IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2019, accepted December 27, 2019, date of publication January 17, 2020, date of current version January 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967219

An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity

OSCAR ARAQUE AND CARLOS A. IGLESIAS
Departamento de Ingeniería de Sistemas Telemáticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain
Corresponding author: Oscar Araque (o.araque@upm.es)

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant agreement no. 740934.

ABSTRACT The Internet has become an important tool for modern terrorist groups as a means of spreading their propaganda messages and recruitment purposes. Previous studies have shown that the analysis of social signs can help in the analysis, detection, and prediction of radical users. In this work, we focus on the analysis of affect signs in social media and social networks, which has not been yet previously addressed. The article contributions are: (i) a novel dataset to be used in radicalization detection works, (ii) a method for utilizing an emotion lexicon for radicalization detection, and (iii) an application to the radical detection domain of an embedding-based semantic similarity model. Results show that emotion can be a reliable indicator of radicalization, as well as that the proposed feature extraction methods can yield high-performance scores.

INDEX TERMS Affective computing, machine learning, natural language processing, terrorism, radicalism.

I. INTRODUCTION

Modern terrorist organizations are continuously changing and adapting to achieve their objectives as well as to overcome current counter-terrorist strategies. Terrorists have effectively used the Internet in their propaganda and recruitment strategy [1]. As a consequence, online radicalization detection is a significant concern. Digital traces on social networks can be used as social signs to detect online radicalism, based on text mining and social analysis techniques [1], which analyze different aspects of these interactions such as sentiment, topics, and the social context of users and semantics [2].

Recent works in social science [3], [4] have highlighted the role of emotions for understanding terrorism. In [3], van Stekelenburg argues that the radicalization process can be conceived as an emotional transformation consisting of three phases. In the first phase, the group perceives injustice from an out-group fueled by anger. In the second phase, the group adopts a moral superiority based on the emotion of contempt. Finally, in the third phase, the group decides to eliminate the out-group based on the emotion of disgust. Rice [4] proposes a research agenda in social science to understand better the role of emotions in the etiology of terrorism since this understanding can heighten the explanatory and predictive capacity of the theories. Nevertheless, to our knowledge, previous works in the development of computational models for online radicalization detection and prediction have used only sentiment analysis. This work aims to bridge this gap by researching on the relevance of the emotions expressed in online radicalization publications.

Undoubtedly, there is a trend of studying these issues employing data exploitation techniques, such as machine learning. It is still an open challenge how such technology-driven approaches can be effectively used in the field of terrorism policing [5].

In the current work, we focus on investigating three main research questions (RQ):

RQ1: Can emotion information be used for radicalization detection? If so, how? It has been observed that emotions are related to hostility toward another group, which can evoke radical behavior [6]. Previous work has used sentiment signals, but exploiting emotion information for this task has not been studied thoroughly. Thus, this paper offers a view of how emotion features can be generated and used in a machine learning system for radicalization detection.

RQ2: Can semantic similarity-based features be used effectively for radicalization detection? In learning-based approaches, it is necessary to extract useful features from the input in order to allow the learning model to differentiate between data samples accurately. Currently, it is not clear

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

which kind of features to extract when performing radicalization detection, and a thorough study in this sense has not been made. Besides, semantic similarity has been successfully used in other Natural Language Processing (NLP) tasks and domains, such as Sentiment Analysis [7] and Entity Disambiguation [8]. In this way, this work explores the use of such a measure for radicalization detection.

RQ3: How can radical vocabularies be obtained and exploited? It is clear that radical content makes use of specific vocabulary to express such ideas. So, it is natural to think that integrating this kind of domain knowledge into the model can enhance its performance. Still, how to use and generate such vocabulary selections is an open challenge that has yet to be deeply studied.

In light of these questions, this paper proposes a machine learning method that aims to detect radical text in two domains: online press and the Twitter social network. For this end, the proposed approach generates distributed representations of the text that are fed into a machine learning classifier. These representations are generated through computing the similarity between the analyzed text and a particular lexicon. The similarity measure is obtained through a pre-trained word embeddings model. In this way, we intend to exploit the knowledge contained in word embedding models as well as in a lexical resource. Additionally, this paper also proposes a novel approach that makes use of an emotion dictionary to compute a statistical summary of the emotions present in the analyzed text. In this way, we aim to exploit the idea existing in the literature that emotion can have a role in the elicitation of radicalism [6].

To extend the current scope of this and future research works, we present a new dataset in this work. Such data has been collected from both radical and neutral online press sources that target ISIS-related topics. The original domain differs from the data usually used in previous works, which may allow for more extensive studies of radicalization in texts.

The rest of the paper is organized as follows. Section II presents an overview of the previous work regarding radicalization detection. Following, Section III presents the collected dataset. In Section IV, the proposed radicalization detection model is described. Later, in Section V, we describe the experimental setup which is aimed at evaluating the proposed model, and present the obtained results and their discussion. Section VI presents a comparison of our proposed model to previous works. Finally, the paper concludes in Section VII by depicting the conclusions drawn from the evaluation, as well as offering an outline of possible future lines of work.

II. RELATED WORK

In this section, we review approaches that are related to our proposal. First, Sect. II-A carries out a literature review that covers the works dealing with affect analysis in online radicalization sources. Then, Sect. II-B introduces related approaches for exploiting affect signs and word embeddings in a text.

A. AFFECT ANALYSIS OF ONLINE RADICALISM

The main problems that have been addressed in previous research for the automatic processing of online radicalization are analysis, detection, and prediction of radicalization [27], [28].

Online radicalization analysis aims at providing actionable information to improve Law Enforcement Agencies (LEA)'s decision making processes. According to Correa and Sureka [27], analysis approaches fall in two main categories: network-based (leaders, communities and topology characteristics) and content-based (authorship identification, stylometric analysis, website activities, affect analysis and usage).

The works focused on detection can be classified into web mining and text mining approaches [27]. Web mining works aim at detecting radical online content using techniques such as focused crawling [29], while text classification techniques aim at developing a binary classification model based on text features [11] combined with other features such as social dynamics [11], [28].

Regarding detection, different types of analysis have been proposed for understanding better online radicalization, including network and content-based analysis. While network analysis takes into account the social interactions, a content-based analysis is focused on analyzing several aspects of radical texts, such as affects, topics or stylistic features.

Finally, several works have addressed prediction. Ferrara *et al.* [30] propose a machine learning framework for detecting extremist supporters, predicting extremist user content adoption (i.e., a retweet of extremist content) and predicting interactions with extremists (i.e., reply direct messages from extremists). The framework considers three types of features: user and activity, timing and network. Agarwal and Sureka [31] presents a survey focused on two problems, the automatic identification of online radicalization (hate promoting content, users and hidden communities) and the prediction of civil unrest related events (protests, riots, public demonstrations). The survey concludes that most works have used spatiotemporal features as discriminatory ones for predicting events. López-Sáncez *et al.* [26] propose a system for predicting radicalization risks. They propose to generate alarms based on the radicalization influence of monitored users and the emotional load of the received retweets.

In this section, we provide a more detailed analysis of the use of affect technologies for analyzing online radicalization. As shown in Table 1, affect analysis has been applied in a wide array of domains, such as radical forums [9], [15], [18], radical magazines [16], and social networks (Twitter [2], [11], [11], [14], [19], Facebook [20] and YouTube [1]).

Regarding the affect model, most works analyze the polarity (i.e., valence or sentiment) using sentiment analysis techniques [1], [2], [14], [16], [18], [19], [24]. Other works

TABLE 1. Summary of literature survey regarding the use of affective technologies for processing online radicalization signals.

Reference	Type	Data	Method	Affect Model
[9]	Analysis	Forum posts	Linear regression	Intensity of hate and violence terms
[1]	Analysis	YouTube Group	Sentiment analysis of most frequent terms	Sentiment (polarity) using SentiWordNet [10]
[11]	Analysis	pro-ISIS Twitter users	Definition of windows per analyzing the interaction and an algorithm for calculation the adoption of a term based on lexical, sharing and interactions homophily	Sentiment analysis (polarity) using MPQA [12] and ArSenL [13]
[14]	Analysis	Twitter	Lexicon based Sentiment analysis	Sentiment (polarity) using SentiWordNet [10]
[15]	Analysis	Dark Web Forum postings	Support Vector Regression ensemble	Manual annotation of the intensity (0-1) of violence, anger, hate, racism and sentiment
[16]	Analysis	Dabiq radical magazine	Sentiment analysis based on LIWC (Linguistic Inquiry and Word Count) [17]	Anger, anxiety and emotion tone from LIWC
[18]	Analysis	Forum posts	Threshold of negative words to calculate sentence polarity	Sentiment (polarity) using SentiWordNet [10]
[19]	Analysis	ISIS fan girls tweets and control datasets	Lexicon based Sentiment Analysis toward some topics identified by keywords	Sentiment (polarity)
[20]	Analysis	Images and text from Facebook posts related to Paris attack	Domain transfer learning with SentiBank [21] for images and LIWC [17] for texts	Sentiment (polarity)
[22]	Detection	pro and anti-ISIS tweets	Sentiment analysis (polarity) based on CoreNLP [23] that uses a Recursive Deep Model over a Sentiment Treebank.	Feature based classification using SVM, Naive Bayes and Adaboost
[24]	Detection	Twitter	Single-class SVM and KNN	Binary model to indicate if a tweet includes terms that indicate a negative emotion
[2]	Detection	pro and anti-ISIS tweets	Lexicon based Sentiment Analysis	Polarity (sentiment): neutral, positive, negative using SentiStrength [25]
[26]	Prediction	Twitter	Alerts based on user influence (followers, followed and messaged retweeted) and sentiment load of received mentions	Sentiment (polarity) using SentiStrength [25]

evaluate the intensity of some terms, such as hate or violence [9], [14]. Finally, other works [15], [20] use LIWC's categories for affective processes [17]: positive emotions, negative emotions, anxiety, anger, and sadness.

