercase, apostrophes and hyphens removal, diacritics removal. Then, if two terms are present in both lexicons, they are merged and added as one unique entry in the final lexicon. The final lexicon comprises 305 entries, including expression, terms, and variances [10].

The authors in [10] also compute the radicalisation influence of different roots over the user determining cosine similarity between the micro-roots and the meso-roots vectors and the generated lexicon. It is not possible to compute cosine similarity for macro vectors due to many sites were already disabled, and it was not possible to collect URLs information. Next Collaborative Filtering (CF) strategies are used to develop an automatic prediction about user's interests by collecting numerous user's preference information, using the following two steps: Search for such users that have a similar rating pattern to other users for whom the prediction is made; use the ratings of user found in the previous step to compute the predictions for the active user. Two publicly available datasets from the Kaggle Data Science Community are used to study radicalisation. One of the datasets contains 17,350 tweets from 112 pro-ISIS accounts. The second dataset is created as the opposite of the previous one. It contains 122,000 tweets from 95,725 users collected on different days.

The corresponding results in [10] show the effectiveness of the proposed algorithms for detecting and predicting the influence of radicalisation with up to 0.9 F-1 of the measurement for detection and between 0.7 and 0.8 precision is obtained for the prediction. The work concluded as the presentation of a computational approach to the detection and prediction of the influence of radicalisation to which a user is exposed, based on the concept of 'roots of radicalisation' identified in social science models.

Detecting radicalisation online faces several challenges. From an accuracy perspective, most of the 'ground truth' datasets used in different works are not reliably verified. Many of these datasets, e.g., [12], [13], [14], are collected using keyword sets, with users tweeting those words would be regarded as in the 'radicalized' 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').

There is still a need to use a gold standard dataset to train recognition models. This dataset must be manually checked by experts to ensure that the cases are real positives and/or real negatives not false positive and/or false negative. One source of manually identified radical accounts is Ctrl-Sec [15], where volunteers report ISIS propaganda on social media. This initiative claimed to have closed more than 200,000 Twitter accounts in three years [10]. While these are critical mechanisms to encounter radicalisation online, still the accounts are closed quickly once identified as radical means that the data cannot be further collected and analyzed to train automated methods.

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. However, adopting or preaching violent radicalisation is a crime [10]. Therefore, considering the above-presented work, our finding is that online radicalisation detection needs a multi-pronged approach(es). Researchers need to focus this research area and developed/proposed more constructive approaches to come up with the best and the most effective ones to prevent society from radicalisation.

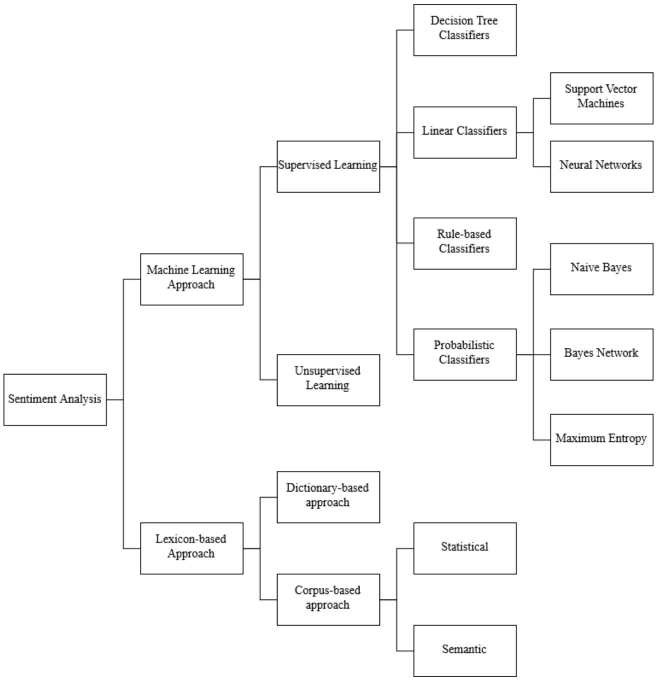

Figure 1. Sentiment Classification Techniques Used in SA [16]

B. Sentiment Analysis

The sentiment, polarity, and opinion mining or sentiment analysis deal with direction-based text analysis, i.e., text with opinions and emotions [17]. '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, event, or topic [16]. Opinion Mining (OM) is not the same as Sentiment Analysis(SA). OM starts by extracting and analysing the opinion about something while SA is more about the sentiment that something causes on people, usually expressed in the text, like or share tweets or Facebook posts [16]. SA can also be observed as a type of text classification but deals with subjective statements that are harder to classify [18].

