Cognitive Computation (2022) 14:48–61
https://doi.org/10.1007/s12559-021-09845-6

An Ensemble Method for Radicalization and Hate Speech Detection Online Empowered by Sentic Computing

Oscar Araque[1] · **Carlos A. Iglesias**[1]

Received: 16 June 2020 / Accepted: 3 February 2021 / Published online: 16 February 2021

Abstract
The dramatic growth of the Web has motivated researchers to extract knowledge from enormous repositories and to exploit the knowledge in myriad applications. In this study, we focus on natural language processing (NLP) and, more concretely, the emerging field of affective computing to explore the automation of understanding human emotions from texts. This paper continues previous efforts to utilize and adapt affective techniques into different areas to gain new insights. This paper proposes two novel feature extraction methods that use the previous sentic computing resources AffectiveSpace and SenticNet. These methods are efficient approaches for extracting affect-aware representations from text. In addition, this paper presents a machine learning framework using an ensemble of different features to improve the overall classification performance. Following the description of this approach, we also study the effects of known feature extraction methods such as TF-IDF and SIMilarity-based sentiment projectiON (SIMON). We perform a thorough evaluation of the proposed features across five different datasets that cover radicalization and hate speech detection tasks. To compare the different approaches fairly, we conducted a statistical test that ranks the studied methods. The obtained results indicate that combining affect-aware features with the studied textual representations effectively improves performance. We also propose a criterion considering both classification performance and computational complexity to select among the different methods.

Keywords Sentic computing · Affective computing · Radicalization detection · Hate speech detection · Machine learning · Natural language processing

Introduction

The rapid growth of users and user-generated content has dramatically increased the quantity of information available on the Web. This content is published in myriad sites, such as websites, social networks, online consumer platforms, online communities, and other collaborative media. Information in this form is spread across all these places and spans a large number of topics, which turns the attention of many actors that aim to distill knowledge from such content [1].

However, extracting knowledge from such a distributed, unstructured, and significant source is an arduous task. Automatically extracting, processing and understanding user-generated language have been sources of tremendous interest for researchers since the applications can widely advance existing artificial intelligence technology [2, 3].

Natural language processing (NLP) has significant importance as a research area that generally addresses these challenges. In particular, we stress the importance of affective computing whose primary aim is to understand human emotion computationally [4]. A relevant challenge lies in exploiting affective computing techniques, models, and insights in different areas that do not initially benefit from these approaches. These areas include business, commerce, health, psychology, learning, and mobility [5].

This paper focuses on the application of affective computing to two distinct areas that can largely benefit from it: radicalization analysis and hate speech detection. While these two areas have been previously enhanced by the addition of affective computing approaches, previous works do not thoroughly study the role of emotions in the detection process. More concretely, this paper addresses the effect of

✉ Oscar Araque
o.araque@upm.es

Carlos A. Iglesias
carlosangel.iglesias@upm.es

[1] Intelligent Systems Group, Universidad Politécnica de Madrid, 28040 Madrid, Spain

incorporating sentic methods and resources, as presented by [6].

Therefore, we focus on the following research questions (RQs):

1. RQ1. Can affect-aware systems (sentic computing approaches) improve performance in NLP application domains?
2. RQ2. Considering the model's performance and computational complexity, can we obtain a criterion for selecting among the proposed methods?

Motivated by these RQs, this paper proposes a machine learning approach that combines known text representations with sentic computing methods to obtain a more robust feature extraction framework. Regarding the text representations, we explore the effect of TF-IDF and SIMON [7], a model that exploits a word embedding model to project input text to a domain lexicon by computing wordwise similarities. Next, we propose two novel methods that extract knowledge from the AffectiveSpace [8] and SenticNet [9] resources. These novel methods aim to maintain low computational complexity while extracting useful affect representations.

The remainder of this paper is organized as follows. Section 2 presents the related work on the application of affective computing to both radical and hate speech detection. Section 3 describes the proposed methods of this paper. Next, in Section 4, we depict the evaluation tasks, including the materials and method used and a thorough analysis of the obtained results. Finally, the paper concludes with Section 5, which provides the insights of this paper and outlines possible future work.

Related Work

Affective Computing and Sentiment Analysis

Affective computing addresses a challenge that entails a broad set of NLP problems that must be addressed to achieve a general understanding of human emotion. In this way, problems related to affective computing and sentiment analysis can be organized in a layered fashion, with three distinct layers that increase in abstraction: syntactics, semantics, and pragmatics [10].