The analysis of sentiment has provided many insights. Abbasi and Chen [9] conclude that there is a linear relationship between the intensity scores for violence and hate affects, which is strong in Middle Eastern forums and weak in U.S. forums.

Another interesting aspect is the process of radicalization. Rowe and Saif [11] notice when analyzing the process of radicalization that users, before being activated, frequently discuss politics, using words like 'Syria' or 'Egypt' with negative sentiment. Once they are activated, they tend to use more religious words and some words as ISIS are used with negative sentiment since they prefer the term 'Islamic State'.

Some analyses that have also considered the temporal evolution of the text with different purposes, such as the comparison of different radical blogs [15] or the language used in Dabiq radical magazine [16].

Finally, another compelling aspect is how public opinion reacts when a terrorist attack has happened. Dewan *et al.* [20] analyzed both the sentiment of images and texts posted on Facebook. They observed that the sentiment for textual posts was negative during the first few hours but gradually moved to the positive side over time. Contrarily, posted images reflected positive sentiment initially but moved towards negative after the first few hours.

There is not an agreement in current research regarding the importance of affect signals. Several studies [18], [22], [24] highlight their importance, while in other experiments [2] sentiment features, which consist of both, word unigrams and their sentiment have no impact on the classification performance compared to using unigrams only.

B. WORD EMBEDDINGS AND AFFECT LEXICONS

Word embedding representations encode both semantic and syntactic regularities that are extracted through training from large amounts of not annotated text [32]; such regularities are expressed in the resulting vector space as relation offsets. Commonly, word embedding approaches are trained using an unsupervised training process where word vectors are generated, forming word representations that capture relevant language knowledge. Such type of word vectors, whose training process relies on co-occurrence data, are denoted pre-trained word vectors. This work makes use of this type of word vectors for one of the proposed feature extraction method.

As traditionally done with bag-of-words features, word embeddings can be used as features for textual representation

in text categorization tasks, including sentiment analysis [32]. Several works tackle the use of word embeddings as feature extractors in text classification. The work described in [33] constitutes a study of the effectiveness of word embeddings when applied to various tasks, including sentiment analysis. This study also provides a summary of unsupervised embedding models and how they can be used to acquire text representations. In the work of [34], word embeddings are used as features for a SVM classifier in the task of polarity detection. After the evaluation, the authors conclude that word embeddings may contain semantic information among words, which can be leveraged for obtaining useful text representations. In addition, word embeddings can be used in combination with more traditional features, such as n-grams, lexicons and lexical features to improve the quality of text representations [35]. Authors also show that the use of embeddings in an ensemble model can produce higher prediction performance. Also, it is worth to consider that techniques based on word embeddings are widely used in public challenges in which competitors aspire to obtain the highest scores in a wide variety of tasks [36]–[38]. Reference [39] introduces an interesting work where word embeddings are used in a relevant analysis of computation complexity reduction. In a compelling experiment, the authors report similar performance metrics in comparison to more complex neural models in several tasks. Furthermore, [40] presents a model that uses both a measure of semantic similarity and embedding representations. For a comprehensive description of how to exploit word embeddings, we refer the reader to [41].

Considering affect lexicons, the challenge of adequately using an affect lexicon is a common issue that frequently appears in Sentiment Analysis literature. The interested reader can consult an interesting survey regarding the use of affect lexicons [42]. There are numerous affect lexicons available, and deciding which qualities drive the final performance of a machine learning system that uses them is currently an open challenge. In this line of work, [43] deals with some of these topics, analyzing a series of affect lexicons and how they can be complemented.

As done in this work, affect lexicons have been used as features in supervised machine learning scenarios. For example, in [44], lexical resources are combined with both specific micro-blogging features and n-grams. In such a way, the authors show that these features are useful in their validation. Similarly, the work described in [45] uses two sentiment lexicons and then combines such information with n-gram and Part-Of-Speech (POS) features. All of these features are used by a SVM classifier, which resulted in a state-of-the-art system in a Sentiment Analysis competition. There are cases where lexicon features have been supplemented with domain-specific features, as done in [46], where a search query settles the domain. Of course, several affect lexicons can also be aggregated, as is the case of [47]. That work integrates several lexicons by means of Markov's logic, using information about the relations between neighboring words.

III. MAGAZINES DATASET COLLECTION

The scarcity of annotated data in the radicalization detection field is notable, which difficult the advancement of research. As an additional issue, we have to consider that the majority of existing datasets are oriented to the Twitter domain [28]. Indeed, Twitter represents a valuable source of information in this area, but generalization concerns have to be taken into account when addressing radicalization detection. Thus, intending to expand the scope of both our proposal and future research, we have collected the *Magazines dataset*.

Firstly, we have considered the Dabiq [48] and Rumiyah [49] online magazines, which are distributed by Islamic State of Iraq and the Levant (ISIS) radical organization [50]. These magazines are written in the English language. Dabiq was released from July 5, 2014, to July 31, 2016, by Al Hayat Media Centre, the branch of ISIS's Ministry of Media which produces material in English. After producing fifteen issues of Dabiq, the organization released the first issue of Rumiyah on September 6, 2016. Thirteen issues of Rumiyah have been released on more or less monthly schedule until September 9, 2017. Most analysts consider that Rumiyah has replaced Dabiq [51]. The change in the name is attributed to a change in ISIS's objectives: from Dabiq, a northern Syrian city where jihadists believe they would be defeated, to Rumiyah (Rome in Arabic), a West battle space outside Syria and Iraq [50].

Since the original Dabiq and Rumiyah publications are not freely accessible, they have been collected from an online resource that tackles terrorism[1]. This online site makes the magazines available for research purposes. The source data is available in PDF format, so the textual data has been extracted. Since we aim to evaluate textual data, the images have been omitted. The resulting data consists of the 15 issues of Dabiq, and the 13 issues of Rumiyah. In total, the number of articles from Dabiq and Rumiyah is 161 and 155, respectively.

As a counterpoise of the previous radical content, we have selected two online newspapers that address ISIS-related issues but are not radical sources: CNN[2] and The New York Times[3]. These sources have been selected since both represent prominent newspapers that frequently mention radical and ISIS-related topics from a neutral perspective. Also, such an addition to the data has been done in the line of previous work [28], [52] where the authors complement radical content with non-radical data. Such content is freely accessible and can be obtained through the newspapers' APIs. The data has been obtained using domain-related keywords (*ISIS*, *Daesh*, *Islamic State*, etc.) on a time frame of 10 months. During this time, articles were automatically collected. In order to enhance the quality of the collected data, manual filtering was done, removing articles that are not related to the topic

[1] http://www.jihadology.net
[2] https://cnn.com (http://developer.cnn.com/docs/)
[3] https://www.nytimes.com (https://developer.nytimes.com/apis)