SA can be viewed as a classification process. It can be divided into three 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. SA is an expressed feeling classified in each sentence at the sentence level. First of all, it must be determined whether the sentence is subjective or objective. If the sentence is subjective, SA determines the polarity of the opinion (positive or negative) [16]. However, using these two levels does not provide the necessary details on all aspects of the entity that are needed in many applications. The aspect-level classifies, taking into account the specific aspects of the entities. Firstly, it is required to identify the entities and their aspects. The opinion holders can give different opinions on different aspects of the same entity [16].

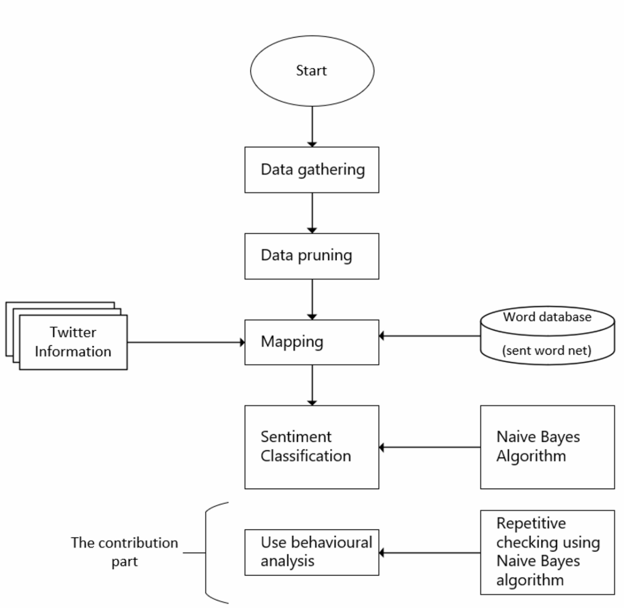

Figure 2. Proposed system development diagram [6]

C. Terrorism Detection using Sentiment Analysis and Machine Learning

In this section, we review an approach based on Sentiment Analysis (SA) for detecting terrorism on social networks. According to [6], social networks have recently been the most crucial channel for people to interact and share ideas. People choose to express their opinion(s) on a particular subject, news, or event due to the rapid spread of information on social media. For example, it is easier to reach more people online and influences the choices of potential users about the top trending topic on Twitter. Contrary, it is also easy for extremist groups and its members to recruit the people sharing the same ideology and views on social media and networks. In 2015 more the 250,000 accounts were linked to terrorism and later the accounts have been deleted and disabled [7].

Existing SA approaches aim to find a tweet that may or may not lead to an extremist user. These approaches are still not practical enough for specific reasons like ambiguity in tweets, synonymy in tweets, use of emotions in tweets, etc. It is also quite familiar for humorists to make fun of people or even joke about terrorism on Twitter just for fun. A big challenge for existing approaches is to classify if a tweet is a real threat or not. The two most general approaches to SA are the lexical and Machine Learning (ML) approach [17]. These two approaches are further sub-classified into more approaches as shown in Figure 1.

The main objective of the proposed work in [6] is to present a system for improvising current techniques for SA through ML to detect terrorist acts on Twitter more accurately. The general structure of the system is shown in Figure 2. The novelty of this research is having divided the sentence into positive, negative, and neutral categories. Then all three categories are compared to the previous sentence for a given account holder based on the sentiment score for the latest and previous sentence. This means a specific account holder's tweet history in each of the categories is extracted, and the sentiment value is calculated. Later, the sentiment score from the above statement will be compared with the sentiment value of the latest identified sentiment. The system is based upon the ML approach rather lexical-based [6]. For better understanding, the functions of each component are presented below.

