IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 October 2023, accepted 8 October 2023, date of publication 20 October 2023, date of current version 1 November 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3326429*

RESEARCH ARTICLE

# Contextualization of a Radical Language Detection System Through Moral Values and Emotions

**PATRICIA ALONSO DEL REAL, (Member, IEEE), AND OSCAR ARAQUE**
Intelligent Systems Group, 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 962547.

**ABSTRACT** The popularity of current communication technologies has boosted the spread of polarization and radical ideologies, which can be exploited by terrorist organizations. Building upon previous research, this work focuses on the task of automatic radicalization detection in texts using natural language processing and machine learning techniques. In this way, we investigate the effectiveness of integrating moral values through the Moral Foundations Theory (MFT). Moral values play a crucial role in identifying ideological inclinations and can have a significant impact on the radicalization detection task. Our approach distinguishes itself in the feature extraction stage, leveraging moral values, emotions, and similarity-based features that utilize word embeddings. Additionally, we thoroughly evaluate the proposed representations with three distinct datasets that model radicalization and use the SHAP method to gain relevant insight into the models' reasoning.

**INDEX TERMS** Machine learning, moral values, natural language processing, radicalization detection, sentiment analysis.

## I. INTRODUCTION

The proliferation of social media and the internet in recent times has accelerated the pace and lowered the expenses associated with information sharing. As a consequence, this has facilitated numerous terrorist organizations to restructure themselves in order to amplify their capabilities to operate autonomously and cause significant harm to individuals, communities, and nations [21]. The spread of extremist beliefs and ideologies can result in violent actions, hate crimes, and social unrest [25]. In this context, the detection of radical propaganda is crucial for preventing radicalization and promoting social harmony. Numerous global organizations and nations have formulated tactics and initiatives to counteract radicalization through social and computational text analysis [79]. The task of performing thorough manual inspections is impractical with the vast amount of text and relationships between information and individuals. Therefore, the development of computational techniques for detecting, analyzing, and preventing radicalization and extremism is essential. In fact, such automated systems need to be constantly updated, since individuals and communities can escape detection by usual means [80].

There are many research studies that investigate the development of efficient automated approaches for detecting extremism [4]. An important illustration of this is the present emphasis of counter-terrorism organizations and governments on automatically detecting extreme language on social media in their attempt to combat extremist social network groups. These works contemplate that by developing information technology systems that can recognize extremist content it would be possible to combat online radicalization [16]. For example, there are numerous works that tackle the study of hate speech in relation to radical content and terrorist organizations, and how terrorist attacks lead to the propagation of hateful language [21].

Another relevant research direction is the study of social networks, that has as basis that the social context influences an individual behaviour. One of these studies addresses the intervention of networks [58], which can be a powerful ally in the fight against radicalization. Another perspective consists in predicting links within a network [5]. While

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

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

radical behaviour is indeed interesting for research, there are works that have analyzed the used case of Islamist terrorist networks [56]. Such studies offer implications that can be used by policymakers and Law Enforcement Agencies (LEAs). For example, it has been found that Islamist networks are resilient and resistant to node removal. Additionally, it is interesting to observe that new studies aim to combine social network analysis and Natural Language Processing (NLP) into a unified framework [53].

There is a growing tendency to utilize data mining approaches, such as machine learning, to investigate relevant challenges related to extremist content [32]. Unfortunately, the efficient implementation of intelligent methods for counter-terrorism remains an ongoing issue.

In an effort to improve current radicalization detection methods, this work explores the effect of moral foundations in the radicalization process. Moral foundations are basic psychological systems that represent the shared bedrock of society's ethics. Many social-ideological differences within a country may stem from variations in the definitions of morality, as suggested by the evidence. For example, according to [39], liberals and conservatives in the US have unique perspectives on the social environment and depend on differing moral frameworks and ideologies. With this in mind, this work aims to contextualize textual representations in learning systems with moral foundations information.

Additionally, and following previous work [7], this work further complements the contextualization effort by means of adding emotion-aware representations. The significance of emotions in comprehending terrorism has been emphasized in recent literature [67], [77].

This investigation is based on two main research questions (RQ):

**RQ1: Does a moral foundations approach help in the task of detecting radicalization? If so, how?** Previous research shows that using moral values to radicalize groups of people is a tendency in terrorist cells [54]. When a group collectively agrees on certain values and morals, it can wield significant influence, including the ability to legitimize and sometimes mandate violence against individuals who are a threat to them. Thus, this work studies the effect of incorporating moral values information through the comparison of two learning models for the task of radicalization detection.

**RQ2: Can emotion and moral values in combination with embedding-based similarity features be used effectively for propaganda detection?** As said, moral foundations and emotions can be useful tools when classifying text to detect radicalization. Word embeddings have also been utilized alone and in combination with other approaches for the same task [7]. With this information, this work analyzes whether the effect of merging these techniques can have a positive outcome in the overall classification performance.

In light of these questions, this paper proposes a NLP system that can identify propaganda in text. For this purpose, the designed system produces text representations that are exploited by a machine learning classifier. These representations are obtained from a variety of information sources: moral values and emotion lexicons, and the SIMilarity-based sentiment projectiON (SIMON) model [11]. Alternatively, we also study simpler text representation baselines such as unigrams, bigrams, and TF-IDF vectors. A thorough assessment has been conducted on three pertinent datasets in order to evaluate the efficiency of this broad range of features. Furthermore, a computational technique that utilizes word frequency distributions to derive a domain-specific vocabulary has been used (*FreqSelect*) [7], which allows us to capture the vocabulary of a specific domain.

More concretely, the contributions of this work are fourfold. Firstly, we propose and implement (i) the use of moral values through a computation model of the Moral Foundations Theory (MFT) for radicalization detection. We argue that moral cues can aid automatic systems in identifying radical content, thus offering a more complete view of the radicalization process in language. Following the previous reasoning, this paper also proposes (ii) the generation of a unified representation for radicalization detection, composed of moral values, emotions, and distributed representations through word embeddings. In order to assess the effectiveness of such representations, we perform a (iii) thorough evaluation using three distinct datasets that address radical language. Finally, the paper (iv) explores a qualitative assessment of the obtained models through the computation SHapley Additive exPlanations (SHAP) values [50]. This allows us to better understand the phenomenon of radicalization in language, offering explanations on word usage and their impact on predictions.

The rest of the article is structured in the following manner. Section II provides an outline of the related work by presenting a summary of the Moral Foundations Theory (MFT) and its application for detecting radicalization, as well as the utilization of emotion lexicons and word embeddings for this task. Section III presents the proposed system and its architecture. Then, Section IV describes the evaluation of the system, including the materials used, the methodology followed, and the results obtained. Finally, Section VI illustrates the conclusions derived from the performed analysis.

## II. RELATED WORK

This section reviews related work in the field of online radicalization analysis and detection. Firstly, Sect. II-A presents the background and similar works when addressing the Moral Foundations Theory. Secondly, Sect. II-B describes previous works on the use of distributed text representations and emotion resources.

### A. MORAL FOUNDATIONS THEORY

Following the line of reasoning that states that perspectives coming from multiple levels of analysis should be acquired prior to understanding the social pattern of radicalization [27], the approach that this work proposes uses the study

line related to morality by analyzing the performance of two different lexicons as tools for classifying the input data based on the Moral Foundations Theory [38]. This psychological theory aims to model the differences in morality across different cultures while also highlighting the presence of resemblances and recurring patterns. According to the theory, there are certain psychological systems that are innate and universally accessible. These foundational elements are then built upon by each culture through the development of virtues, narratives, and institutions. The foundations that have been taken into account in this work are 'care/harm', 'fairness/cheating', 'loyalty/betrayal', 'authority/subversion' and 'sanctity/degradation'.

Other studies have also been developed using the Moral Foundation Theory for investigation in diverse fields. One of the most representative tools used by computation approaches is the Moral Foundations Dictionary (MFD) [39], which allows researchers to assess moral foundations by means of a lexical resource based on the Linguistic Inquiry and Word Count (LIWC) framework [65]. In [39], four studies using four different methods were carried out in order to develop a theory that states the psychological foundations upon which political groups build their moral codes. In more recent work, the MFD was utilized to conduct a manual analysis of 12 years' worth of coverage in the New York Times, with a focus on political discussion in the United States [22]. Reference [76] analyzed the discourse on the potential exit of Greece from the European Union through the examination of approximately 8,000 tweets related to the topic. A comparison was made between the effectiveness of basic machine learning models, such as Maximum Entropy (ME) and Naive Bayes (NB), in utilizing unprocessed MFD features. Both studies came to the conclusion that machine learning in its pure form is more desirable than dictionary-based approaches as it achieves similar predictive accuracy with fewer assumptions.