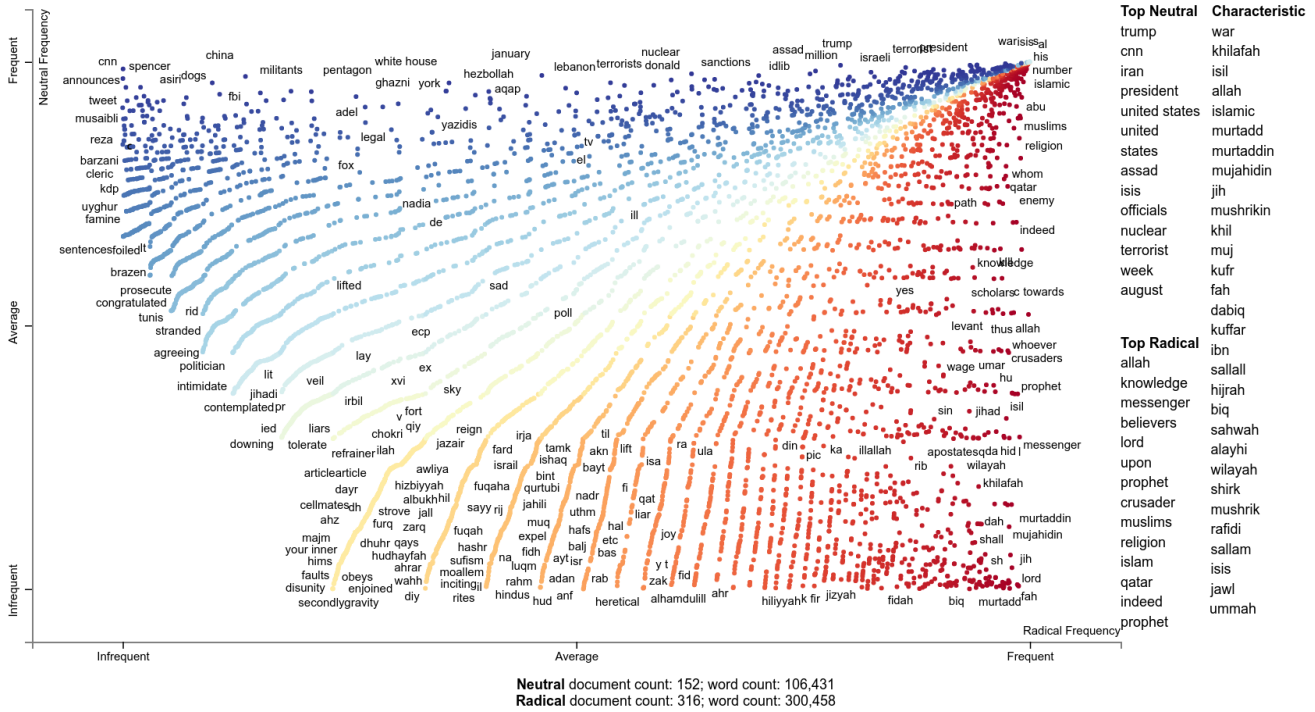

FIGURE 1. Frequency of words for both neutral and radical categories. On the right, most frequent words for neutral (Top Neutral) and radical (Top Radical), and both (Characteristic).

TABLE 2. Statistics of the collected dataset.

	CNN	NYT	Dabiq	Rumiyah
No. of articles	129	23	161	155
Avg. no. of words	860	271	1097	35
Avg. no. of sentences	35	10	35	32
Avg. no. of word appearances				
Allah	0.08	0.0	9.05	14.83
knowledge	0.09	0.08	0.55	0.90
Trump	1.87	0.09	0.01	0.0
Iran	2.46	0.30	0.22	0.07

at hand. Images, links, and other media are removed from the articles, leaving the text, both titles and bodies. In total, 129 articles have been collected from CNN, and 23 from The New York Times.

The same preprocessing has been applied to all the obtained data: normalization of numbering, capital letters, contractions (e.g., *I've*, *we'll*) is done, resulting in documents composed of lower case tokens. Table 2 presents a summary of some statistics of the resulting dataset.

Figure 1 shows a visualization of the introduced dataset, as described in [53]. This graph consists of a frequency scatter plot of the dataset words accordingly to the neutral and radical category. We consider the Dabiq and Rumiyah texts to be radical, in opposition to the CNN and New York Times articles, which we label as neutral. The color indicates the frequency in relation to each class: blue for neutral, red for radical. For each category, the frequency of each different word to that category is computed. A subset of the word labels is placed along the figure due to space constraints. In this way, the x-axis indicates the frequency in the radical category: if a word frequently appears in radical texts, it is placed on the right area.

Similarly, the y-axis encodes the frequency in the neutral category. A word that frequently appears in neutral texts will be placed in the top area. As a consequence, it is of particular interest the areas where the more frequent words appear: top left (frequent in neutral texts), bottom right (frequent in radical texts) and top right of the figure (frequent in both neutral and radical texts). In those areas, the most characteristic words for the neutral, radical, and both categories are presented. These areas offer a view of how the words are used in these two categories. For example, common radical words are *Allah*, *knowledge*, and *messenger*, that stress the religious tone that can be found in the narratives of the Dabiq and Rumiyah magazines. In contrast, the articles found in the neutral newspapers have a more informative tone, frequently using words such as *Trump*, *CNN*, and *Iran*. Given the implication of the United States in the affairs related to ISIS, it is understandable that such words are being frequently used in the neutral class.

IV. EXPLOITING EMOTION FEATURES AND SEMANTIC SIMILARITY FOR RADICALIZATION DETECTION

This work proposes a model that exploits several sources of information, aiming at improving the final performance in

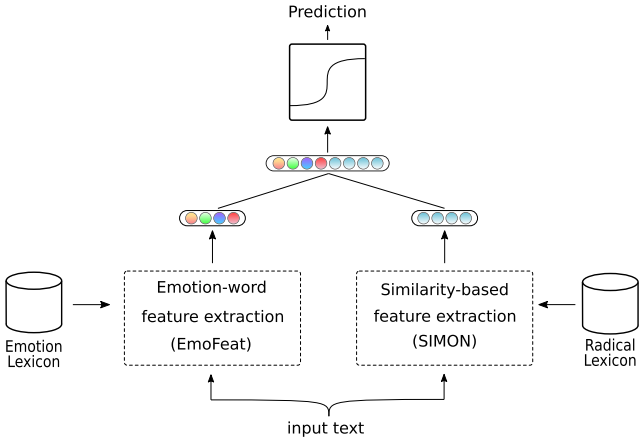

FIGURE 2. General architecture representation of the proposed model.

the task of radicalization detection. For this end, a machine learning system is introduced that consists of two submodules: (i) emotion-based (Sect. IV-A) and (ii) embedding word similarity feature extraction (Sect. IV-B).

Figure 2 shows a diagram of the proposed model. As shown, the analyzed text is processed by the two processing modules. Both of them use as input the natural language, yielding as output a feature vector that represents the input text. These feature vectors are concatenated and fed to a machine learning classifier, which outputs a prediction based on the information given by the features mentioned above. As classifiers, in this work, both Logistic Regression and Linear SVM are considered.

A. EMOTION BASED FEATURES

It has been discussed in the literature that emotions can play a role in the elicitation of radical behavior [6]. Sentiment information has been used in a machine learning system in order to distinguish between radical and non-radical content [54]. Nevertheless, to the extent of our knowledge, the effect of the emotions present in the text over a learning system has not been thoroughly studied in the task of radicalization detection.

In this manner, this work proposes the use of an emotion lexicon in order to extract emotion-driven features that are, as explained, fed to a machine learning algorithm. With this, we intend to investigate whether this kind of information is relevant for the task of radicalization 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. For simplicity, we denote this method as *EmoFeat* (Emotion Features).

Consider an emotion lexicon, composed by a vocabulary $W^{(l)} = \{w_1^{(l)}, \ldots, w_i^{(l)}, \ldots, w_P^{(l)}\}$ and a vector of numeric annotations $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots \boldsymbol{l}_i, \ldots \boldsymbol{l}_P]$. Note that such lexicon has an annotation $\boldsymbol{l}_i$ for each word w_i, with a total of P $(w_i, \boldsymbol{l}_i)$ pairs. Note also that $\boldsymbol{l}_i$ is a vector expressing the intensity of each emotion for word $w_i^{(l)}$ in the lexicon. An emotion vector has dimensionality $\boldsymbol{l}_i \in I\!R^m$, and thus the emotion lexicon annotation matrix is $\boldsymbol{L} \in I\!R^{P\times m}$, where m is the number of emotions considered in the lexicon. Following, let $W^{(i)} = \{w_1^{(i)}, \ldots, w_j^{(i)}, \ldots, w_I^{(i)}\}$ be the set of length I composed by the input words.

For each word w_k of the intersection $W^{(l)} \cap W^{(i)}$, the associated emotion vector is extracted from $\boldsymbol{L}$. This process outputs a matrix that contains the emotion annotation for all the input words that appear in the lexicon. Next, several statistical measures are taken in order to represent such a matrix as a feature vector. The proposed measures are average, maximum, and median. Please note that these measures can be used independently: for example, only average and median could be used, removing maximum. As a result, a feature vector is obtained with dimension $n \cdot m$, where n is the number of statistical measures selected.

Algorithm 1 EmoFeat

Require: Emotion lexicon composed by a vocabulary $W^{(l)}$ and annotations $\boldsymbol{L}$
Ensure: $\boldsymbol{v} \in I\!R^{n\cdot m}$, the final feature vector
1: **for all** $w_k \in W^{(l)} \cap W^{(i)}$ **do**
2: $\quad \boldsymbol{E}_{k,:} \leftarrow$ emotionAnnotation$(w_k, \boldsymbol{L})$
3: **end for**
4: **for** $i \leftarrow 1, n$ **do**
5: $\quad$ **for** $j \in 1, m$ **do**
6: $\quad\quad index \leftarrow i + n(j-1)$
7: $\quad\quad \boldsymbol{v}_{index} \leftarrow$ statMeasure$(i, \boldsymbol{E}_{:,j})$
8: $\quad$ **end for**
9: **end for**

The proposed feature extraction method can be expressed as an algorithm, as shown in Algorithm 1. The function *emotionAnnotation* extracts the emotion annotation vector corresponding to word w_k from matrix $\boldsymbol{L}$. Also, the function *statMeasure* computes the corresponding statistical measure from the column j of matrix $\boldsymbol{E}$. The index i indicates which statistical measure is applied (e.g., average if $i = 1$, maximum if $i = 2$, etc.).