- **Data gathering:** The target source for data collection is Twitter due to its popularity, and even it is used for communication about terrorism. Even compared to Facebook and other popular blogs, Twitter has recorded more significant problems related to acts of terrorism. The data are gathered from the Twitter streaming API. For this work [6], the authors used keywords e.g., ISIS, Bomb, etc. to obtain data related to terrorism. If tweet(s) match(es) the user's criteria directly, these tweets are sent directly to the user in JSON format, a JavaScript object notation.
- **Data pruning:** After data collection, it is preprocessed for normalization. Removal of URLs, @tags, hashtags, uppercase and lowercase letters, misspellings, etc. are some examples of data pruning.
- **Mapping:** SentiWordNet [19] is used as a dataset for mapping. It is made up of thousands of English words that have a positive or negative score for each word. Tweets are compared and computed with SentiWordNet. Since the word alone is not enough to make a decision, the total score is calculated based on the sentence context.
- **Sentiment Classification:** Twitter sentences are classified into positive, negative, or neutral class for Sentiment Classification. Naïve Bayes is used because it is commonly used for SA. Bayes theorem is used to predict the probability that a given set of features will belong to a particular label. The Naïve Bayes classifies the statement as positive, negative, or neutral based on the result of the sentiment assessment.
- **User Behavioral Analysis:** User Behavioral Analysis is carried out using Snapbird tool [6] to track the previous tweets of a particular user. When tweets become classified to their polarity based on the sentiment score, all three classes are checked repetitively. For double-checking, tweets on each category are compared with tweets history. The purpose of this repetitive checking is to find better results on the understanding if tweets are leading towards terrorism or not [6].

If the score is negative after re-checking and results in the same class, it can be concluded that the account holder may lead to acts of terrorism. The purpose of reviewing the user's previous tweets is to analyze the user's tweet patterns. As mentioned above, the user can be a humorist or just joke about terrorism, so the pattern of user tweets can be related to jokes. However, if the user seriously discussing to support terrorism and wanted to convince or influence other readers about terrorism support, then that user is categorized in the terrorist category [6]. The use of Naïve Bayes has been proven and had the potential to be implemented [6]. Hence, Bayes theorem is being applied to predict a class for any giver text from tweets. The authors [6] applies Bayes theorem to predict the class of any tweet using the Equation 1.

$$P(label|features) = \frac{P(label)P(features|label)}{P(features)} \tag{1}$$

Where *P(label)* is the class (i.e., positive/negative/neutral) of the tweets while *P(features)* is the tweet. *P(label—features)* is the result of the application of the techapprichnique. By using 1, we get 2:

$$P(positive|tweet) = \frac{P(positive)P(tweet|positive)}{/P(tweet)} \tag{2}$$

The process has to be repeated for all three categories (positive/negative/neutral). Finally, the highest-ranked class is chosen to label the document [6]. The initial results show that there are more than 50 words indicated as terrorism keywords, e.g., jihad, bomb, radical Al-Qaeda etc. Among the top eight words in the list are terrorism, jihad, bomb, radical, Abu Sayyaf, ISIS and extremist [6].

To conclude it, Naïve Bayes algorithm-based system is proposed for terrorism detection on Twitter. Naive Bayes approach appears as a medium accuracy comparing with support vector machine and neural network. To further enhance the accuracy of Naive Bayes, the element of user behavioral analysis has been proposed to embed into the algorithm after sentiment classification process have been performed.

D. Learning with Hashtags

Here we present an interesting supervised approach [20] based upon for learning hashtags, hashtag patterns, and phrases associated with five emotions: AFFECTION, ANGER/RAGE, FEAR/ANXIETY, JOY, and SADNESS/DISAPPOINTMENT. It is usual for users to express an emotional state using hashtags (e.g., **#inlove, #hatemylife**) on Twitter. Few hashtags consist of a single word like **#Faith**, others composed of multiple words (e.g., **#FaithinhumanityRestored**), or it can even be a creative spelling, e.g.,**sk8** or **cantwait4tmrw**. As these hashtags are continuously created through various infinite open combinations, it is not easy to identify such hashtags using sentiments or emotions lexicons.