Deep learning approaches have marked a path for improvement by increasing performance in all related problems, and affective computing applications. As a relevant advantage, these systems have the ability to include different sources of information and generate useful representations in a semisupervised or even unsupervised manner. These trends take the field from a bag-of-words to a bag-of-concepts perspective, as described by [11].

One of the mentioned information sources is semantic resources, which contain hierarchical and organized knowledge that can be used by other systems. A relevant example is SenticNet [12], which was constructed using a combination of logical reasoning with deep learning architectures. Similarly, an ontology that allows for common sense reasoning is OntoSenticNet [13], which was built on top of SenticNet.

While semantic models and resources can be effectively leveraged [14], deep learning has made advancements in many other areas. In multimodal sentiment analysis, where text is analyzed in conjunction with audio and video, neural models can be effectively used [15, 16].

Additionally, advances have been made in the relevant challenge of cross-lingual sentiment analysis, where systems are trained in a language using abundant available data and are later adapted to make predictions in a target language. Chen et al [17] presented an adversarial neural network to transfer the knowledge to target languages with no labels. In a novel work, Esuli et al [18] addressed cross-lingual sentiment quantification using a neural architecture. Transfer learning is an active area of research in affective computing and sentiment analysis [19].

Another relevant trend is semisupervised learning as there are many domains where data are scarce and there are few annotations. Semisupervised learning utilizes both labeled and unlabeled data for model training. Following this line of research, Hussain and Cambria [20] presented a semisupervised framework for reasoning that improves emotion recognition and polarity detection. Park et al [21] propose a model that introduces a semisupervised sentiment-aware objective function using distant sentiment annotations for computing sentiment-oriented document representations. Interestingly, semisupervised learning that detects multilingual sentic patterns has also been applied for polarity estimation in English variants [22].

One of the problems that appear in sentiment analysis is word polarity disambiguation. This task is context- and domain-dependent, which makes it a nontrivial problem. To address this issue, Xia et al [23] presented an approach using a Bayesian model that exploits intra- and inter-opinion features to estimate the probabilities of word polarities. In another work, Vechtomova [24] explored the approach to this problem from an information retrieval perspective. This method computes the similarity of the query to the documents and finds the polarity that best matches.

In this work, we use an ensemble to augment the performance of the system. Recently, stacked ensembling has been more prominent in affective computing since it has been demonstrated to be an effective method [25]. As reported by Emre Isik et al [26], a novel model that uses two different classifiers and feature extraction methods successfully improves the performance of other approaches.

Akhtar et al [27] proposed an ensemble of several deep learning and classical models using a multilayer perceptron network. The networks used are based on convolutional neural network (CNN), long short-term memory (LSTM), and gated recurrent unit (GRU) architectures while the classical methods use a support vector machine (SVM) model. Among other applications, stacked ensembles are used to improve the results in low-resource languages, such as Arabic [28], Bengali [29], and Moroccan [30].

Conversely, in affective computing, other recent advances have been presented that do not use neural networks. In this field, generating sentiment and emotion lexica remains an open challenge that has profound effects in affective computing since these resources represent a reliable source of subjective knowledge. For example, [31] presented a method for generating domain-specific emotion lexica through the Unigram Mixture Model (UMM). Another approach for generating lexica was presented by [32] and describes the DepecheMood++ resource.

Affective Computing in Radicalism

Previous works that address the automatic processing of online radicalism can be divided into three categories: analysis, detection, and prediction [33, 34].

The objective of online radicalization analysis is to provide information that improves law enforcement agencies (LEA) in their decision-making process. Correa and Sureka [33] described that analysis solutions can be classified into two main categories: network-based and content-based. The first focuses on online communities, their leaders, and topological characteristics; the second tackles website activities, stylometric analysis, and authorship identification, and both affect and usage analysis.

The research focused on detection includes Web and text mining approaches [33]. Web mining solutions aim to detect radical online content by means of different techniques, such as focused crawling [35]. Text mining approaches are oriented to developing a machine learning-based classification model that exploits textual features, as presented by Rowe and Saif [36]. Such a representation can be combined with different features, such as social dynamics [34, 36].

Regarding detection, different types of analyses, including content and network-based analysis, have been proposed to improve the understanding of online radicalization. Content analysis focuses on analyzing various aspects of radical texts, such as stylistic features, aspects, and topics. In contrast, network-based analysis explores the social interactions in a community.