B. EMBEDDING BASED SEMANTIC SIMILARITY

A well-known advancement of the NLP field is that distributed representation based techniques have become the state-of-the-art models on text processing problems [55]. Among the different methods for computing distributed representations, we highlight word embeddings, as they arguably represent the most popular choice of such text representations. In light of this, it is natural to presume that word embeddings can be leveraged for the task of radicalization detection effectively. Nevertheless, pre-trained word embedding models do not contain specific knowledge of the task at hand, since these models are trained from large corpora in an unsupervised manner.

One possible solution that has been studied before [56] would be to train a word embedding model using a

sufficiently large corpus with explicit radical detection annotations. Unfortunately, a data repository that fulfills these requirements is not currently available for the research community [57].

In this way, we propose the application of the SIMilarity-based sentiment projectiON (SIMON) method [7] as a feature extractor in radicalization detection. This method makes use of a word embedding model and orients the extracted features to a particular domain utilizing a domain-centered lexicon. In this work, the SIMON method is adapted to extract features for radicalization detection by using radical-oriented lexicons.

The main idea of the SIMON method is that given a domain lexicon, the input text is measured against it, computing a vector that encodes the similarity between the input text and the lexicon. With such a model, the model can exploit the knowledge contained in word embeddings, as well as the domain information that the lexicon offers. As an additional remark, this method does not need large corpora to be trained and thus can be used in problems where annotated data is scarce.

As stated, the application of SIMON towards this domain is achieved by composing domain-oriented lexicons. The generation of domain adapted lexicons is a challenge in itself [58], but in this work, we proposed a simple method for the generation of a domain-specific lexicon. Please note that, for the use of a lexicon made by the SIMON method, numeric annotations are not needed. As previously demonstrated [7], using such annotations does not necessarily lead to an improvement of results.

Thus, a collection of words is generated through filtering by the frequency of appearance in the training data. This selection of words is used as a lexicon for the SIMON method. Such a selection method is oriented to represent a simple baseline to capture the vocabulary of a specific domain. We refer to this frequency-based selection as *FreqSelect* in the rest of the paper.

Additionally, the SIMON method allows the generated collection of words to be subsequently filtered. Such filtering is done by means of an ANOVA-based features filtering, attending to the feature relevance for the classification task [7].

V. EVALUATION

The evaluation of the proposed model has been done through a text categorization task, where given a piece of text, the aim is to detect whether it does contain radicalization evidence. In this way, the evaluation is done by performing a binary classification task. As such, the proposed approaches have been validated using the datasets, lexicons, and embeddings listed in Section V-A, and following the methodology described in Section V-B. Results of such experiments are shown in V-C.

A. MATERIALS

The datasets used for the evaluation are presented in Table 3. Apart from the dataset presented in Section III, other two datasets have been used, which contain data from Twitter.

TABLE 3. Statistics of the datasets used for the evaluation: number of positive (radical), negative and total instances.

Dataset	Positive	Negative	Total
Pro-Neu [28]	112	112	224
Pro-Anti [11]	566	566	1,132
Magazines (this work)	316	152	468

TABLE 4. Statistics of the twitter datasets used in the evaluation.

Statistic	Pro-Neu [28]		Pro-Anti [11]	
	Pro	Neu	Pro	Anti
Total no. of tweets	17,350	197,743	602,511	1,368,827
Avg. no. of tweets	154	1,765	1,065	2,418
Total no. of words	271,842	3,151,107	3,945,815	9,375,841
Avg. no. of words	2,427	28,134	6,971	16,570

Pro-Neu. This dataset is composed of two different datasets, both in English, collected by [28]. The first one contains 17,350 tweets extracted from 112 distinct Twitter pro-ISIS accounts, which can be found online[4]. Through a study of three months, a set of users was identified using a selection of keywords (e.g., *Dawla*, *Amaq*, *Wilayat*), and filtered according to their use of images (e.g., radical leaders images, ISIS flags) and their network of followers. The second one, which comprises 122k tweets from more than 95k accounts, has been used as a counterexample from the pro-ISIS instances since it contains ISIS-related messages (that can be either neutral or anti-ISIS). This set of tweets was collected using ISIS-related keywords (e.g., *ISIS*, *ISIL*, *Daesh*, *IslamicState*, *Raqqa*, *Mosul*). From the original accounts, a selection of 112 is made, as done in [28]. This additional filtering is done filtering accounts that are not active nowadays from the original dataset, thus ensuring that the remaining accounts are not pro-ISIS. For comparison purposes, the same selection and split from [28] are performed.

Pro-Anti, formed by 1,132 Twitter accounts and their timelines, which have been collected in [11]. This dataset is in English. In the mentioned work, the authors identified users as pro-ISIS by attending to their sharing activity of incitement material from known pro-ISIS accounts, as well as the use of extreme language. Initially, the authors had 727 accounts identified, but 161 of these Twitter accounts were either suspended or hidden from public access. Such status prevents from accessing the profile information. Consequently, these 161 accounts were removed, obtaining 566 pro-ISIS users in total [11]. In order to balance the data, the authors added 566 anti-ISIS users. The annotation of anti-ISIS accounts has been done by observing the use of anti-ISIS language. This dataset has been used in [2], a fact that enables comparison with our work.

Magazines, which is the dataset presented in this work, as described in Section III.

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

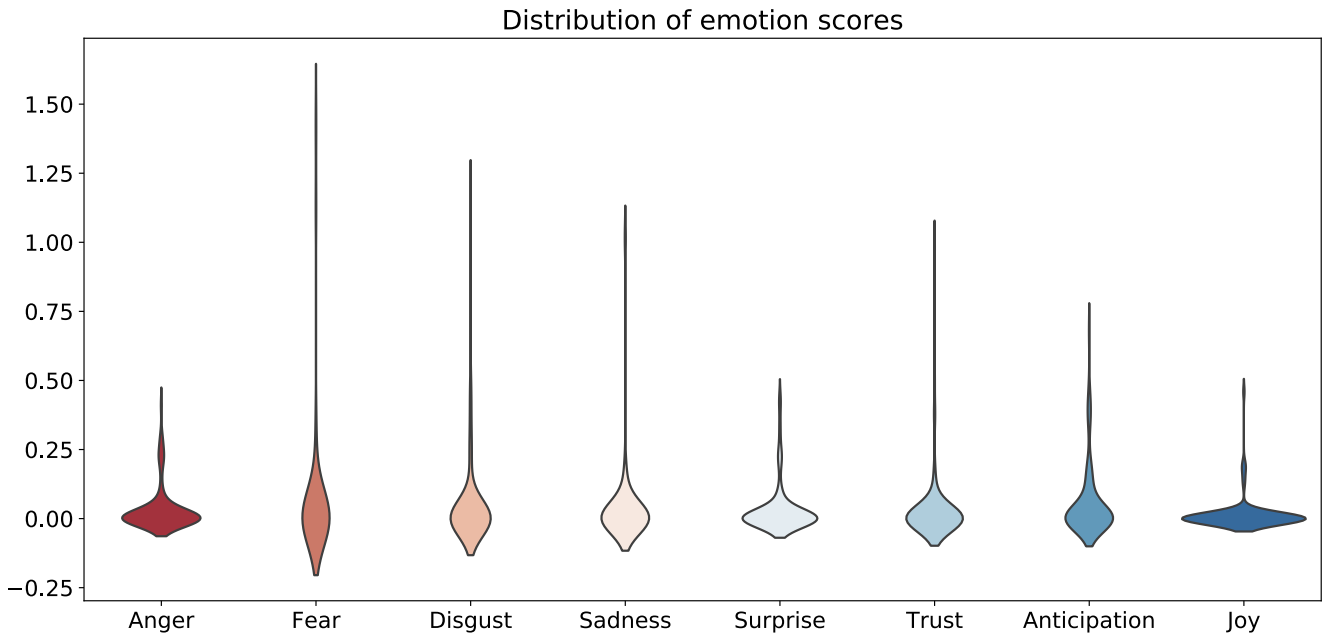

FIGURE 3. Distribution of raw emotion scores for all considered emotions.

TABLE 5. Statistics of the word vectors used for the evaluation: name of the vectors, reference, vocabulary size (no. of words), and training domain.