In [20], a bootstrapping framework for learning emotion hashtags is proposed as described above, which has been further improved to learn more general hashtag patterns. The emotion phrases are extracted from the hashtags and the hashtag patterns to classify contextual emotions. The first step is to find the common prefix in the hashtags. For example, **#Angryatlife** and **#Angryattheworld** have the same prefix

angry at that predicts ANGER emotion. As a result, certain hashtags generalize into hashtag patterns that match hashtag with the same prefix. A critical challenge here is to identify the same prefixes in hashtags with different emotions that could lead to incorrect emotion. For example, #anger pattern generally points out angry tweets. However, hashtag as #angrybirds refers to a game, not the emotion of the writer. AFFECTION can be determined as 'I love you' followed by the person (e.g., #loveyoufather). This can be related to JOY in other contexts (e.g., #loveyoulife). The authors use the probability estimates to determine certain hashtag patterns that are reliable indicators to an emotion [20]. Also, if there is a negation, it can toggle the polarity of the tweet (e.g., **not love life** can suggest SADNESS instead of JOY).

1) Learning Hashtags: The authors used and collapsed parrot emotion taxonomy [20] into only five emotions that occur more often in tweets and are easily distinguishable from each other, i.e., AFFECTION; ANGER/RAGE; FEAR/ANXIETY; JOY SADNESS/DISAPPOINTMENT. Adding another class, 'None of the above' that does not express any emotion. For each one of the five classes, the five common identified hashtags are strongly associated with the emotion and these hashtags are used as seeds.

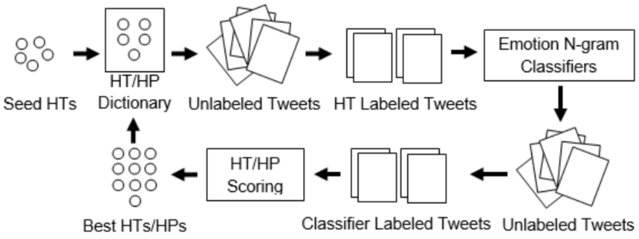

Figure 3. Bootstrapped Learning (HT = hashtag; HP = hashtag pattern) [20]

The general architecture of the framework is shown in Figure 3. The process starts with tweets containing the seed hashtags and marks with the appropriate emotion. There are 323,000 tweets received from at least one of the seed hashtags. Additionally, more than 2.3 million untagged tweets are collected using Twitter's streaming API that contains at least one hashtag (an average of 1.29 hashtags per tweet and 3.95 tweets per hashtag). Tweets are preprocessed with the CMU tokenizer and normalized against the case. The tagged tweet is then used to train a series of emotion classifiers. A logistic regression classifier is trained for each emotion class [20].

Each emotion classifier applies to unlabeled tweets. For each emotion e, the obtained tweets are classified as e, and the hashtags are extracted from such tweets to create a candidate pool of hashtags H_e for that emotion e. Next, the candidate hashtag h is assigned a score by calculating the average probability of the same emotion e obtained from the logistic regression classifier for the entire tweet containing the candidate hashtag h. From the untagged tweets, all tweets with one of the learned hashtags are then added to the training instance, and the process continues. To reduce the number of potential candidates, hashtags that appear less than ten times, those with a single character, and those that appear more than twenty times are discarded.

2) Learning Hashtag Patterns: In this phase, the hashtag is expanded into a sequence of words using an N-gram based word segmentation algorithm [21]. The Prefix Tree data structure is used for the representation of all possible prefixes of the expanded hashtag. Then, the tree is traversed, and all possible prefixes are considered as candidates of hashtag patterns. Later, each pattern is assigned a score as the way it is done with hashtags. The authors calculate the average probability of classifier, and for each emotion class, ten hashtag patterns with the highest scores are selected. For unlabeled tweets, all tweets with hashtags are added, which match one of the learned hashtag patterns to the training instances, and the bootstrapping process continues.

3) Creating Phrase-based Classifiers: The final type of emotion classifier aims to acquire is emotion phrases. Right at the end of the bootstrapping process, the word segmentation algorithm is applied to all hashtags and hashtag patterns to separate them into phrases (e.g., #lovemylife → 'love my life'). It is assumed that the obtained phrase has the same emotion as the original hashtag. Nevertheless, it will have low precision due to the presence of a phrase yields, and the surrounding context must also be taken into account [20]. Finally, a logistic regression classifier is trained for each emotion that classifies a tweet about its emotion based on the presence of learned phrases for the emotion, as well as a context window of size six around the obtained phrase, three for each side of the phrase.