As previously mentioned, the third category for the study of online radicalism is prediction. Ferrara et al [37] proposed a machine learning-based system that detects extremist supporters by addressing two different problems. The first problem is the prediction of the adoption of extremist user content by measuring retweets of extremist content. The second problem is the prediction of the interaction toward extremist users by analyzing the replies to direct extremist messages. Therefore, this system considers three types of features: user metadata and activity, timing, and network statistics. In another relevant work, Agarwal and Sureka [38] presented a survey focused on two challenges. The first challenge addresses the automatic identification of online radicalization and studied hate promoting content, as well as users and hidden communities. The second challenge revolves around the prediction of civil unrest related events, such as protests, riots, and public demonstrations. This survey indicated that most works found that spatiotemporal features are effective at predicting events. Additionally, [39] described a system for the prediction of radicalization risk. The authors proposed generating alarms based on monitored users' radicalization influence and emotional loads of received tweets.

In this work, we exploit affective information to enhance the performance of a radicalization detection system. This approach has been followed by previous works, offering a range of insights. Affect analysis has been used in a variety of domains, including radical forums [40–42], radical magazines [43], and social networks such as Twitter [36, 44–46], Facebook [47], and YouTube [48].

As for the affective model used, many works exploit the sentiment analysis polarity (e.g., valence and sentiment) [41, 43–46, 48, 49]. However, other works utilize the intensity of terms concerning hate and violence [40]. Other works [42, 47] use LIWC's [42] categories for affective processing to exploit the information gathered from positive and negative emotions, anxiety, anger, and sadness.

In the process of applying sentiment analysis techniques, some interesting insights have been found in previous works. Abbasi and Chen [40] indicated a relationship between violence intensity scores and hate affects, remarking that the latter is weaker in publications in occidental forums than in those in Middle Eastern forums. Subsequently, Rowe and Saif [36] note that users immersed in the radicalization process often discuss political subjects, Syria and Egypt, with a negative tone. Next, once the users are considered radical, they tend to address more religious topics. Interestingly, the authors report that this kind of user tends to use the term ISIS with a negative sentiment, preferring using the alternative Islamic State. Another path analyzes the temporal evolution of text by comparing different radical blogs [40] or the language used in Dabiq, a radical magazine [43].

An interesting research topic is analyzing the public's reaction to a terrorist attack. Dewan et al [47] analyzed the sentiment of both images and texts extracted from Facebook. In this study, the authors observed that although the sentiment is initially negative during the first hours, it

changes toward a positive valence over time. Conversely, they also report the contrary effect for images.

In general, we can state that there is no agreement in the current research works that determines the importance of affect information in radicalization. However, several studies [41, 49, 50] that highlight its importance while other works [46] report the negative results of applying sentiment analysis. In this paper, we apply known affective techniques in order to improve radicalization detection performance.

Automatic Hate Speech Detection

Hate speech detection works can be categorized according to the type of hate speech addressed. Although the majority of previous works claim to tackle "general hate speech," some other authors refine their aim, addressing racism, sexism, and religion [51]. Regardless, the most descriptive way to analyze previous works in this research area is by studying their computational approaches.

A broad tendency in hate speech detection is adapting known text mining methods to this specific domain. This approach, as seen, is also common in radicalization detection. A prevalent method relies on the use of dictionaries, which consist of domain-related terms, possibly using their frequencies of appearance in real data. For the specific case of hate speech, known repositories are noswearing.com[1] and hatebase.org[2]. In addition to these types of terms, which include insults, reaction words, and swear words, previous works extract profane words [52], verbal abuse and stereotypical terms. Similarly, the Ortony lexicon, which contains a list of words that carry a negative connotation but not directly represent profane terms, has been generated [53]. This kind of resource can be exploited by considering the total number of words per document [52].

A specific characteristic of this domain is that offending words may be obscured with intentional misspellings, that is, frequent character alterations (e.g., b1tch, 4ss, and nagger) [54]. A distance-based metric can be used to detect these terms, and it can be complemented with dictionary-based approaches [55].

As one could expect, fundamental NLP methods have been successfully adapted to this domain. The bag-of-words (BoW) method has been used as a primary text representation method [56–58]. Of course, given BoW's limitation of ignoring the word order, some authors have used n-gram representations [56, 57, 59–61]. An interesting approach is computing character n-grams, which is more robust than word n-grams to spelling variations. In fact, [62] observed that character-based n-grams are more effective in detecting hate speech than word n-grams. As a natural extension of the n-gram representations, [53] used TF-IDF features, thus incorporating frequency information into their pipeline.