Name	Reference	Vocabulary size	Training domain
Word2Vec	[32]	3,000,000	News
FastText	[59]	1,999,995	Wikipedia
Glove	[60]	400,000	Wikipedia + Gigaword[5]

Also, as described in Section IV, word embeddings are extensively used in the proposed model. In order to assess the differences in performance among various pre-trained models, we make use of three popular resources extensively used in the Natural Language Processing community. Table 5 shows the characteristics of such vectors. Please note that, in order to normalize the dimension of the vectors, in all three variants we use 300-dimensional vectors.

Another resource is the emotion lexicon used, as explained in Section IV-A. For this work, the selected lexicon is the NRC Hashtag Emotion Lexicon [61], which has been curated specifically for the Twitter domain. This resource contains 16,862 words annotated with eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Figure 3 shows the distribution of emotions scores along with a union of the evaluation data. Although a relevant portion of the scores is close to 0 (no emotional content) it can be observed that for all emotions, there is a long tail in the distribution. Such an elongated distribution contains a variation that is exploited by the emotion feature extraction model. We have examined the emotion distribution along to each dataset and annotation, confirming that there are no relevant differences between those. This indicates that our dataset is balanced in terms of emotion distribution and that there are not any biases that may corrupt the evaluation.

B. EVALUATION METHODOLOGY

The proposed model (Section IV) can be implemented with a variety of resources and methods. In order to thoroughly evaluate the effectiveness of the model, an extensive experimental setup has been performed. In all experiments, the weighted average of the F1-Score is used as the performance metric. To evaluate an instance of our model, cross-validation via k-fold is used, with a $k = 10$.

Concerning the SIMON features, the experiments consider three variations, attending to which (i) word vectors and (ii) collection of words are used, as well as (iii) the percentage of the word collection retained. As explained in Section IV-B, a collection of words is used in order to compute the similarity to the analyzed text. This collection is extracted from the training dataset by measuring the appearance frequency of the dataset vocabulary: this is done by the FreqSelect method. Although, in order to compare to other word selections, we include in the experiments the use of two lexicons: Bing Liu's [62], a popular sentiment lexicon (as done in [7]); and a radical-oriented lexicon that is the result of merging different radical-related lexicons (for more information, please consult the original publication [28]). With respect to the third parameter, it expressed the percentage of the collections of words that is retained after performing an ANOVA based filtering. In this way, the extracted features can be filtered, attending to how relevant are for the classification problem [7]. A percentage of 100 would indicate that no filtering is performed.

As for the EmoFeat method, an evaluation of the number of emotions considered (parameter m) and the number of

the statistical measures selected (parameter n) is done. Thus, we use the Recursive Feature Elimination (RFE) method [63] to select among all the possible combinations of these parameters.

In order to evaluate the use of the different features by our model, the experiments explore the performance using the following selection of features: (i) EmoFeat, (ii) SIMON, and (iii) EmoFeat combined with SIMON. Also, we have tested two classifier algorithms: Logistic Regression and Linear SVM, since the primary objective is to evaluate the effectiveness of the feature extraction methods.

C. RESULTS

Firstly, in order to evaluate the influence of the number of emotions considered (parameter m) and the number of statistical measures selected (parameter n), the RFE method has been used [63]. This method outputs a feature ranking (the lower, the better) that can be used to eliminate features. Consequently, we explore such a raking, as given by RFE, when applied to the features generated by the proposed EmoFeat approach. In this way, taking into account the selected emotion lexicon (NRC Hashtag Emotion Lexicon [61]), the maximum number of emotions considered is $m = 8$, and $n = 3$, which outputs a maximum number of features $m \cdot n = 24$. Following, the RFE strategy has been applied to these features across all considered datasets and machine learning classifiers. Attending to the parameter n, the *median* measure has a lower rank on average. Thus, we evaluate the F-Score obtained using the three measures (*average* + *maximum* + *median*), which raises to 79.01%, against the performance removing the *median* (*average* + *maximum*), that yields 79.18%. Please note that these F-Scores are averaged over all datasets and classifiers. Next, the parameter m is evaluated. The emotion that has the lower ranking, as computed by the RFE method, is *Joy*. Similarly, a comparison is made between the averaged performance including all emotions, which is again 79.18%, and the performance obtained when removing *Joy*, that outputs 78.97%. It can be seen that excluding the *Joy* emotion decreases the overall performance of the system. As a conclusion, for the rest of the experiments, we consider $m = 8$ (all emotions) and $n = 2$, selecting the average and maximum measures.

Secondly, we evaluate the three variations relevant to the SIMON approach: word vectors, collection of words, and percentage of filtering over this collection. Figure 4 shows the results. The percentage of selection (horizontal axis) presents a general trend, where the performance improves proportionally with the percentage. The best scores, with only one exception, are obtained when using a percentage of 100%; that is, with no filtering.

As for the word collections used, it can be seen the difference in performance among the three variations. As expected, the domain-oriented FreqSelect outperforms the other two collections in all cases. This improvement indicates that adapting the word collection leads SIMON to an increase in performance in the radical classification task. Nevertheless, the radical collection used as baseline performs poorly in comparison to the sentiment-oriented collection. This accuracy difference could be explained attending to the composition of the radical lexicon, that contains a large number of Arabic terms that do not appear in the word vectors. In such cases, the similarity between the analyzed text and the out of vocabulary words cannot be computed, which leads to a loss of information that can be relevant to the problem at hand.

As for the word vector comparison, we observe that the best performance is obtained by using the FastText word vectors in most cases. Nevertheless, this difference can be seen more easily by attending to Tables 6 and 7. In these tables, the F-Score is shown for each combination of our models, and the best scores are marked in bold. Please note that the different word collections (sentiment, radical and FreqSelect) do not apply when uniquely evaluating the emotion-based features (Sect. IV-A), since this approach does not make use of such collections.

As mentioned, the best performance is obtained in almost all cases when using FastText embeddings. Also, in the cases where this is not the case, the difference in performance to that using FastText word vectors is small, being 0.45% the highest. This difference could be explained attending to training method of the embedding models: FastText does take into account subword information, while Word2Vec and Glove do not [59].

Paying attention to the extracted features, it is clear that although solely using the EmoFeat method does not yield the best results, its performance is competitive. This observation is especially true in the pro-neu dataset, where the F-Score obtained is 92.41% when using a Linear SVM classifier. This result enforces the idea that emotion analysis can play a relevant role in radical content detection. The scores obtained when using emotion-based features surpass that of specific approaches that merge social and computational concepts [28].

Following, the combination between the emotion-based and SIMON features has been evaluated. As shown, combining these two types of features does not always improve over the performances obtained by using SIMON uniquely. This is quite evident when looking at the best performances (in bold). Such an effect could be due to overfitting, since the combination augments the number of features, causing the performance score to stop improving, or even to decrease. The reduced size of the datasets further supports this hypothesis. Nevertheless, there is still a large number of cases where this combination improves over the separate methods, which indicates that a combination of the proposed features can lead to an improvement in the performance.

In order to further study the impact of the proposed features and their performance, we have performed the Friedman statistical test [64]. As a result, the Friedman test outputs a ranking of methods attending to their performance in the different datasets. The lower the ranking, the better a specific method performs in comparison to the rest. This test has been computed with $\alpha = 0.01$. For simplicity, Table 8 shows