The results in [20] show that the learned set of emotional indicators causes a substantial improvement in F-scores, ranging from + % 5 to + % 18 to basic classifiers, The result also showed that the combination of the emotion indicators learned with an N-gram classifier in a hybrid approach significantly improves performance in 5 emotion classes. This work [20] proposed three types of emotional indicators. The approach is categorized as weakly supervised monitored bootstrapping: hashtags, hashtag patterns, and phrases. Once the emotion indicators are trained using hashtags, and the hashtags can gain form in any language. Moreover, these indicators can also be applied to any language as a language-independent method since it does not depend on a specific corpus.

III. Supervised, Unsupervised and Language-Independent Approaches

In the previous Section II, we discuss three different studies in detail for extremism, collective radicalisation detection and detecting sentiment on Twitter using hashtags. The purpose of this study to create a state-of-the-art that aims to understand extremism and collective radicalisation, sentiment analysis, and to develop an unsupervised and language-independent system using machine learning models. To achieve it, we will rely on probabilistic approaches that can be applied to any language, or even in a mix of languages. On the unsupervised part, we aim to create a system that can detect extremist or radical tweets by itself without much human intervention. In this section, we overview a few NLP approach that we can use to achieve our desire goal.

A. Supervised Natural Language Processing Approach

Supervised machine learning involves labeling or commenting on a series of text documents with examples of what the machine is looking for and how it should interpret this aspect. Researchers use datasets to train a statistical model that is then given unlabeled text for analysis. Later, more extensive or better datasets can be used to retrain the model as it learns more about the documents it analyzes. For example, one can deploy a supervised learning system to train a model on Twitter tweets and then use it for various purposes. Several methods have heretofore been used in a supervised approach, e.g., Support Vector Machines, Bayesian Networks, Maximum Entropy Conditional Random Field, Neural Networks/Deep Learning. An interesting supervised approach based on word embedding for the sentiment classification of Twitter is shown in [22]. We aim to develop a similar approach to detecting extremism and collective radicalisation based on sentiment analysis (SA). For SA, it is essential to examine them at different levels for extremism and radicalisation detection. A user may be talking on a topic that represents extremism but is not an extremist. For example, as ISIS became more active on social networks, some accounts unrelated to extremism groups were temporarily deleted from Twitter. Hence, it is essential to identify an extremist person or someone who is involved in a radicalisation process.

B. Unsupervised Natural Language Processing Approach

An unsupervised approach refers to a system where training inputs are not necessary to discover the target point of the learning. The system needs to train itself without human supervision and intervention or with human intervention only if there is a need to add or change the functionalities. Topic modelling is another core task of NLP. Let say you have a bunch of books; you want to categorize them according to (of course) the topics they talk about it, how you solve the challenge without reading all the books. An unsupervised approach discussed in [23], which uses matrix factorization to extract latent (or hidden) topics from the text; this approach is unsupervised as there is no model trained and tested, one just set the parameters in a trial and error to achieve the best results. The work discussed in subsection II-D, a weekly supervised system, and a supervised language-independent probabilistic is proposed in [24] for twitter sentiment analysis. Therefore, our focus is to use probabilistic methods to develop an unsupervised language-independent system for user's sentiment analysis.

C. Language-Independent Approach

A language-independent approach refers to a single system that is applied to different natural languages (e.g., English, Chinese, German, etc.), the results keep being satisfactory and the experiment values represent the reality in a viable way. Once an algorithm has been developed for a specific language, the question arises, it can be trivially extended to another language; All it is needed an adequate amount of training data for the new language. It is a virtue. However, the typical approach to developing language-independent systems is to avoid using any particular linguistic knowledge in their development. The approaches presented in [25][26][27] are a few examples of such an approach.

Hence, we aim to propose a language-independent system beside unsupervised. For example, if we collect tweets from the streaming API, having as a criterion the geo-localization as Portugal, the tweet would not only in Portuguese but probably on other languages. Therefore, we would like our system to analyse the tweet regarding the language it is written.