Another study [28] employed a Latent Dirichlet Allocation (LDA) model to explore the existing variations between conservative and liberal ethical codes. This facilitated the unsupervised identification of topics related to morality. In [69], the authors applied the same structure to investigate moral argumentation in text, focusing on the US Federal shutdown of 2013. The study analyzed the influence of morals on intra- and inter-community disparities in political party retweets [70]. In addition, [46] proposes Latent Semantic Analysis (LSA) for representing the moral values of text through a model that uses a multiset of words in order to calculate a co-occurrence matrix, and subsequently, word vectors are extracted from it.

There are also some recent studies where the MFD has been utilized to identify moral values in lengthy political speeches over a period of time, such as in [34]. Likewise, [1] presented a technique called Distributed Dictionary Representations (DDR), which involves merging psychological dictionaries and semantic similarity to assess the prevalence of moral rhetoric on a particular subject. This technique has been utilized in other studies with the aim of identifying morals in the donation to charity [42] and also to include demographic embeddings within the language representations by expanding it [35]. Another study where the MFD and the DDR technique are combined and the *MoralStrength* lexicon is taken into account is [81], where the moral divergence between candidates from the Republican and Democrat parties are analyzed through the quantification of presidential debaters' moral judgments. In the line exploiting a lexical resource that models morality, in [37] the authors use word embeddings and an extended version of the MFD to assign morality orientations to movie synopses. It is indeed insightful to assess moral values from text, as shown by a study that assesses moral concerns on language, indicating that language usage is related to morality [47]. The presented work is related to the mediatization of opinions, which can lead to their polarization.

In this work we explore the performance of two different lexicons coming from distinct works: the *MoralStrength* [6] and *eMFDscore* [43] lexicons. The *MoralStrength* resource expands the original MFD. After an initial preprocessing step was conducted on the word corpus obtained, wherein forms that matched the search but did not relate to a moral trait were eliminated. This procedure was carried out manually, taking into account both the gloss for the lemma provided by WordNet [61] and the moral trait associated with that word. Then, a division was done separating the obtained word corpus in ''virtue'' and ''vice'' lemmas so that an association strength between word and moral trait can be provided with value ranges from 1 to 9 (1 for words commonly linked to vices and 9 for words commonly linked to virtues).

The MoraStrength lexicon can be used in a variety of applications, such as in [57], where the authors monitorize the responses to the mask mandate due to COVID-19 and analyze the moral values behind each argument proposed, as well as the political leaning of people with one opinion or the other. This is based on the conceptual perspective of Moral Foundation Theory and Hofstede's cultural aspects [41].

The other resource has also been used in this work is the *eMFDscore* lexicon [43]. With this approach, every word is assigned both foundation probabilities, that indicate the likelihood that each word is linked with each one of the five moral foundations, and sentiment scores, that reflect the average sentiment of the context related to the moral foundation where each word is used.

### B. WORD EMBEDDINGS AND EMOTION LEXICONS

Word embeddings and emotion lexicons are NLP methods that have been previously used for detecting radicalism in natural language processing. Word embeddings represent words as vectors in high-dimensional spaces [60]. This approach can be useful for capturing the meaning and context of words within a given text corpus. On the other hand, emotion lexicons are dictionaries that label a certain vocabulary with affective dimensions, such as joy or fear [66].

There are a number of previous works that use NLP techniques to assess radicalization, as we do in this work. As an application example, LEAs can use such systems in their process of making decisions. In [23], authors explain that solutions resulting from analysis can be categorized into two primary types: network-based and content-based. The first type concentrates on virtual communities and graph characteristics, while the second deals with online behavior, linguistic style analysis, authorship identification, emotion analysis and usage mining. As a reference, this work falls into the second category.

A common approach for many NLP works is to utilize bag-of-words text representations in classification tasks. More refined features are word embeddings, which are unsupervised text representation models that can be used for subsequent learning tasks [59]. Continuing this research direction, various works address the incorporation of word embeddings as a component of feature extraction techniques for text categorization. In [83], word embeddings are utilized to compute features that are used by a Support Vector Machine (SVM) classifier in order to identify polarity in texts. This last work shows that word embeddings may encompass semantic knowledge among words, thereby enabling the acquisition of valuable text representations. In [64], semantic meanings are captured by a similar approach, using the Word2Vec model to create a system that can differentiate whether a Twitter message endorse the Islamic State of Iraq and Syria (ISIS) or not. Similarly, the authors of [74] examine the efficacy of utilizing word embeddings for sentiment analysis and provide a summary of unsupervised embedding methods and their potential in obtaining text representations. In addition, methods based on this approach are widely employed in open competitions, where participants strive to get the highest scores in a wide range of tasks [31], [40], [62]. Another interesting work [44] proposes the use of word embeddings within a novel framework designed to minimize computational complexity and demonstrates comparable evaluation metrics to those of more intricate neural models across multiple tasks. Besides, in [10], a merge of semantic similarity and word embeddings approaches is presented.

Recently, the detection of hate speech in a forum dedicated to white supremacy has been explored [26]. The presented model has been trained and evaluated on a balanced subset of a dataset containing roughly 2,000 sentences sourced from the Stormfront forum. Learning models like SVM, Long-short Term Memory (LSTM), and Convolutional Neural Network (CNN) were utilized to identify hate speech. One significant constraint of the mentioned study is that it involves annotating sentences taken from paragraphs without any extra information that could aid in the comprehension of the context of the sentences for precise labeling.

Delving into affect analysis for radicalization analysis, it has been exploited in different fields such as radical forums [2], extremist magazines [78], and social media platforms like Twitter [68]. In this sense, there are different ways challenges that have been explored. Several studies utilize the polarity of sentiment analysis (e.g., valence) [15], [71]. In other cases, the strength of the vocabulary related to hatred and aggression is used [2]. Besides, LIWC [75] categories are used to extract information related to positive and negative emotions as well as sadness, anxiety, and anger.

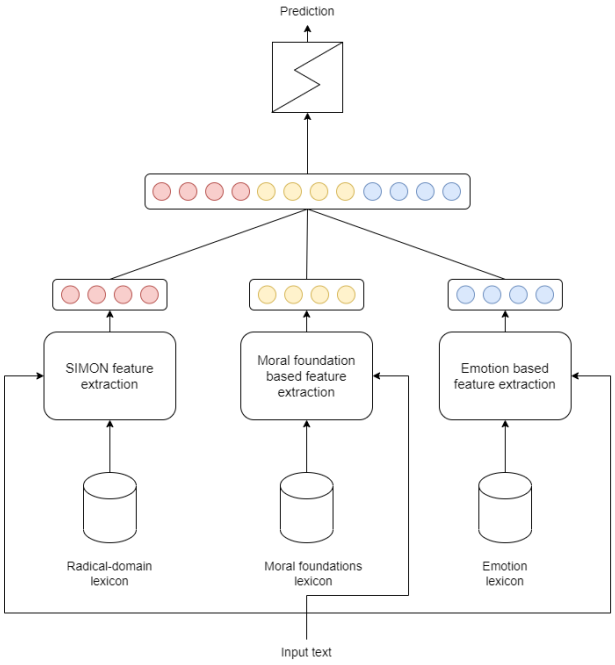


**FIGURE 1.** Diagram of the emotion feature extraction approach.

Additionally, there is an interesting research direction that tracks the response of the general public to a terrorist attack [29]. This last work conducted a sentiment analysis on both texts and images that were extracted from Facebook and noticed that while the sentiment was initially negative in the first few hours, it gradually shifted toward a positive valence as time passed. This happened with text, but for images, the opposite effect took place.

Data extracted from social media and magazines has been analyzed with the aim of detecting radicalization by merging embedding and lexicon approaches [7], [8]. The obtained results ratified the potential that the combination of these techniques has in this kind of text classification, as the F1 score increased in comparison to its value when only using one of the methods.