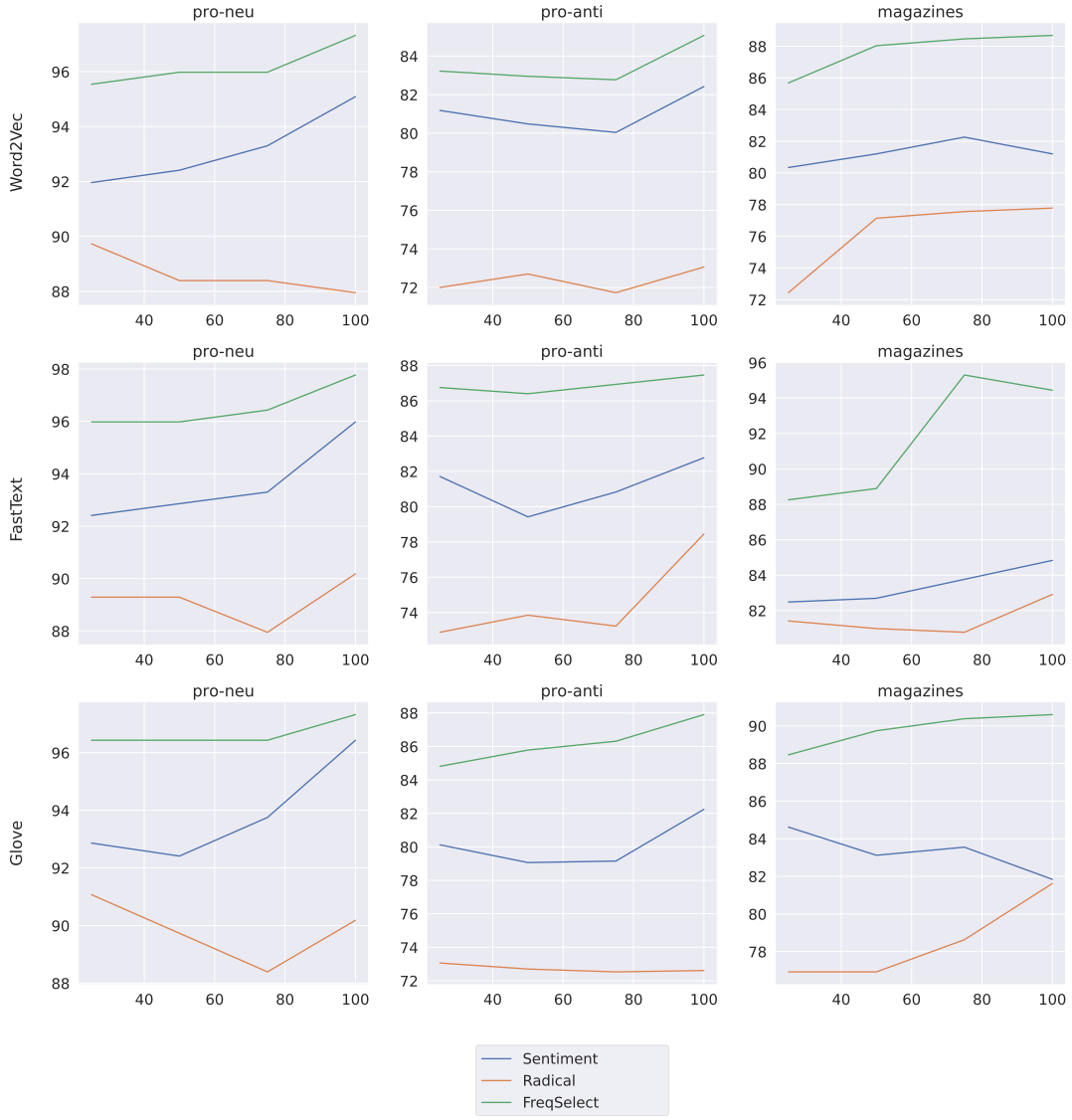

FIGURE 4. Evaluation of SIMON variations on the three datasets. Rows express the word vectors used, columns differentiates the datasets, and the word selections are shown in color. Vertical axis expresses the weighted F1-Score, while horizontal axis indicates the percentage of filtering.

TABLE 6. Averaged F1-Scores for the model using emotion, SIMON, and SIMON + emotion, using logistic regression as classification algorithm.

LOGISTIC REGRESSION		PRO-NEU			PRO-ANTI			MAGAZINES		
EmoFeat		91.07			73.95			68.59		
		Sentiment	**Radical**	**FreqSelect**	**Sentiment**	**Radical**	**FreqSelect**	**Sentiment**	**Radical**	**FreqSelect**
SIMON	Word2Vec	95.09	87.95	97.32	82.42	73.06	85.07	81.20	77.78	88.68
	Facebook	95.98	90.18	**97.77**	82.77	78.45	87.46	84.83	82.91	**94.44**
	Glove	96.43	90.18	97.32	82.24	72.61	**87.90**	81.84	81.62	90.60
SIMON + Emotion	Word2Vec	95.54	88.39	97.32	82.77	75.00	85.16	81.84	80.34	89.53
	Facebook	95.09	90.62	97.32	82.69	79.59	**87.90**	85.47	83.97	**94.44**
	Glove	96.43	91.07	**97.77**	82.77	73.67	87.10	83.33	82.26	91.03

the best six approaches, as computed by the Friedman test. As can be seen, the lower rank is obtained by two methods that make use of a Linear SVM classifier, the FastText word vectors, as well as the FreqSelect word collection. More importantly, the features used are both SIMON, and SIMON combined with emotion. This is consistent with the current

TABLE 7. Averaged F1-Scores for the model using emotion, SIMON, and SIMON + emotion, using linear SVM as classification algorithm.

LINEAR SVM		PRO-NEU			PRO-ANTI			MAGAZINES		
EmoFeat		92.41			77.21			72.22		
		Sentiment	**Radical**	**FreqSelect**	**Sentiment**	**Radical**	**FreqSelect**	**Sentiment**	**Radical**	**FreqSelect**
SIMON	Word2Vec	95.54	80.80	97.77	80.12	65.72	83.92	77.78	75.43	88.46
	Facebook	95.09	89.29	**99.11**	78.71	69.52	86.22	83.12	79.49	**94.02**
	Glove	97.32	86.16	98.66	79.15	65.46	85.07	79.49	79.91	88.25
SIMON + Emotion	Word2Vec	95.09	88.39	97.77	79.51	64.75	83.75	79.27	79.27	87.39
	Facebook	95.09	91.07	**99.11**	78.80	72.53	**86.57**	83.12	83.12	93.80
	Glove	96.88	87.95	98.66	79.42	70.23	85.42	80.34	78.85	89.10

TABLE 8. Friedman rank for the six best methods. EF is the proposed EmoFeat method.

Classifier	Features	Embedding	Collection	Rank
Lin. SVM	SIMON	FastText	FreqSelect	3.5
Lin. SVM	SIMON + EF	FastText	FreqSelect	3.5
Log. Reg.	SIMON	FastText	FreqSelect	3.66
Log. Reg.	SIMON + EF	FastText	FreqSelect	4.67
Log. Reg.	SIMON + EF	GloVe	FreqSelect	5.17
Log. Reg.	SIMON	GloVe	FreqSelect	6.17

TABLE 9. Averaged F1-Scores for the model using emotion, SIMON, and SIMON + emotion, using logistic regression as classification algorithm in a cross-dataset evaluation.

		PRO-NEU	PRO-ANTI	MAGAZINES
		EmoFeat		
PRO-NEU		-	65.81	63.25
PRO-ANTI		73.66	-	66.24
MAGAZINES		45.54	47.53	-
		SIMON		
PRO-NEU	Sentiment	-	69.61	71.79
	Radical	-	66.25	68.80
	FreqSelect	-	61.75	70.30
PRO-ANTI	Sentiment	55.80	-	65.17
	Radical	66.07	-	67.74
	FreqSelect	54.02	-	60.90
MAGAZINES	Sentiment	66.07	63.43	-
	Radical	64.73	42.93	-
	FreqSelect	63.39	38.34	-
		SIMON + EmoFeat		
PRO-NEU	Sentiment	-	69.61	72.22
	Radical	-	66.61	67.31
	FreqSelect	-	61.75	71.79
PRO-ANTI	Sentiment	57.59	-	65.38
	Radical	71.43	-	71.79
	FreqSelect	56.70	-	68.38
MAGAZINES	Sentiment	60.27	56.63	-
	Radical	51.52	41.17	-
	FreqSelect	58.93	38.25	-

discussion and sheds light in the fact that the combination of features does not necessarily improve over the use of SIMON separately.

Next, a cross-dataset experiment has been performed, with the aim of studying how the idiosyncrasy of the data affect the generalization performance. Table 9 shows the results of this experiment, where the rows indicate the training dataset and the columns the test dataset.

As expected, the performance drop is relatively high. The differences in the datasets easily explain this. For example, a model trained in the pro-neu data suffers a performance decrease when predicting on pro-anti data; as seen, the non-radical class in the first leans towards the neutrality, while in the second it is composed of anti-radical messages. These differences affect cross-dataset performance. Additionally, when attending to the pro-neu/magazines experiment, we see that in general, the performance is higher than in the pro-neu/pro-anti case. Similarly, this can be explained by the observation that the magazines dataset is also composed of radical and neutral texts, which aligns with the pro-neu data. Also, it can be seen that the peculiarities of the magazines data make it difficult for the classifiers to generalize, especially to the pro-anti data.

An additional consideration can be made when comparing the SIMON + EmoFeat model against the isolated SIMON and EmoFeat counterparts. It can be seen that in many cases combining these features lead to an improvement of the F-Score (see pro-anti/pro-neu and pro-anti/magazines results). In light of this, we can assume that under certain circumstances, the feature combination can be used to augment the robustness of the proposed method. Nevertheless, there are cases where such a combination does not improve but decreases the performance. This result can be observed in the cases where the training is done with the magazines data and the testing on the twitter data. Based on this observation, we hypothesize that the learning process done with the magazines is not applicable to the Twitter domain. This is to be expected since they represent two very different domains, and this last experiment is a cross-dataset evaluation.

VI. DISCUSSION

When comparing the presented method with previous works, we can attend to two different criteria: the method, and the affect model (Table 1). In the sense of the

method, some works do not make use of machine learning techniques to assess the radicalization of texts [11], [14], [16], [18], [19], [26].

In contrast, many of the related works include some machine learning techniques to study radicalization. In this regard, our method uses two popular machine learning algorithms, logistic regression and Support Vector Machine (SVM). Some works also use these or similar learning algorithms, such as [1], [2], [9], [15], [20], [22], [24]. The main difference with our work can be found in the feature extraction phase. The proposed method exploits a word embedding model and a domain knowledge lexicon to calculate a set of features that are fed to the learning algorithm. To our knowledge, this type of approach has not been proposed before in radicalization assessment.