IV. Models

The current known models for solving NLP problems are based on Supervised Machine Learning (SML). The basic idea behind SML models is to follow automatically induces rules from training data. The most common ML models commonly used to resolve ambiguities in language knowledge with the main tasks of NLP are Hidden Markov Model (HMM), Conditional Random Fields (CRF), Maximum Entropy (MaxEnt), Support Vector Machines (SVM), Decision Trees (DT), Naïve Bays (NB) and Deep Learning (DL) [28]. Apart, the following models also explain possible ML techniques in [29] that can also be considered for the development of our desire system:

Naïve Bayesian: Naïve Bayesian (NB) classifier is constructed using Bayes' theorem with assumptions of independence between predictors. A NB model is easy to develop without complicated iterative parameter estimation, which makes it particularly useful for huge big datasets. Despite its simplicity, the NB classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods [29]. The classifier is stated as:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)} \tag{3}$$

Where $P(A)$ is the prior probability of B, $P(A|B)$ is the conditional probability of A, given B called the posterior probability, $P(B|A)$ is the conditional probability of B given A and $P(B)$ is the prior probability of $B4$.

The NB classifier is based on the assumption of conditional class independence. If conditional class independence is assumed, the effect of an attribute value on a particular class is independent other attributes values [28]. The contribution of Naïve Bayes technique in computational linguistic is minimal. Recently, few research works reported based on Naïve Bayes technique for NLP tasks are [30][31][32] respectively.

Neural Networks: The biological neurons of brain structures inspire Neural Networks (NNs). Individual neuron models can be combined into several networks made up of many individual nodes, each with their variables. These networks have an input layer, an output layer, and one or more hidden layers. Hidden levels provide connectivity between entrances and exits. The network can also receive feedback using the result variables as input to the pre-processing nodes [29]. A network of interconnected functional elements, each with several inputs/one output as specified in equation 4:

$$y(x_1, \ldots, x_n) = f(w_1x_1 + w_2x_2 + \ldots + w_nx_n) \tag{4}$$

w_n, x_n are parameters of equation, f is the activation function of equation 4, crucial for learning that *addition* is used for integrating the inputs.

K-Nearest Neighbor: In the K-Nearest Neighbor (KNN) model, there is no learning phase as the training set is used every time a classification is performed. The NN search, also known as an approximate search, similarity search, or closest point search, is an optimization problem to find the closest points in metric spaces. K nearest neighbor is used to simulate daily precipitation and other meteorological variables [29].

Decision Trees: The Decision Tree (DT) is one of the standard classification algorithms currently used in ML. The DT is a new field of ML that involves the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees, and discrimination networks or production rules [29].

Each of these models has its pros and cons. The NB model is quick to train and classify, but also it is assumed to be independent of features approach [29]. For NN, they are not sensitive to irrelevant properties in contrast to NB. NNs are manufactured as specialized hardware systems. This is also advantageous for network learning. On the contrary, this is too large a black box technique, and it is not probabilistic [29].

KNN is an appropriate model once we collected data from Twitter. Since the data can be quite noisy, this model is robust for noisy training data, even for a large amount of training data. On the other hand, the KNN value needs to be determined, which is not easy to identify. Furthermore, it has a high computation cost [29]. Finally, the TD that offers an easy way to understand and interpret calculations, and it can always be used with other decision techniques. As mentioned before, these techniques are supervised, but this can be an initial point, and to exploit one of the model to develop an unsupervised system.

Native unsupervised approaches generally *Lexicon based, Dictionary-based, and Corpus-based approaches*. *Lexicon-based approaches* use insights obtained on the ground of words polarity composing a sentence. With this approach, one can create a categorical polarity (Positive, Neutral, Negative), or one can calculate a score [33]. Two most famous lexicons are: Sentiwordnet [34] and SenticNet [35]. *Dictionary-based approaches* follow two main steps: a small amount of manually collected opinion words with known instructions; expand this set by searching the WordNet dictionary [36] for synonyms and antonyms. The newly found words are added to the seed list, and the next iteration starts. The iterative process stops when no more new words are found [37]. *Corpus-based approaches* rely on syntactic patterns in large corpora. This approach can generate opinion words with relatively high accuracy. Most of the corpus-based approaches need extensive labeled training data. This approach has a significant advantage compared to dictionary-based approaches as it can help find domain-specific opinion words and their orientations [38].