## III. CONTEXTUALIZING TEXT REPRESENTATIONS FOR RADICALIZATION DETECTION

This work proposes a machine learning model that exploits the combination of a variety of methods that connect different study lines. In this way, the proposed model is composed of three sub-modules: (i) morality, (ii) emotion-based, and (iii) embedding word similarity text representations. An illustration of the proposed model is shown in Figure 1, where it can be seen that the three sub-modules extract their own features from the natural language input. Each feature set

**Algorithm 1** MoralStrength Feature Extraction

**Require:** Moral lexicon composed by a vocabulary $T^{(s)}$ and annotations $\boldsymbol{A}$
**Ensure:** $\boldsymbol{v} \in IR^{n \cdot m}$, the final feature vector
**for** $i \leftarrow 1, n$ **do**
  **for** $j \leftarrow 1, m$ **do**
    **for all** $t_k \in T^{(s)} \cap T^{(i)}$ **do**
      $\boldsymbol{M}_{k,:} \leftarrow$ moralAnnotation$(t_k, \boldsymbol{A})$
      $\boldsymbol{S} \leftarrow$ sum$(\boldsymbol{M}_{k,:})$
      $\boldsymbol{D} \leftarrow \boldsymbol{S}/size(T^{(s)} \cap T^{(i)})$
      $\boldsymbol{v} \leftarrow$ append(D)
    **end for**
  **end for**
**end for**

is represented as a vector in the model output. These vectors are concatenated to construct a unified representation that is later fed to a machine learning classifier, which generates a prediction. Regarding the classifier algorithms, Logistic Regression and a linear Support Vector Machine model have been used.

### A. MORAL FOUNDATION BASED FEATURES

Propaganda often appeals to people's moral values in order to persuade them to adopt a certain belief or point of view [24]. By analyzing the moral foundations present in text, we hypothesize that tactics used to manipulate people can be identified. For example, propaganda may use language that evokes feelings of loyalty and patriotism in potential recruits to motivate them to support a specific political program. On this basis, we propose the use of two resources to extract different features regarding morality: the *MoralStrength* [6] and *eMFDscore* [43] lexicons.

In *MoralStrength*, the procedure has several steps. First, the association strength between each word and a certain moral foundation is obtained. Such strengths are modelled as numeric values between 1 and 9. Then, for each text, the acquired moral values for each foundation are summed and then divided by the number of words that appear in the lexicon. This process is repeated for each of the five moral foundations, so as result, a matrix of $n$ rows and five columns is obtained, being $n$ the number of texts in the dataset and $m$ the number of moral foundations. $T^{(i)}$ is the input text of a single instance of the dataset. Algorithm 1 presents the pseudo-code that expresses the logic of the model.

In *eMFDscore*, the resulting data includes, among others, moral foundation probabilities, which indicate the mean probability of each text's association with any of the five moral foundations. They are obtained by counting the frequency of a word's annotation with a particular foundation and then dividing it by the total number of times the word was seen by annotators with this foundation assigned. Other metrics are included in the result that this method offers, but only the one mentioned is used so that it can be compared

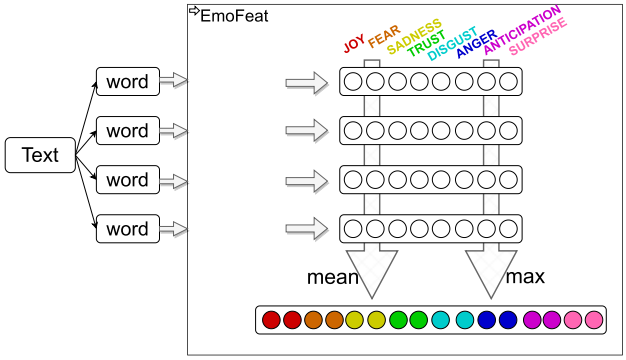


**FIGURE 2.** Diagram of the emotion feature extraction approach.

to *MoralStrength*. As a result, the *eMFDscore* outputs a representation vector of five components, one for each moral foundation.

### B. EMOTION BASED FEATURES

Radical speech and extremist discourses can entail the use of emotionally charged language and rhetoric with the aim of manipulating and radicalizing potential members of the organization [52], [55]. This is why we consider that the study of the emotions expressed in text is a relevant step in the path to detecting radicalization. Also, sentiment analysis plays an important role in this task as it can identify the overall sentiment of a text, which can also offer an understanding of the underlying radicalization process.

In this work, we propose the utilization of an emotion lexicon to enrich the contextualization of text representations. As seen in previous research [7], the emotion-based approach can be useful when detecting extremism in text. Following this research, this work proposes the use of a lexicon-based approach that utilizes statistical measures for encoding emotional attributes within textual content.

The algorithm that makes this possible, which we will refer to as *EmoFeat (Emotion Features)*, considers an emotion lexicon formed by a group of words $W^{(l)} = w_1^{(l)}, ..., w_i^{(l)}, ..., w_P^{(l)}$ and a vector of numeric annotations $L = [l_1, ...l_i, ...l_P]$. This lexicon has an annotation $l_i$ for each term $w_i$, so there are P pairs of $(w_i, l_i)$ values. Additionally, the vector $l_i$ indicates the strength of each emotion for word $w_i^{(l)}$ in the lexicon. Besides that, as an emotion vector has dimensionality $l_i \in IR^m$, the computed emotion annotation matrix is $L \in IR^{Pm}$, where the $m$ columns indicate the number of emotions taken into account in the lexicon. Finally, $W^{(i)} = w_1^{(i)}, ..., w_j^{(i)}, ..., w_I^{(i)}$ is defined as the set of I elements formed by the input words.

Taking into consideration the intersection $W^{(l)} \cap W^{(i)}$, the related emotion vector is extracted from L for each word $w_k$. The result of this process is a matrix with emotion annotations for all the input words that exist in the lexicon. Then, different statistical metrics are employed to depict the matrix as a feature vector. This work utilizes the mean and maximum statistical metrics. Consequently, a feature vector is derived with a dimension of $n \cdot m$, where $n$ is the number of statistical functions and $m$ is the number of emotions

considered. In Figure 2, a graphical representation of the process implemented by EmoFeat is presented. As indicated, we have calculated the mean and the maximum for each of the eight emotions.

### C. EMBEDDING BASED SEMANTIC SIMILARITY

Distributed representations have become increasingly popular in NLP since they provide various benefits over more traditional approaches [82]. One of the main advantages of these models is that they can capture the richness and complexity of word semantics in a way that can not be achieved with simpler and more symbolic depictions. Word embeddings are one of the most outstanding techniques for computing distributed representations [48]. They usually involve training an unsupervised neural model to predict specific aspects of the context in which a word appears using an internal vector representation of the said word. The problem that arises here is the fact that pre-trained word embedding models' content does not entail any task-specific information as these models are trained from extensive datasets using unsupervised techniques.

Besides, there is currently an issue of data scarcity in the extremist language detection domain [7]. As a consequence, training specific word vectors in this domain does not represent an interesting direction. To avoid this limitation, we use the SIMilarity-based sentiment projectiON (SIMON) model [11]. SIMON is a feature extractor whose main proposal is the representation of a certain term that could be absent from the lexicon database. This approach involves projecting this word onto a group of sentiment words that have been extracted from a sentiment lexicon. This projection is done by making use of the semantic similarity between words by means of a word embedding model. This exploits the idea that an embedding model contains semantic and syntactic information by converting the text into vectors of a predetermined length. The input text is measured against a specific vocabulary: this outputs a vector that encodes the similarity between the input and lexicon words. In this way, the method exploits the knowledge included in both a word embedding model and a domain lexicon. This specific lexicon is extracted from the training dataset by computing the frequency of appearance of the vocabulary within the dataset. This method has been originally denoted as *FreqSelect* [7]. Its aim is to serve as a basic reference to capture the lexicon of a particular domain in a straightforward manner.

Furthermore, it is relevant to notice that the SIMON model can be used when a large corpus is not available. Some examples of the successful use of SIMON include radicalization detection [7], moral value estimation [6], and hate speech analysis [14].

## IV. EVALUATION

The assessment of radicalization has been modeled as a binary classification task, being non-radical and radical the negative and positive classes, respectively. Thus, the techniques used for this task have been developed by leveraging the datasets, embeddings and lexicons described in Section IV-A, and adhering to the methodology outlined in Section IV-B. The results of this evaluation are detailed in Section IV-C.

### A. MATERIALS

The datasets used in this study and their main characteristics are shown in Table 1. These datasets are the following.

***Pro-anti.*** This English-written dataset is the collection of tweets from 1,132 Twitter accounts done by [68] and can be divided into two groups. The first group contains 566 instances which were categorized as pro-ISIS, as they are users that share propaganda material from established pro-ISIS accounts that aim to incite or provoke. In an initial version of the dataset, there were 727 identified accounts, but 161 were found either closed down or unavailable for public access, so they were removed. The second group contains another 566 different instances extracted from anti-ISIS accounts. That is, a categorization done through the analysis of language use in accounts that oppose ISIS.