When attending to the affect model described in previous work (Table 1), it can be observed that there is a trend of using lexicons. Indeed, our proposal also makes use of such kind of resource. Nevertheless, previous work exploit the lexicons by directly matching words in the analyzed text to the ones contained in the lexicon [14], [16], [18], [19]. In this way, out-of-vocabulary terms can not be properly modeled by these approaches. In contrast, our model can handle terms that are not in the lexicon by computing the similarity of the analyzed text to the lexicon, and thus taking advantage of the wide vocabulary contained in word embedding models [7].

An additional difference with the body of previous works is that our model uses three different sources of knowledge to perform the radicalization detection. Those are the emotion lexicon, the domain-adapted radicalization lexicon, and the word embeddings model. It is relevant to stress that the combination of both an affect and radicalization lexicon is novel, and has not been previously studied in related work.

In related literature authors regularly use sentiment rather than emotion signals (Sect. II). This largely limits the capacity of the models, as taking into account the emotions offers a more detailed categorization of the affect information. In addition, we implement a distributed emotion representation, which not only considers the most prominent emotion but rather exploits all emotions simultaneously.

It is also worth noting that although embeddings have been well known by the NLP community, its role in radicalization detection is not as extended as in other fields. Still, some recent works incorporate embeddings to their learning models. In [52], the authors explore the use of sentence embeddings for classifying texts into *radical* and *benign*, representing the text directly with the Doc2Vec model [65]. A more recent work [66] also makes use of word embeddings, using Word2Vec vectors to represent the text. Such representations are combined with psychological and behavioral features to perform radicalization detection. Our work computes a similarity measure of the analyzed text to a domain lexicon. For the generation of this domain-adapted lexicon, we propose a simple but effective approach, *FreqSelect*. Such a method generates a selection of words relevant to the domain that is extracted from the datasets.

Previous research has studied the Twitter datasets used in this work, which allows us to compare the performance of our approach to those that use the same datasets. The Pro-Neu was used in [28] where the authors present a two-dimensional model for classifying users radical and non-radical users, reporting 90.60% as F-Score. In comparison, the proposed emotion model reaches 91.07%, while the full model (EmoFeat and SIMON) yields 99.11%. In [28], the authors present a model that computes a two-dimensional metric indicating radicalization from individual and social influence. In contrast, the model presented in this paper uses a richer representation, combining semantic similarity and a distributed emotion representation.

When attending to the Pro-Anti dataset, the comparison is made against the work presented in [2]. This last work introduces a learning model that makes use of 5 different types of information sources, including the inferred knowledge from the semantic graph, contextual network information, and entity knowledge. When using only sentiment information through the SentiStrength lexicon [25], the authors report an F-Score of 85.40%. Our model, in comparison, outputs 87.90%, indicating the effectiveness of the proposed model. If attending to the full system, which includes semantic pattern detection and network features, their score raises to 92.30%. This is to be expected, since extracting contextual information (e.g., network information) is beneficial for classification tasks [67]. Nevertheless, exploiting the social context falls outside the scope of this paper.

The lack of resources related to radicalism has been a challenge as well as a limitation for conducting this research. Regarding social networks, we have taken advantage of public datasets provided by researchers and Kaggle, since radical accounts are banned. Nevertheless, this is a limitation of our study, since radical messages are in continuous evolution. In the same way, we have provided a cured dataset of radical texts that have been balanced with standard newspapers. Still, a broader selection of sources could be done to incorporate additional sources.

VII. CONCLUSION

This work proposes a machine learning model that exploits two kinds of information sources: (i) emotion and (ii) embedding-based semantic similarity features. The first approach has been designed to take advantage of an existing emotion dictionary in order to generate features that can be used for radicalization detection. Such an approach generates simple but effective features by computing a statistical summary of the emotion valence of the words in the text. As for the second approach, the SIMON model has been adapted to extract radicalization-oriented features. In this work, a method for generating a domain word collection is proposed (*FreqSelect*), which leads to consistent improvements in comparison with other existing collections. Also, as an additional contribution, the Magazines dataset is presented, which expands the scope of this and future research. Intending to conduct a thorough and comparable

evaluation, we use this new dataset, as well as other two existing ones that have been previously studied. As a further study, a statistical test verifies the conclusions drawn from the experiments. Besides, a cross-dataset evaluation is also done.

Previously, we presented three research questions that were inherent to the task of radicalization detection (Sect. I). In the first one, RQ1, we centered our attention on how emotion information can be used for radicalization detection. In this sense, a novel approach that exploits the presence of emotions in the text has been presented. As seen in the experiments, these emotion-based features can yield interestingly high scores, reaching an F-Score of 92.41% in a known dataset. It is reasonable to assume that these emotional features can be useful when performing radicalization detection. Also, results from the cross-validation experiments indicate that the proposed combination can help with generalization issues. As seen in Table 9, combining the emotion features in a cross-dataset setting can improve the final performance, resulting in a more robust system.

In relation to RQ2, it is asked whether semantic similarity features can be used in radicalization detection. For addressing this issue, we have applied an original feature extraction method (SIMON) to the radicalization detection domain. In order to do this, a novel method has been presented that generates a domain-oriented vocabulary, which can be used by the SIMON model. Given the experimental results, it is clear that this kind of features can accurately represent the analyzed text, allowing simple classifiers to obtain high classifications metrics: 99.11%, 86.22%, and 94.02% for the pro-neu, pro-anti and magazines datasets, respectively. Thus, given these promising results, we conclude that the application of this method can be further studied.

Following, RQ3 raises attention around the challenge of generating and using domain-oriented vocabularies. As presented, we proposed a method that effectively generates this kind of vocabularies: FreqSelect. Moreover, we have evaluated the use of this vocabulary, comparing it with similar selections. Seeing the results, we believe that such an approach performs positively.

To summarize, this paper presents a machine learning approach for categorizing radical content in both social media and magazines, exploiting an emotion dictionary, a word embeddings model, and a domain-adapted word selection. The presented advancement in automatic identification of online radicalization can be beneficial to counter-extremist agencies, since these organizations have as a key priority the use of intelligent technologies in this direction.

Although the obtained results are very positive, we believe that as future work, the effect of word embeddings can be further studied. That is, obtaining domain-oriented word vectors could largely improve the text representations. Besides, studying the effect of temporal effects on the radical narratives can be interesting, since the use of the language is prone to change as time progresses.

ACKNOWLEDGMENT

The authors would like to thank the Trivalent Team, specially M. Fernandez, H. Alani, R. Denaux, J. M. Gomez-Perez, R. Barbado, and D. Suarez for their help in this research.

REFERENCES

[1] A. Bermingham, M. Conway, L. Mcinerney, N. O'hare, and A. F. Smeaton, "Combining social network analysis and sentiment analysis to explore the potential for online radicalisation," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2009, pp. 231–236.

[2] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, "A semantic graph-based approach for radicalisation detection on social media," in *The Semantic Web* (Lecture Notes in Computer Science), vol. 10249, E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, Eds. Cham, Switzerland: Springer, 2017.

[3] J. Van Stekelenburg, "Radicalization and violent emotions," *APSC*, vol. 50, no. 4, pp. 936–939, Oct. 2017.

[4] S. K. Rice, "Emotions and terrorism research: A case for a social-psychological agenda," *J. Criminal Justice*, vol. 37, no. 3, pp. 248–255, May 2009.

[5] R. Pelzer, "Policing of terrorism using data from social media," *Eur. J. Secur. Res.*, vol. 3, no. 2, pp. 163–179, Oct. 2018.

[6] C. Mccauley and S. Moskalenko, "Mechanisms of political radicalization: Pathways toward terrorism," *Terrorism Political Violence*, vol. 20, no. 3, pp. 415–433, Jul. 2008, doi: 10.1080/09546550802073367.

[7] O. Araque, G. Zhu, and C. A. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis," *Knowl.-Based Syst.*, vol. 165, pp. 346–359, Feb. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118305926

[8] G. Zhu and C. A. Iglesias, "Exploiting semantic similarity for named entity disambiguation in knowledge graphs," *Expert Syst. Appl.*, vol. 101, pp. 8–24, Jul. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418300897

[9] A. Abbasi and H. Chen, "Affect intensity analysis of dark Web forums," in *Proc. IEEE Intell. Secur. Informat.*, May 2007, pp. 282–288.

[10] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, vol. 10, 2010, pp. 2200–2204.

[11] M. Rowe and H. Saif, "Mining pro-isis radicalisation signals from social media users," in *Proc. 10th Int. AAAI Conf. Web Social Media (ICWSM)*, 2016, pp. 329–338.

[12] L. Deng and J. Wiebe, "MPQA 3.0: An entity/event-level sentiment corpus," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 1323–1328.

[13] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale arabic sentiment lexicon for arabic opinion mining," in *Proc. EMNLP Workshop Arabic Natural Lang. Process. (ANLP)*, 2014, pp. 165–173.