V. Conclusion

Social media have a significant role in the process of extreme ideas dissemination all over the world. People have the dissemination of similar information, which can lead to collective radicalisation and extremism. In this study, we discussed three different research areas, i.e., collective radicalisation, extremism, and sentiment analysis, and analysed three different approaches for their detection. Our aim of this study was to provide the state-of-art to construct an unsupervised and language-independent system for collective radicalisation and extremism detection using SA. To do this, we also presented a few supervised NLP models and discussed lexicon, dictionary and corpus based approaches that can be integrated to achieve this goal. The area of extremism and/or radicalisation does not have much previous work. However, there are few works based on SA classification. With our knowledge, this study is the first attempt to provide a depth review related to these research areas.

Furthermore, this study paper gave a generic structure and guidelines for developing a new unsupervised language independent-system for addressing radicalisation and extremism issue. This study intended to cover supervised, unsupervised and language-independent techniques 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, both to learn what is necessary to know and to advance further the integration of supervised and language-independent techniques with different machine learning models.

Acknowledgment

This work was supported by National Founding from the FCT- Fundacão 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.

References

[1] L. Williamson, “Gilets jaunes: Anger of yellow vests still grips france a year on,” 2019, URL: https://www.bbc.com/news/world-europe-50424469#:~:text=But%20three%2Dquarters%20of%20French,sausages%20over%20a%20small%20fire. [accessed: 2019-09-18].

[2] X. Crettiez, “”gilets jaunes”: la violence, l’arme des bavards, est aussi celle des silencieux”,” 2018, URL: https://www.bbc.com/news/world-europe-50424469#:~:text=But%20three%2Dquarters%20of%20French,sausages%20over%20a%20small%20fire. [accessed: 2018-12-04].

[3] C. Reis, “Polícia não vai deixar manifestação da extrema-direita chegar à sede do bloco de esquerda,” 2019, URL: www.dn.pt/pais/policia-nao-vai-deixar-manifestacao-da-extrema-direita-chegar-a-sede-do-bloco-de-esquerda-10487434.html [accessed: 2019-01-25].

[4] H. Becker, M. Naaman, L. Gravano et al., “Selecting quality twitter content for events.” ICWSM, vol. 11, 2011, pp. 442–445.

[5] J. S. Krumm, “Influence of social media on crowd behavior and the operational environment,” Army Command and General Staff College Fort Leavenworth Ks School of Advanced Military Studies, Tech. Rep., 2013.

[6] B. S. Iskandar, “Terrorism detection based on sentiment analysis using machine learning,” Journal of Engineering and Applied Sciences, vol. 12, no. 3, 2017, pp. 691–698.

[7] D. Yadron, “Twitter deletes 125,000 isis accounts and expands anti-terror teams,” 2016, URL: https://www.theguardian.com/technology/2016/feb/05/twitter-deletes-isis-accounts-terrorism-online [accessed: 2016-02-05].

[8] R. Scruton, The Palgrave Macmillan dictionary of political thought. Springer, 2007.

[9] A. W. Kruglanski, M. J. Gelfand, J. J. Bélanger, A. Sheveland, M. Hetiarachchi, and R. Gunaratna, “The psychology of radicalization and deradicalization: How significance quest impacts violent extremism,” Political Psychology, vol. 35, 2014, pp. 69–93.

[10] M. Fernandez, M. Asif, and H. Alani, “Understanding the roots of radicalisation on twitter,” in Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 1–10.

[11] A. P. Schmid, “Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review,” ICCT Research Paper, vol. 97, no. 1, 2013, p. 22.

[12] S. Agarwal and A. Sureka, “Using knn and svm based one-class classifier for detecting online radicalization on twitter,” in International Conference on Distributed Computing and Internet Technology. Springer, 2015, pp. 431–442.