Figure 3 shows a visualization of the *Pro-anti* dataset. To avoid excessive computer costs, a random subsampling has been performed over the mentioned dataset. The visualization depicted in Figure 3 is a scatter plot that presents the frequency distribution of words in the dataset categorized as radical and non-radical. Each point on the plot represents a word, and its colour indicates its frequency in relation to each category: blue for radical and red for non-radical. To analyze the frequency within each category, the occurrence of different words is computed. Due to limited space, only a subset of word labels is displayed along the figure. The y-axis represents the frequency within the radical category, so words that frequently appear in radical texts are positioned towards the upper region of the plot.

Likewise, the x-axis represents the frequency within the non-radical category. Words that commonly appear in non-extremist texts are positioned towards the right side of the graph. Notably, the areas that help the most in this study are the top left (words frequent in radical texts), bottom right (words frequent in non-radical texts), and top right (words frequent in both neutral and radical texts) sections of the visualization as they reveal the most distinctive words associated with the neutral, radical, and overlapping categories. Examining these regions provides insights into how words are employed in these two categories. For instance, notable radical words found in the dataset include 'khilafah' and 'ummah'. The first one means 'caliphate', which denotes the position held by the leader responsible for the political affairs of the Muslim community or state, in particular during the period from 632 to 1258. 'Ummah' is the Muslim community itself. In contrast, non-radical texts frequently feature words such as 'twitterkurds' and 'kurdistan'. Kurdish people have suffered violence and injustice from the Islamic State; in fact, there exist militia groups against ISIS in this country [84], so it is not surprising that these terms are found in the non-radical category.

**TABLE 1.** Statistics of the datasets: number of instances, category balance (percentage), average number of words per instance and source.

| Dataset | No. of instances | Category balance (%) | Avg. no. of words | Source |
|---|---|---|---|---|
| Pro-anti | 1,132 | 50/50 | 36,352 | Twitter |
| Magazines | 468 | 68/32 | 950 | Magazines |
| Jacobs | 5,000 | 50/50 | 519 | Dark Web |

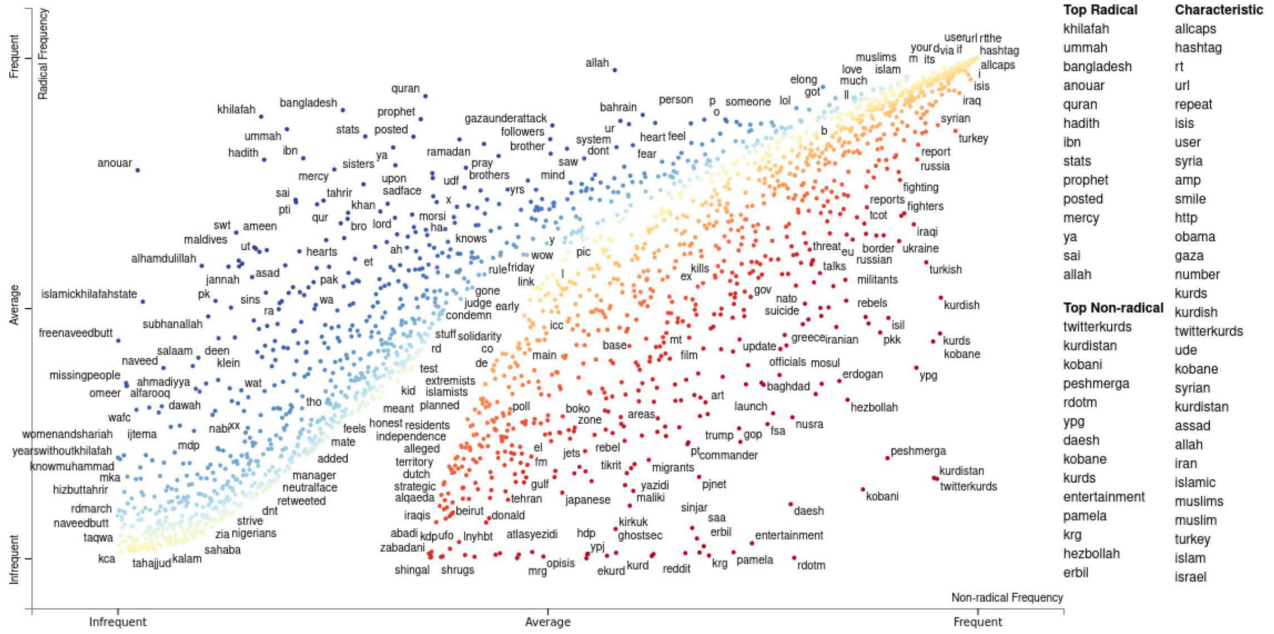


**FIGURE 3.** Word frequency for both radical and non-radical categories regarding half of the *Pro-anti* dataset. On the right, the most frequent words for radical (Top Radical), non-radical (Top Non-radical) and both (Characteristic).

***Magazines.*** This English-written dataset is presented in [7] and two parts can be differentiated. For the first one, the data came from two online magazines shared by the Islamic State of Iraq and the Levant radical organization [13]: Dabiq [33] and Rumiyah [51] magazines. 166 articles from 13 editions have been extracted from Dabiq (released in July 2014 and lasted two years) and 155 articles corresponding to 15 editions from Rumiyah (released in September 2016 and lasted one year). All content has been originally extracted from `jihadology.net`, a digital platform that addresses terrorism.[1] As a balance of Dabiq and Rumiyah's texts, additional data is considered. Concretely, texts from two digital newspapers that deal with matters related to ISIS from a non-radical perspective: Cable News Network (CNN)[2] and The New York Times,[3] which are both sources that make their content available through their APIs. To gather the data, a keyword-based filtration (e.g., *ISIS, Daesh*) was done for a 10-month period. For an increase in the value of the categorization, texts that did not address the ISIS issue were removed, as well as links, images and other media. As a result, 129 instances were added to the *Magazines* dataset from CNN and 23 from The New York Times.

***Jacobs.*** This resorce was created by Scanlon and Gerber [73]. It is composed of a group of jihadist posts from private forums with origin in the dark web, which were already assembled in the Dark Web Portal Project [20]. These forums belong to the Ansar AlJihad Network, an extremist organization with ties to Al-Qaeda and well-liked among Western jihadists [73]. The classification carried out in this dataset is binary: propaganda and non-propaganda.

Following on the rest of the resources used for the evaluation, we have described how word embeddings have been used in the context of the SIMON [9], [11] method. Following previous research [7], we use the FastText embedding model [17]. Said model contains 300-dimension vectors with a vocabulary size of 1,999,995 words in which the training domain is Wikipedia.

In addition, as described above, an objective of this work is to explore the impact of contextualization through domain-relevant lexicons. We study the effect of the following lexicons.

The ***MoralStrength*** lexicon is a resource containing 2,845 annotated words and their association strengths with each moral trait (care, fairness, loyalty, authority, purity). For each word, the lexicon provides a numerical evaluation of moral valence. This lexicon is an expansion of the lemmas from the original Moral Foundations Dictionary (MFD), as presented in [6].

[1] http://www.jihadology.net/
[2] https://cnn.com/
[3] https://nytimes.com/

***NRC Hashtag Emotion Lexicon*** [63] is an inventory of 16,862 English words and their relationship with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive), selected concretely for the Twitter platform. As observed in previous works [7], emotion can have an impact on radicalization detection.

***eMFD Lexicon*** [43] is made up of 3,270 English words. Every word is allocated 5 probability scores that indicate their association with each moral foundation, as well as 5 sentiment ratings that encode the average sentiment of the foundation context in which each word appears.

Finally, it is worth mentioning that a specific lexicon has been generated and utilized for each dataset to extract dataset-oriented features using SIMON feature extractor. We use the *FreqSelect* method, presented in [7].

### B. METHODOLOGY

We design a thorough evaluation to assess the different proposed models and their effectiveness in the task of radicalization detection. In every experiment, we use the macro-averaged F1 score for performance assessment. For each dataset, a k-fold cross-validation is performed, being $k = 10$. To study the effect of contextualization through diverse representations, we use different feature extractors, concatenating their resulting vectors. The ones used as the baseline are unigrams, bigrams, the TF-IDF method and SIMON. Unigrams provide speed, transparency, flexibility, and accuracy as a radical propaganda detection technique. Analyzing bigrams offered a more nuanced comprehension of propaganda techniques, a higher precision, and aid in the identification of complex propaganda methods that might evade detection by solely analyzing unigrams. The TF-IDF method is also used as it can provide a more targeted analysis and the detection of important themes.