[14] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur. Informat.*, vol. 4, no. 1, p. 9, 2015.

[15] H. Chen, "Sentiment and affect analysis of dark Web forums: Measuring radicalization on the Internet," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jun. 2008, pp. 104–109.

[16] M. Vergani and A.-M. Bliuc, "The evolution of the ISIS' language: A quantitative analysis of the language of the first year of Dabiq magazine," *Sicurezza, Terrorismo e Societa Secur., Terrorism Soc.*, vol. 2, no. 2, pp. 7–20, 2015.

[17] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway, Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001. 2001.

[18] T. Chalothorn and J. Ellman, "Affect analysis of radical contents on Web forums using sentiwordnet," *Int. J. Innovation Manage. Technol.*, vol. 4, no. 1, pp. 122–124, 2013.

[19] S. Ghajar-Khosravi, P. Kwantes, N. Derbentseva, and L. Huey, "Quantifying salient concepts discussed in social media content: A case study using Twitter content written by radicalized youth," *J. Terrorism Res.*, vol. 7, no. 2, p. 79, May 2016.

[20] P. Dewan, A. Suri, V. Bharadhwaj, A. Mithal, and P. Kumaraguru, "Towards understanding crisis events on online social networks through pictures," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*. New York, NY, USA: ACM, 2017, pp. 439–446.

[21] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*. New York, NY, USA: ACM, 2013, pp. 459–460.
[22] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting Jihadist messages on Twitter," in *Proc. Eur. Intell. Secur. Informat. Conf.*, Sep. 2015, pp. 161–164.
[23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. Mcclosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.
[24] S. Agarwal and A. Sureka, "Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter," in *Distributed Computing and Internet Technology* (Lecture Notes in Computer Science), vol. 8956, R. Natarajan, G. Barua, and M. R. Patra, Eds. Cham, Switzerland: Springer, 2015.
[25] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social Web," *J. Amer. Soc. Inf. Sci.*, vol. 63, no. 1, pp. 163–173, Jan. 2012.
[26] D. López-Sáncez, J. Revuelta, F. de la Prieta, and J. M. Corchado, "Towards the automatic identification and monitoring of radicalization activities in Twitter," in *Knowledge Management in Organizations* (Communications in Computer and Information Science), vol. 877, L. Uden, B. Hadzima, and I. H. Ting, Eds. Cham, Switzerland: Springer, 2018.
[27] D. Correa and A. Sureka, "Solutions to detect and analyze online radicalization: A survey," Jan. 2013, *arXiv:1301.4916*. [Online]. Available: https://arxiv.org/abs/1301.4916
[28] M. Fernandez, M. Asif, and H. Alani, "Understanding the roots of radicalisation on Twitter," in *Proc. 10th ACM Conf. Web Sci. (WebSci)*. New York, NY, USA: ACM, 2018, pp. 1–10, doi: 10.1145/3201064.3201082.
[29] S. Agarwal and A. Sureka, "Topic-specific YouTube crawling to detect online radicalization," in *Databases in Networked Information Systems* (Lecture Notes in Computer Science), vol. 8999, W. Chu, S. Kikuchi, and S. Bhalla, Eds. Cham, Switzerland: Springer, 2015.
[30] E. Ferrara, W. Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Social Informatics* (Lecture Notes in Computer Science), vol. 10047, E. Spiro and Y. Y. Ahn, Eds. Cham, Switzerland: Springer, 2016.
[31] S. Agarwal and A. Sureka, "Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats," Nov. 2015, *arXiv:1511.06858*. [Online]. Available: https://arxiv.org/abs/1511.06858
[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[33] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proc. 2015 Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 298–307.
[34] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, Mar. 2015.
[35] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Syst. Appl.*, vol. 77, pp. 236–246, Jul. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417300751
[36] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. The 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–17.
[37] C. Van Hee, E. Lefever, and V. Hoste, "SemEval-2018 task 3: Irony detection in English tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 39–50.
[38] K. Gábor, D. Buscaldi, A.-K. Schumann, B. Qasemizadeh, H. Zargayouna, and T. Charnois, "SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 679–688.
[39] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," Jul. 2016, *arXiv:1607.01759*. [Online]. Available: https://arxiv.org/abs/1607.01759
[40] O. Araque, G. Zhu, M. García-Amado, and C. A. Iglesias, "Mining the opinionated Web: Classification and detection of aspect contexts for aspect based sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 900–907.
[41] S. Lai, K. Liu, S. He, and J. Zhao, "How to generate a good word embedding," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 5–14, Nov./Dec. 2016.
[42] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl.-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.
[43] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Meta-level sentiment models for big social data analysis," *Knowl.-Based Syst.*, vol. 69, pp. 86–99, Oct. 2014.
[44] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!" in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 538–541.
[45] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the State-of-the-art in sentiment analysis of tweets," Aug. 2013, *arXiv:1308.6242*. [Online]. Available: https://arxiv.org/abs/1308.6242
[46] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.-Assoc. Comput. Linguistics*, vol. 1, 2011, pp. 151–160.
[47] C. Zirn, M. Niepert, H. Stuckenschmidt, and M. Strube, "Fine-grained sentiment analysis with structural features," in *Proc. 5th Int. Joint Conf. Natural Lang. Process.*, 2011, pp. 336–344.
[48] H. K. Gambhir, "DABIQ: The strategic messaging of the Islamic State," *Inst. Study War*, vol. 15, pp. 1–12, Nov. 2014.
[49] R. Mahzam, "Rumiyah–jihadist propaganda & information warfare in cyberspace," *Counter Terrorist Trends Analyses*, vol. 9, no. 3, pp. 8–14, 2017.
[50] N. A. Azman, "'Islamic State' (IS) propaganda: Dabiq and future directions of 'Islamic State,'" *Counter Terrorist Trends Analyses*, vol. 8, no. 10, pp. 3–8, 2016. [Online]. Available: http://www.jstor.org/stable/26351457
[51] P. Wignell, S. Tan, K. O'Halloran, and R. Lange, "A mixed methods empirical examination of changes in emphasis and style in the extremist magazines dabiq and rumiyah," *Perspect. Terrorism*, vol. 11, no. 2, pp. 2–20, 2017.
[52] A. H. Johnston and G. M. Weiss, "Identifying sunni extremist propaganda with deep learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov./Dec. 2017, pp. 1–6.
[53] J. S. Kessler, "Scattertext: A Browser-based tool for visualizing how corpora differ," Mar. 2017, *arXiv:1703.00565*. [Online]. Available: https://arxiv.org/abs/1703.00565
[54] B. S. Iskandar, "Terrorism detection based on sentiment analysis using machine learning," *J. Eng. Appl. Sci.*, vol. 12, no. 3, pp. 691–698, 2017.
[55] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[56] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for Twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. for Comput. Linguistics (Long Papers)*, vol. 1, 2014, pp. 1555–1565.
[57] M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on Twitter," in *Proc. Workshop Semantic Web Social Good (SW4SG), 17th Int. Semantic Web Conf. (ISWC)*, Monterey, CA, USA, Oct. 2018. [Online]. Available: http://ceur-ws.org/Vol-2182/
[58] O. Araque, M. Guerini, C. Strapparava, and C. A. Iglesias, "Neural domain adaptation of sentiment lexicons," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Oct. 2017, pp. 105–110.
[59] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Jul. 2016, *arXiv:1607.04606*. [Online]. Available: https://arxiv.org/abs/1607.04606
[60] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[61] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, 2015, doi: 10.1111/coin.12024.
[62] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, 2004, pp. 168–177.
[63] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
[64] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[65] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[66] M. Nouh, J. R. C. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on Twitter," May 2019, *arXiv:1905.08067*. [Online]. Available: https://arxiv.org/abs/1905.08067

[67] J. F. Sánchez-Rada and C. A. Iglesias, "Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison," *Inf. Fusion*, vol. 52, pp. 344–356, Dec. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253518308704

OSCAR ARAQUE received the graduate and master's degrees in telecommunication engineering from the Technical University of Madrid (Universidad Politécnica de Madrid), Spain, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree. He is currently a Teaching Assistant with the Universidad Politécnica de Madrid. His research interest includes the application of machine learning techniques for natural language processing. The main topic of his thesis is the introduction of specific domain knowledge into machine learning systems in order to enhance sentiment and emotion analysis techniques.

CARLOS A. IGLESIAS received the telecommunications engineering degree and the Ph.D. degree in telecommunications from the Universidad Politécnica de Madrid (UPM), Spain, in 1993 and 1998, respectively. He is currently an Associate Professor with the Telecommunications Engineering School, UPM, where he has been leading the Intelligent Systems Group, since 2014. He has been the Principal Investigator on numerous research grants and contracts in the field of advanced social and the IoT systems, funded by the regional, national, and European bodies. His primary research interests include social computing, multiagent systems, information retrieval, sentiment and emotion analysis, linked data, and web engineering.