Following the trend in applying NLP methods to hate speech detection, [57] included part-of-speech (POS) features in a hate speech detection system. Similarly, Dinakar et al [53] were able to detect relevant POS-aware bigrams for hate speech. Interestingly, Burnap and Williams [63] detected meaningful text pieces such as send them home, should be hung, and get them out. However, POS-based features can cause confusion in hate category identification [51].

Another traditional NLP approach that has been applied to hate speech is word sense disambiguation (WSD). Warner and Hirschberg [64] employed a WSD technique to extract new knowledge from text to infer whether certain words are antisemitic or not. Concerning distilling information from text, Agarwal and Sureka [65] employed topic classification techniques to discover the topics that occur in a document. Specifically, several linguistic features were used to choose between race and religion topics.

As conducted in this paper, many authors have incorporated affect information into learning systems for hate speech. Agarwal and Sureka [65] added emotion features to an ensemble of an array of feature types. For emotion analysis, the authors used the IBM Tone Analyzer API[3]. Similarly, Davidson et al [60] added sentiment features to their feature combination. They used VADER, a sentiment lexicon, and a method that annotates Twitter messages with a sentiment score [66]. Del Vigna et al [67] studied hate speech in the Italian language and incorporated an Italian sentiment lexicon and two other English lexicons that have been translated. Additionally, the authors included a resource that has been created using word embeddings. To incorporate subjectivity analysis, Gitari et al [68] proposed a method that uses known sentiment lexicons in conjunction with semantic features to generate an additional lexicon used for hate speech detection. Another work that combines sentiment features with other classical text representations was presented by Liu and Forss [61]. This work exploited the SentiStrength sentiment lexicon [69], finding that negative sentiments are a better discriminator than neutral sentiments for hate speech detection.

Finally, concerning this paper, some works have used word embeddings for hate speech detection. Djuric et al [70] used a paragraph2vec [71] approach to classify the language of user comments as abusive or clean and to predict the central word belonging to a specific message. As in this paper, [59] used the FastText model [72] and propose several models based on CNN and

[1] http://www.noswearing.com/

[2] https://hatebase.org/

[3] https://tone-analyzer-demo.mybluemix.net/

LSTM architectures. A known problem when using word embeddings is that the classification is performed at the document level. Thus, when extracting word embeddings, a dimensionality problem appears. That is, one dimension must be reduced in order to feed the extracted features to classical classifiers. In this line of work, [54] explored the average over the word vectors, showing that although it is a practical solution, its performance is not very high. As a possible solution, [70] proposed a method for computing comment embeddings.

Regarding the use of deep learning techniques, [73] explored the effectiveness of several neural approaches at providing improvements over more traditional approaches. In particular, they experimented with CNN, LSTM, biLSTM, and multilayer perceptron architectures. Another recent work that addresses the use of neural networks was presented by [74], who proposed using a stacked architecture of CNN and gated recurrent units. This work reported that the proposed model can detect implicit features that aid in the detection of hate speech signals.

In an interesting work, [75] collected a massive dataset and studied hate speech spread across users. Their work used a lexicon to annotate the dataset instances. From that annotation, the authors presented several user and network characteristics that can be used to analyze the spread of hate speech. Such an approach can be combined with NLP techniques to improve the robustness of learning systems.

Proposed Methods

This paper proposes the use of both affect resources and generic textual representations to tackle the challenges of radicalization and hate speech analysis detection. Previous works in these two domains indicate that using sentiment and emotion information can improve the detection performance due to the nature of the tasks. Therefore, we use the sentic resources described in Sect. 3.1. Regarding the generic textual representations, we use two approaches, which are presented in Sect. 3.2.

All these feature extraction methods compute a set of representations that are later combined in an ensemble fashion by means of vector concatenation. This combined vector is then fed into a machine learning algorithm trained to predict the task labels using these representations. An overview of the proposed architecture is shown in Fig. 1.

AffectiveSpace and SenticNet Exploitation

This paper proposes exploiting AffectiveSpace [8] and SenticNet [9] through two novel methods that generate text representations. These proposed methods utilize the specific characteristics of the mentioned resources in order to compute affect-aware features. In this sense, these models do not aim to obtain complete representations of the text but rather are oriented to distilling affect knowledge.