At first, the only results taken into account were the ones obtained from extracting features with unigrams, bigrams and the TF-IDF method. These outputs were compared with those generated using the SIMON method. Following, we merged these two approaches to see if an improvement had taken place compared to the baseline. The process that followed this stage of the evaluation was the combination of the feature extractors with the used lexicons. Additionally, we include an evaluation of the effectiveness of a unified representation that combines SIMON, emotion and morals.

Regarding the classifier algorithms, two have been used: logistic regression and linear Support Vector Machine.

### C. RESULTS

Tables 2 and 3 present the results of the evaluation considering the described models and their combinations with the contextual lexicons.

It is noticeable that the F1 scores do not defer much from one classifier to the other. As seen in tables 2 and 3, it can be seen that the best scores are consistently obtained by combinations of representations.

**TABLE 2.** Results with the logistic regression classifier.

| Method | Pro-anti | Magazines | Jacobs |
|---|---|---|---|
| Unigrams | 88.78 | 95.38 | 91.00 |
| Unigrams + MoralStrength | 88.60 | 94.64 | 90.84 |
| Unigrams + NRC | 88.60 | 95.11 | 91.04 |
| Unigrams + SIMON | 88.87 | 95.43 | 91.02 |
| Unigrams + eMFD | 88.78 | 94.66 | 90.94 |
| Bigrams | 84.97 | 96.86 | 86.93 |
| Bigrams + MoralStrength | 84.88 | 97.10 | 87.42 |
| Bigrams + NRC | 86.13 | 95.16 | 88.07 |
| Bigrams + SIMON | 86.13 | **97.58** | 90.50 |
| Bigrams + eMFD | 85.06 | 96.13 | 86.99 |
| TF-IDF | 85.76 | 86.93 | 91.71 |
| TF-IDF + MoralStrength | 84.61 | 85.67 | 91.32 |
| TF-IDF + NRC | 87.24 | 80.80 | 91.58 |
| TF-IDF + SIMON | **89.09** | 95.64 | 91.48 |
| TF-IDF + eMFD | 85.48 | 86.81 | **91.73** |
| SIMON | 88.03 | 94.35 | 90.04 |
| SIMON + MoralStrength | 88.48 | 93.11 | 90.14 |
| SIMON + NRC | 88.85 | 94.93 | 90.32 |
| SIMON + eMFD | 88.39 | 95.43 | 90.26 |
| SIMON + MoralStrength + NRC | 88.93 | 93.86 | 90.26 |

**TABLE 3.** Results with the linear SVM classifier.

| Method | Pro-anti | Magazines | Jacobs |
|---|---|---|---|
| Unigrams | 84.16 | 92.99 | 90.88 |
| Unigrams + MoralStrength | 83.85 | 93.94 | 90.82 |
| Unigrams + NRC | 85.05 | 93.88 | 90.90 |
| Unigrams + SIMON | 87.42 | 94.86 | 90.26 |
| Unigrams + eMFD | 84.23 | 93.69 | 90.88 |
| Bigrams | 87.99 | 95.88 | 86.53 |
| Bigrams + MoralStrength | 85.32 | 93.20 | 86.87 |
| Bigrams + NRC | 85.77 | 94.37 | 87.27 |
| Bigrams + SIMON | 88.15 | **97.33** | 90.06 |
| Bigrams + eMFD | 85.24 | 94.43 | 86.43 |
| TF-IDF | 88.32 | 95.19 | 93.20 |
| TF-IDF + MoralStrength | 87.70 | 94.96 | 93.30 |
| TF-IDF + NRC | **91.42** | 91.29 | **93.52** |
| TF-IDF + SIMON | 89.20 | 93.39 | 92.82 |
| TF-IDF + eMFD | 88.77 | 93.02 | 93.24 |
| SIMON | 86.43 | 93.39 | 89.86 |
| SIMON + MoralStrength | 86.62 | 91.69 | 89.96 |
| SIMON + NRC | 85.92 | 94.39 | 90.28 |
| SIMON + eMFD | 86.98 | 93.18 | 90.20 |
| SIMON + MoralStrength + NRC | 85.74 | 93.18 | 90.14 |

In *Pro-anti*, Table 2 shows that the strongest result is obtained by joining SIMON and features obtained through the TF-IDF method. There is a very noticeable increase with respect to just TF-IDF, while a less noticeable increase with respect to just SIMON. Such a result indicates that the sparsity of the TF-IDF representations complements the SIMON model. Although the combinations with *MoralStrength* and *eMFD* decrease this baseline, this does not happen when using SIMON as the baseline, where both moral foundations and emotions approaches help to get better scores even when using them at the same time (SIMON +*MoralStrength*+ NRC). In the case of unigrams, only SIMON has an added value and in the case of bigrams only *MoralStrength* decreases the baseline. On the other hand, Table 3 has the best score when combining the TF-IDF method and emotions. When using SIMON as the baseline, only moral foundation-based lexicons help to increase the performance. This does not happen when the baselines are unigrams or bigrams: in this cases, only SIMON increases the overall performance in both cases.

In *Magazines*, Tables 2 and 3 show that the most outstanding result is the one resulting from the combination of bigrams and the SIMON method. Nevertheless, *MoralStrength* contributes to get a nearly as high mark when using the logistic regression classifier. Continuing with this classifier, an interesting observation is that SIMON is the only one that, in combination with the baseline, increases the score when unigrams or TF-IDF approaches are used. With this method as the baseline, moral features seem to elevate the result obtained when *eMFD* lexicon is utilized. Emotions have an important role too in these results as they also enhance the score. Moving on to the SVM classifier, it is worth mentioning that emotions and moral foundations improve the baseline results of unigrams. Although the combination of semantic similarity with *MoralStrength* and emotions does not provide the highest result, it is still a competitive score.

In the *Jacobs* dataset, the best approach that can be used in combination with others is the TF-IDF method. This can be seen in tables 2 and 3, where its combination with *eMFD* and NRC gives the best scores in each of them, respectively. When using unigrams as the baseline, only emotions can improve it when using the SVM classifier, but if bigrams or SIMON are utilized, morality (specially *MoralStrength*) and the affect lexicon have an additional positive impact for both classifiers.

For a more visual understanding of the impact each combination of methods has on text classification, SHapley Additive exPlanations (SHAP) has been applied to the models with the best scores [50]. The SHAP method offers a thorough framework for interpreting the predictions generated by any machine learning algorithm. It entails explaining the output of machine learning models by assigning an importance score to each input feature that indicates how much each feature contributes to the prediction. SHAP is based on the Shapley value concept from cooperative game theory, where a fair assignation of the contribution of each player takes place in a cooperative game. In this context, the ''players'' are the input features, and the ''game'' is the prediction task.

In this work, we exploit the information contained in 'beeswarm plots', which are a SHAP visualization tool for explaining the results obtained. Each point represents an instance of the dataset and its position on the x-axis indicates its SHAP value for a specific feature, which is an indicator of the contribution of that feature to the final model prediction for that instance. The density of the dots is shown through the y-axis, which gives information about the distribution of the feature values in the dataset. Also, the colour of the dot indicates the magnitude of the feature for each instance. For a better understanding, different colours have been applied to different categories of extracted features: yellow for moral foundations, green for emotions, blue for SIMON features and red for bigrams.

Figures 4 and 5 show that moral values, emotions and features obtained through the SIMON model have an important role in the classification process (the words that appear are the most important ones).