[13] M. Rowe and H. Saif, “Mining pro-isis radicalisation signals from social media users,” in Proceedings of the tenth international AAAI conference on web and social media (ICWSM 2016), 2016, pp. 329–338.

[14] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, “Detecting jihadist messages on twitter,” in 2015 European Intelligence and Security Informatics Conference. IEEE, 2015, pp. 161–164.

[15] Twitter, “Ctrlsec,” 2020, URL: https://www.bbc.com/news/world-europe-50424469#:∼:text=But%20three%2Dquarters%20of%20French,sausages%20over%20a%20small%20fire.[accessed: 2019-09-18] [accessed: 2020-09-18].

[16] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, vol. 5, no. 4, 2014, pp. 1093–1113.

[17] H. Ismail, S. Harous, and B. Belkhouche, “A comparative analysis of machine learning classifiers for twitter sentiment analysis.” Res. Comput. Sci., vol. 110, 2016, pp. 71–83.

[18] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, “Sentiment analysis and the complex natural language,” Complex Adaptive Systems Modeling, vol. 4, no. 1, 2016, pp. 1–19.

[19] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining.” in LREC, vol. 6. Citeseer, 2006, pp. 417–422.

[20] A. Qadir and E. Riloff, “Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1203–1209.

[21] P. Norvig, “Natural language corpus data: Beautiful data,” 2020, URL: http://norvig.com/ngram [accessed: 2020-09-18].

[22] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, “Learning sentiment-specific word embedding for twitter sentiment classification,” in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2014, pp. 1555–1565.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, 2003, pp. 993–1022.

[24] S. Narr, M. Hulfenhaus, and S. Albayrak, “Language-independent twitter sentiment analysis,” Knowledge discovery and machine learning (KDML), LWA, 2012, pp. 12–14.

[25] O. Araque and C. A. Iglesias, “An approach for radicalization detection based on emotion signals and semantic similarity,” IEEE Access, vol. 8, 2020, pp. 17 877–17 891.

[26] M. Nouh, R. J. Nurse, and M. Goldsmith, “Understanding the radical mind: Identifying signals to detect extremist content on twitter,” in 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2019, pp. 98–103.

[27] L. Povoda, R. Burget, and M. K. Dutta, “Sentiment analysis based on support vector machine and big data,” in 2016 39th International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2016, pp. 543–545.

[28] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, “A survey on the state-of-the-art machine learning models in the context of nlp,” Kuwait journal of Science, vol. 43, no. 4, 2016.

[29] K. Suresh and R. Dillibabu, “Designing a machine learning based software risk assessment model using naïve bayes algorithm,” TAGA Journal, vol. 14, 2018, pp. 3141–3147.

[30] V. K. Jain, S. Kumar, and S. L. Fernandes, “Extraction of emotions from multilingual text using intelligent text processing and computational linguistics,” Journal of computational science, vol. 21, 2017, pp. 316–326.

[31] V. Malik and A. Kumar, “Sentiment analysis of twitter data using naive bayes algorithm,” International Journal on Recent and Innovation Trends in Computing and Communication, vol. 6, no. 4, 2018, pp. 120–125.

[32] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, “Aspect-based sentiment analysis to review products using naïve bayes,” in AIP Conference Proceedings, vol. 1867, no. 1. AIP Publishing LLC, 2017, p. 020060.

[33] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, “Sentiment analysis of events from twitter using open source tool,” IJCSMC, vol. 5, no. 4, 2016, pp. 475–485.

[34] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.” in Lrec, vol. 10, 2010, pp. 2200–2204.

[35] E. Cambria, D. Olsher, and D. Rajagopal, “Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis,” in Proceedings of the twenty-eighth AAAI conference on artificial intelligence, 2014, pp. 1515–1521.

[36] G. A. Miller, “Wordnet: a lexical database for english,” Communications of the ACM, vol. 38, no. 11, 1995, pp. 39–41.

[37] S. Vohra and J. Teraiya, “A comparative study of sentiment analysis techniques,” Journal JIKRCE, vol. 2, no. 2, 2013, pp. 313–317.

[38] O. Nasraoui, “Web data mining: Exploring hyperlinks, contents, and usage data,” ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, 2008, pp. 23–25.