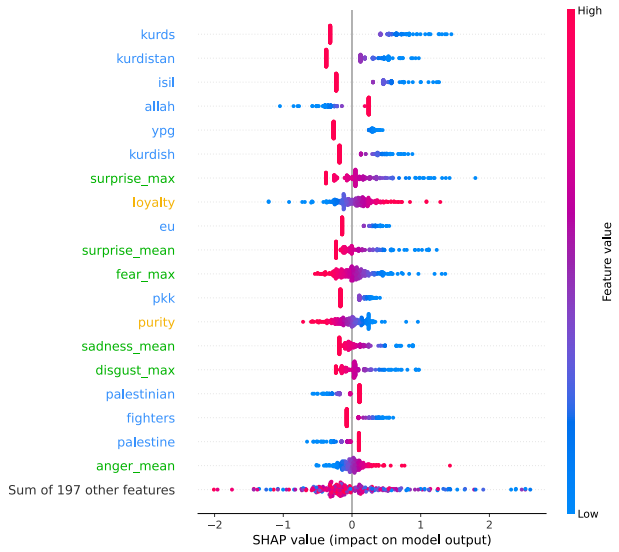


**FIGURE 4.** SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Pro-anti* dataset.

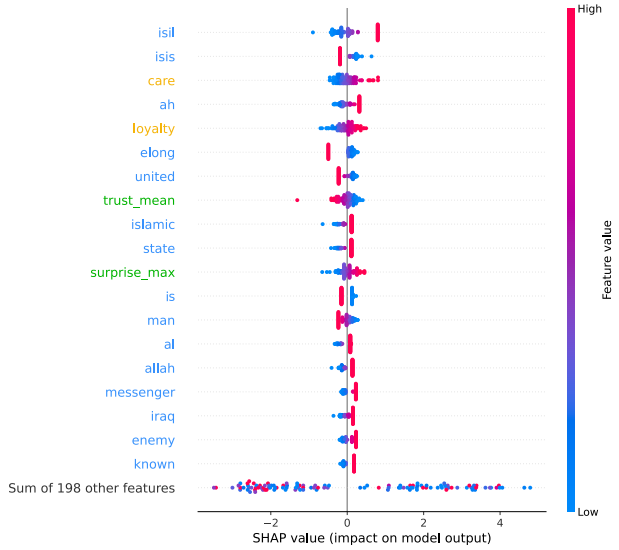


**FIGURE 5.** SHAP beeswarm plot for SIMON + MoralStrength + NRC with *Magazines* dataset.

Additionally, figures 4 and 5 show that the moral value 'loyalty' is a feature with red positive SHAP values, which means that it is a feature that contributes to the final prediction of the model with an inclination towards extremism. We understand this foundation in radical speeches as a virtue that requires a continuous commitment to a particular ideology, cause or leader. The final aim of these discourses is to enforce conformity and suppress dissent within the group or movement. It is worth mentioning the appearance of the word 'purity' in Figure 4 as a feature with positive blue SHAP values. A possible explanation for this would be that purity-oriented individuals may associate extremist ideologies with immoral or impure actions that conflict with their own moral values.

Regarding the emotive effect on the classification, Figure 4 shows that 'anger' is an emotion that helps to classify

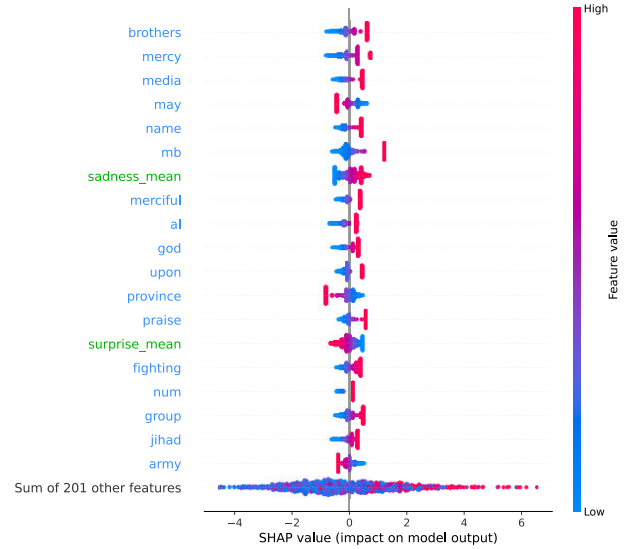


**FIGURE 6.** SHAP beeswarm plot for SIMON + MoralSrength + NRC with *Jacobs* dataset.

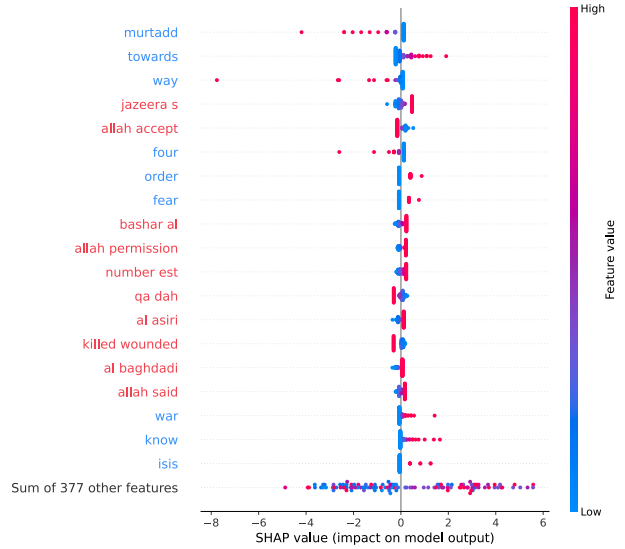


**FIGURE 7.** SHAP beeswarm plot for bigrams + SIMON with *Magazines* dataset.

instances with a radical perspective as the extremist speech resides in hatred and fury. At the same time 'sadness', 'fear' and 'disgust' contribute to the classification of texts with a non-radical discourse, which is a predictable result. In Figure 6, the most outstanding feature is the 'anticipation' emotion. This can be due to the fact that radical discourses sometimes can show a promise of a future outcome or transformation. They may stress the need for immediate action for apparent injustices that require extremist solutions.

Features obtained through SIMON method, such as 'ypg' in Figure 4, have interesting explanations with contributions to non-extremist discourses. The 'ypg' feature refers to ''People's Protection Units'', a militia group from Kurdistan active since 2011 that has been the partner of the United States coalition in Syria against the ISIS [72]. This also explains that 'kurds', 'kurdistan' and 'kurdish' have high SHAP values regarding non-radicality. In contrast, 'allah', a feature shown in Figure 5, has not a high positive SHAP value concerning radical texts. It is an Arabic word that means 'God' and holds important religious and cultural relevance for Muslims; in some cases, individuals or groups with radical or extremist ideologies may use religious language, including references to Allah, to legitimize their actions. 'Praise' and 'mercy' are similar examples shown in Figure 6.

Figure 7 shows the beeswarm plot for the model in which bigrams and features from the SIMilarity-based sentiment projectiON (SIMON) model are combined, as it is a fusion that has provided a very positive result in the experiments. For example, bigrams like 'al baghdadi' (the former leader of the Islamic State of Iraq and Syria) and 'bashar al' (the president of Syria) seem to have helped (with a low SHAP value) to the categorization of radical text. SIMON features like 'war', 'isis' or 'fear' have also carried out this function. The feature 'murtadd' (an Arabic word that refers to an individual who rejects or abandons their previously embraced religious beliefs or faith) would be anticipated to operate as the previous ones described, but surprisingly, it has very low SHAP values for text classification towards the radical side.

## V. DISCUSSION

This work follows the research done in [7], where the detection of radicalization is carried out through the exploratory combination of emotions and similarity-based features with positive results. Here, we add moral values based on the Moral Foundations Theory (MFT) y and see how they affect the results as they have an important role in the task of identifying political inclinations [34]. It is worth mentioning that the scores obtained in this paper are macro-averaged, which involves taking the arithmetic mean (unweighted mean) of the F1 scores for each class and treating all classes equally regardless of their prevalence in the dataset.

Regarding the method used in this investigation, other studies do not use machine learning techniques to evaluate the radicalization present in texts [18], [36], [45], [49], [68], [78]. Our approach employs two commonly used machine learning algorithms: logistic regression and linear SVM. They were selected for this work since the objective is the classification of text according to their level of extremism, employing a richer variety of features but avoiding the complexity of more advanced techniques. These or comparable learning algorithms are also utilized in previous studies [2], [3], [12], [15], [19], [29], [71]. Our approach differs mainly in the feature extraction stage. The proposed method utilizes moral values, emotions and similarity-based features that leverage the extensive lexicon contained in word embeddings (apart from unigrams, bigrams and features generated by the TF-IDF method, which also have important results in tables 2 and 3). As far as we are aware, this type of approach for evaluating radicalization has not been previously studied. What previous works have done is utilize lexicons by directly comparing words in the analyzed text with those included in

the lexicon. Thus, such approaches fail to adequately model out-of-vocabulary terms.

Regarding the technology that has been used in general terms and how it can be further improved in future work, this research has selected the automated detection path. This path has two ways: graph-based and machine learning approaches. According to [32], the graph-baed approach is not as popular as the ML approach among the studies that have been carried out as it is employed to discover followers of known radicals and to identify the extent of extremist influence, but fails when classifying news services and new radical users. On the contrary, the most popular machine learning techniques use algorithms that efficiently discover newly radicalized users and continuously learn from the transforming extremist content.

The limitations that have been perceived in this work are the lack of resources available in the radicalization domain. On the one hand, there are not many datasets that model radicalization available, as some texts come from magazines intended for a very specific public. Others come from Twitter accounts, being a relevant amount banned or deleted. On the other hand, it is not possible to work with a corpus of significant size containing explicit annotations for detecting radicalization, which greatly adds difficulty to the task of training domain-oriented word embedding models [30].

## VI. CONCLUSION

This paper describes a machine learning system that utilizes various categories of characteristics to accomplish the objective of identifying radical propaganda in texts that come from both social media and magazines: moral features, emotions, similarity-based features obtained through word embeddings, unigrams, bigrams and features acquired from the TF-IDF method. Besides, these approaches are combined with the aim of improving the overall task performance. Regarding the moral and emotion representations, they provide supplementary information that gathers moral and emotional knowledge and adds context to the analyzed text creating features that can be very useful in the classification process. On the other hand, the SIMON approach is used with a domain lexicon, obtained by selecting words based on their frequency in the training data. Such a method is denoted *FreqSelect* [7]. Finally, unigrams, bigrams, and features from the TF-IDF method also have an added value as they help to identify noteworthy word associations.

We evaluate all presented techniques using three datasets that contain radical language. To do so, we design a thorough evaluation that studies the performance of the system by means of an ablation test, as well as the study of the interpretability of the obtained solutions using SHAP values.

Section I raises two essential research questions. The first one (**RQ1**) asks **whether moral values are useful for identifying radical propaganda**. To answer RQ1, in general terms, Figure 2 shows that when using a logistic regression classifier, the F1 score improves with respect to the baseline when using bigrams or the SIMON model. Table 3 shows a noteworthy enhancement of the F1 score in the *Pro-anti* dataset when using unigrams (5,49% more with *eMFDscore*) and also the SIMON model (7,91% more with *MoralStrength*). Besides, when utilizing bigrams, *MoralStrength* increases the baseline result with every dataset and *eMFDscore* with *Pro-neu* dataset with an improvement of 11,55%. All these experimental results indicate that the assessment of moral values in text can enhance the performance of an automatic system for radicalization detection. These positive outcomes open a promising avenue for the field, suggesting that comprehensive text representations, including morality cues, can be incorporated into system that analyze radical language.

Following, **RQ2** addresses the **effectiveness of combining moral values and emotion information**. In this sense, as seen in Table 2, using a logistic regression classifier the F1 score improves in *Pro-anti* and *Magazines* datasets with respect to the baseline. If the classifier used is a linear SVM, an enhancement can be seen in *Pro-anti* dataset. We have seen that the values obtained through SIMON model are already very high, so it is a difficult task to elevate them. While the experiments indicate that the combined use of moral values and emotions can offer more informed representations, it is necessary to further study their effect on different learners.

Despite the fact that the results achieved are quite promising, we consider that it is important to keep a thorough study on the use of language as it progressively varies with the passing of time. An essential task that should be addressed in future work is the gathering of new data that incorporates the time component in the radicalization process. In addition, how this data is classified can also be enhanced: the binary categorization of texts (radical and non-radical) can be substituted by a classification based on a wider spectrum. This would provide more detailed distinction and a greater understanding of the nature and impact of radical content.

Technologies that the research community should keep investigating are advanced deep learning architectures such as neural networks and transformer models in order to effectively capture the connections and meaningful associations within propaganda texts. Also, the implementation of active learning techniques would ease the path of researchers as the models would be updated with new information. Consequently, this would decrease the dependence on manually annotated data.

## ACKNOWLEDGMENT

The author Oscar Araque would like to thank the ''ETSI Telecomunicación'' of ''Universidad Politécnica de Madrid'' through the initiative ''Primeros Proyectos'' under ''AFRICA—Detecting and Analyzing Affective and Moral Factors in Radicalization and ExtremIsm: a MaChine learning Approach.''

## REFERENCES

[1] J. Garten, J. Hoover, K. M. Johnson, R. Boghrati, C. Iskiwitch, and M. Dehghani, ''Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis,'' *Behav. Res. Methods, Instrum., Comput.*, vol. 50, no. 1, pp. 344–361, Feb. 2018.

[2] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *Proc. IEEE Intell. Secur. Informat.*, May 2007, pp. 282–288.
[3] S. Agarwal and A. Sureka, "Using knn and SVM based one-class classifier for detecting online radicalization on Twitter," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.*, 2015, pp. 431–442.
[4] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Online extremism detection in textual content: A systematic literature review," *IEEE Access*, vol. 9, pp. 42384–42396, 2021.
[5] A. Anil, D. Kumar, S. Sharma, R. Singha, R. Sarmah, N. Bhattacharya, and S. R. Singh, "Link prediction using social network analysis over heterogeneous terrorist network," in *Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 267–272.
[6] O. Araque, L. Gatti, and K. Kalimeri, "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105184.
[7] O. Araque and C. A. Iglesias, "An approach for radicalization detection based on emotion signals and semantic similarity," *IEEE Access*, vol. 8, pp. 17877–17891, 2020.
[8] O. Araque and C. A. Iglesias, "An ensemble method for radicalization and hate speech detection online empowered by sentic computing," *Cognit. Comput.*, vol. 14, no. 1, pp. 48–61, Jan. 2022.
[9] O. Araque, J. F. Sánchez-Rada, and C. A. Iglesias, "GSITK: A sentiment analysis framework for agile replication and development," *SoftwareX*, vol. 17, Jan. 2022, Art. no. 100921.
[10] 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 Int. Conf. Data Mining Workshops (ICDM Workshops)*, Barcelona, Spain, C. Domeniconi, F. Gullo, F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou, and X. Wu, Eds., Dec. 2016, pp. 900–907.
[11] 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.
[12] 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.
[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.
[14] D. Benito, O. Araque, and C. A. Iglesias, "GSI-UPM at SemEval-2019 task 5: Semantic similarity and word embeddings for multilingual detection of hate speech against immigrants and women on Twitter," in *Proc. 13th Int. Workshop Semantic Eval.* Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 396–403.
[15] 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. ASONAM*, Jul. 2009, pp. 231–236.
[16] E. Bodine-Baron, T. C. Helmus, M. Magnuson, and Z. Winkelman, *Examining ISIS Support and Opposition Networks on Twitter*. Santa Monica, CA, USA: RAND Corporation, 2016.
[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2016, *arXiv:1607.04606*.
[18] T. Chalothorn and J. Ellman, "Using sentiwordnet and sentiment analysis for detecting radical content on web forums," in *Proc. 6th Conf. Softw., Knowl., Inf. Manag. Appl. (SKIMA)*, Chengdu Univ., Sep. 2012.
[19] 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.
[20] H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor, *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*, vol. 18. New York, NY, USA: Springer, 2008.
[21] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression Violent Behav.*, vol. 40, pp. 108–118, May 2018.
[22] S. Clifford and J. Jerit, "How words do the work of politics: Moral foundations theory and the debate over stem cell research," *J. Politics*, vol. 75, no. 3, pp. 659–671, Jul. 2013.
[23] D. Correa and A. Sureka, "Solutions to detect and analyze online radicalization: A survey," 2013, *arXiv:1301.4916* .
[24] G. D. S. Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, and P. Nakov, "A survey on computational propaganda detection," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2021, pp. 4826–4832.
[25] C. E. Dauber and C. K. Winkler, "Radical visual propaganda in the online environment: An introduction," in *Visual Propaganda and Extremism in the Online Environment*, 2014, pp. 1–30.
[26] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," in *Proc. 2nd Workshop Abusive Lang. Online (ALW)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20.
[27] J. Decety, R. Pape, and C. I. Workman, "A multilevel social neuroscience perspective on radicalization and terrorism," *Social Neurosci.*, vol. 13, no. 5, pp. 511–529, Sep. 2018.
[28] M. Dehghani, K. Sagae, S. Sachdeva, and J. Gratch, "Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the ground zero Mosqu," *J. Inf. Technol. Politics*, vol. 11, no. 1, pp. 1–14, Jan. 2014.
[29] 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: Association for Computing Machinery, 2017, pp. 439–446.
[30] M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on Twitter," in *Proc. Semantic Web Social Good Workshop, Int. Semantic Web Conf. (SW4SG)*. CEUR, Oct. 2018. [Online]. Available: https://sw4sg.github.io/ISWC2018/
[31] 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.* New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 679–688.
[32] M. Gaikwad, S. Ahirrao, S. Phansalkar, and K. Kotecha, "Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools," *IEEE Access*, vol. 9, pp. 48364–48404, 2021.
[33] H. K. Gambhir, "Dabiq: The strategic messaging of the Islamic state," Inst. Study War, Washington, DC, USA, Tech. Rep., Apr. 2014.
[34] J. Garten, R. Boghrati, J. Hoover, K. M. Johnson, and M. Dehghani, "Morality between the lines: Detecting moral sentiment in text," Tech. Rep.
[35] J. Garten, B. Kennedy, J. Hoover, K. Sagae, and M. Dehghani, "Incorporating demographic embeddings into language understanding," *Cognit. Sci.*, vol. 43, no. 1, pp. 1–18, Jan. 2019.
[36] 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.
[37] C. González-Santos, M. A. Vega-Rodríguez, C. J. Pérez, J. M. López-Muñoz, and I. Martínez-Sarriegui, "Automatic assignment of moral foundations to movies by word embedding," *Knowl.-Based Syst.*, vol. 270, Jun. 2023, Art. no. 110539.
[38] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Chapter two—Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in Experimental Social Psychology*, vol. 47. New York, NY, USA: Academic, 2013, pp. 55–130.
[39] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations," *J. Personality Social Psychol.*, vol. 96, no. 5, pp. 1029–1046, May 2009.
[40] 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.
[41] G. Hofstede, *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Newbury Park, CA, USA: Sage, 2001.
[42] J. Hoover, K. Johnson, R. Boghrati, J. Graham, and M. Dehghani, "Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation," *Collabra, Psychol.*, vol. 4, no. 1, pp. 1–18, 2018.
[43] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (EMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text—Behavior research methods," *Behav. Res. Methods*, vol. 53, pp. 232–246, Jul. 2020.
[44] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*.
[45] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur. Informat.*, vol. 4, no. 1, pp. 1–13, Dec. 2015.

[46] R. Kaur and K. Sasahara, "Quantifying moral foundations from various topics on Twitter conversations," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 2505–2512.
[47] B. Kennedy, M. Atari, A. M. Davani, J. Hoover, A. Omrani, J. Graham, and M. Dehghani, "Moral concerns are differentially observable in language," *Cognition*, vol. 212, Jul. 2021, Art. no. 104696.
[48] Y. Liu and M. Zhang, *Neural Network Methods for Natural Language Processing*. Yoav Goldberg: Springer, Jun. 2022.
[49] 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*, L. Uden, B. Hadzima, and I.-H. Ting, Eds. Cham, Switzerland: Springer, 2018, pp. 589–599.
[50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 4765–4774.
[51] R. Mahzam, "Rumiyah—Jihadist propaganda & information warfare in cyberspace," *Counter Terrorist Trends Analyses*, vol. 9, no. 3, pp. 8–14, 2017.
[52] G. D. S. Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. D. Pietro, and P. Nakov, "A survey on computational propaganda detection," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, C. Bessiere, Ed., Jul. 2020, pp. 4826–4832.
[53] M. Bérubé, L.-A. Beaulieu, P. Mongeau, and J. Saint-Charles, "Identifying key players in violent extremist networks: Using socio-semantic network analysis as part of a program of content moderation," *Stud. Conflict Terrorism*, vol. 53, pp. 1–19, May 2021.
[54] C. McCauley and S. Moskalenko, "Mechanisms of political radicalization: Pathways toward terrorism," *Terrorism Political Violence*, vol. 20, no. 3, pp. 415–433, Jul. 2008.
[55] C. McCauley and S. Moskalenko, "Mechanisms of political radicalization: Pathways toward terrorism," *Terrorism Political Violence*, vol. 20, no. 3, pp. 415–433, Jul. 2008.
[56] R. M. Medina, "Social network analysis: A case study of the Islamist terrorist network," *Secur. J.*, vol. 27, no. 1, pp. 97–121, Feb. 2014.
[57] Y. Mejova, K. Kalimeri, and G. D. F. Morales, "Authority without care: Moral values behind the mask mandate response," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 17, no. 1, 2023, pp. 614–625, doi: 10.1609/icwsm.v17i1.22173.
[58] J. Mellon, J. Yoder, and D. Evans, "Undermining and strengthening social networks through network modification," *Sci. Rep.*, vol. 6, no. 1, p. 34613, Oct. 2016.
[59] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[60] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2013, pp. 746–751.
[61] G. A. Miller, "Wordnet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[62] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–17.
[63] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, May 2015, doi: 10.1111/coin.12024.
[64] M. Nouh, J. R. C. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on Twitter," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 98–103.
[65] J. W. Pennebaker, M. E. Francis, R. J. Booth, *Linguistic Inquiry and Word Count*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2001.
[66] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.
[67] 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.
[68] M. Rowe and H. Saif, "Mining pro-ISIS radicalisation signals from social media users," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 1, Aug. 2021, pp. 329–338.
[69] E. Sagi and M. Dehghani, "Measuring moral rhetoric in text," *Social Sci. Comput. Rev.*, vol. 32, no. 2, pp. 132–144, Apr. 2014.
[70] E. Sagi and M. Dehghani, "Moral rhetoric in Twitter: A case study of the U.S. Federal Shutdown of 2013," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, vol. 36, 2014. [Online]. Available: https://escholarship.org/uc/item/9sw937kk
[71] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, "A semantic graph-based approach for radicalisation detection on social media," in *Proc. Extended Semantic Web Conf.*, 2017, pp. 571–587.
[72] E. Savelsberg, "The Kurdish PYD and the Syrian civil war," in *Routledge Handbook on the Kurds*. Evanston, IL, USA: Routledge, 2018, pp. 357–365.
[73] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," *Secur. Informat.*, vol. 3, no. 1, pp. 1–10, Dec. 2014.
[74] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 298–307.
[75] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.
[76] L. Teernstra, P. van der Putten, L. Noordegraaf-Eelens, and F. Verbeek, "The morality machine: Tracking moral values in tweets," in *Advances in Intelligent Data Analysis XV*, H. Boström, A. Knobbe, C. Soares, and P. Papapetrou, Eds. Cham, Switzerland: Springer International Publishing, 2016, pp. 26–37.
[77] J. van Stekelenburg, "Radicalization and violent emotions," *PS, Political Sci. Politics*, vol. 50, no. 04, pp. 936–939, Oct. 2017.
[78] 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*, vol. 2, pp. 7–20, 2015.
[79] L. Vidino, "Countering radicalization in America," JSTOR, United States Inst. Peace, Washington, DC, USA, Tech. Rep., 2010.
[80] M. Waniek, T. P. Michalak, M. J. Wooldridge, and T. Rahwan, "Hiding individuals and communities in a social network," *Nature Hum. Behav.*, vol. 2, pp. 139–147, Jan. 2018.
[81] M. Xu, L. Hu, and G. T. Cameron, "Tracking moral divergence with DDR in presidential debates over 60 years," *J. Comput. Social Sci.*, vol. 6, no. 1, pp. 339–357, Apr. 2023.
[82] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[83] 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.
[84] M. Stephens, "Facing ISIS: The Kurds of Syria and Iraq," in *IeMed Mediterranean Yearbook. European Institute of the Mediterranean*, Barcelona, Spain, 2015.

**PATRICIA ALONSO DEL REAL** (Member, IEEE) is currently pursuing the bachelor's degree in telecommunication technologies and services engineering with Universidad Politécnica de Madrid (UPM). Her research interest includes machine learning, especially techniques used in natural language processing and cybersecurity. She is a member of the IEEE Women in Engineering.

**OSCAR ARAQUE** is currently an Assistant Professor with Universidad Politécnica de Madrid (UPM). His research interest includes the application of machine learning techniques for natural language processing. The main topic of his thesis is introducing specific domain knowledge into machine learning systems to enhance sentiment and emotion analysis techniques and their applications to new domains, such as radicalization narratives. His work has received four distinguished prizes, such as the Most Cited Scientific Paper Award 2020 by Universidad Politécnica de Madrid, the Prize for the Most Cited Scientific Article Originating from a UPM Doctoral Thesis, in 2021, the ISDEFE Award for the Best Doctoral Thesis in Security and Defense, in 2020, and the Extraordinary Doctoral Thesis Award-ETSIT UPM 2022.