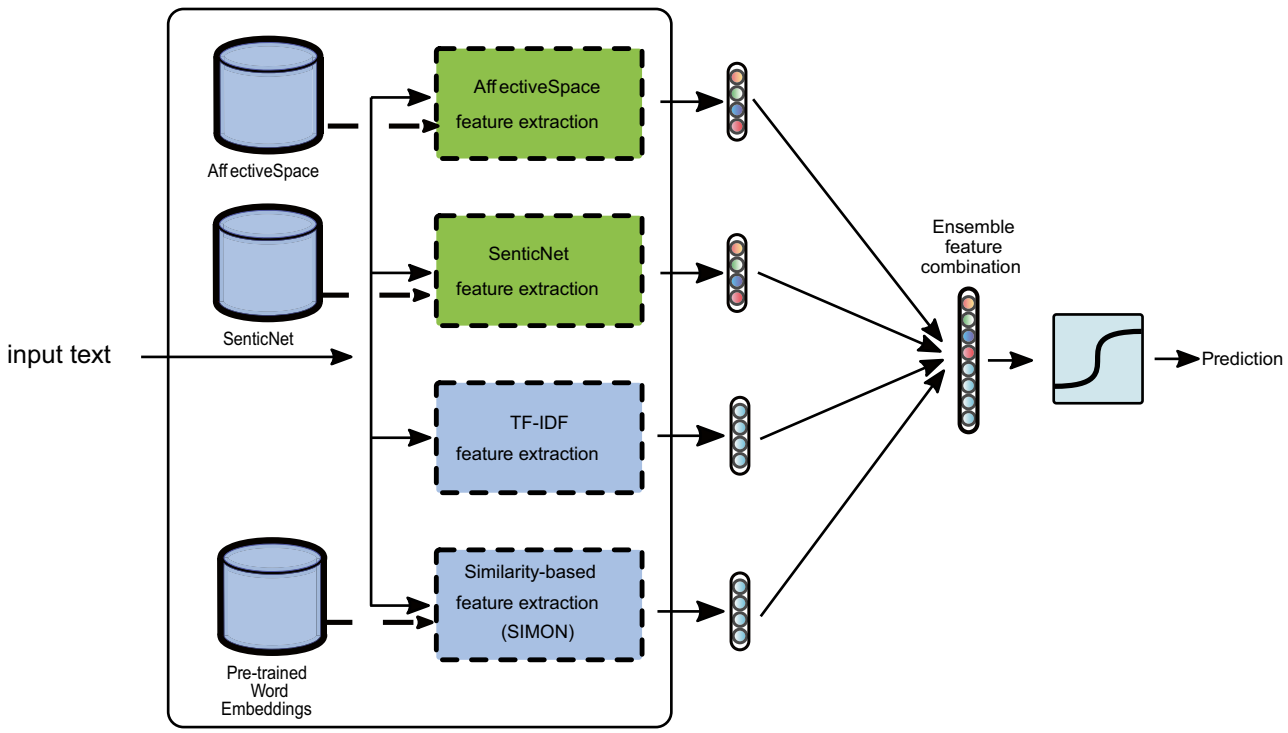

Fig. 1 Proposed architecture

Table 1 AffectiveSpace and SenticNet vocabulary distribution of n-grams

n-gram	No. of occurrences
1	39,889
2	51,859
3	7,773
4	455
5	23

AffectiveSpace First, we utilize the AffectiveSpace resource, developed by [8], which contains a large number of concepts embedded into a vector space. This resource represents an NLP framework that can potentially be embedded in a general purpose cognitive system. AffectiveSpace has a total of 100,000 concepts that are represented by vectors with a dimension of 100. These concepts are described by terms that range from 1-grams to 5-grams with the distribution shown in Table 1.

To utilize AffectiveSpace, it is necessary to perform the following steps: (i) lemmatize the text words and then (ii) extract the n-grams from the processed text. Once this process in complete, we extract the resource's vector representations for an analyzed document D composed of m n-grams. We then construct a matrix with dimensions $m \times d$, where d is the dimension of the AffectiveSpace vectors. Finally, to reduce the dimensionality of the matrix and obtain a feature vector representing D, we compute the average of the m vectors, resulting in a vector of d dimensions.

SenticNet Another sentic resource we use in this paper is SenticNet[4], developed by [9]. SenticNet contains an extensive database that represents the same 100,000 concepts described above. These concepts are linked through a semantic network, offering an additional dimension over other affect resources. Instead of using embedded representations, each concept has an associated list of characteristics: the pleasantness, attention, sensitivity, and aptitude values; a primary and a secondary mood; the polarity label and value; and the relation to five other concepts as given by the semantic network.

As in previous works, we use this resource to extract features that can be used for text classification. Thus, it is also necessary to compute the lemmatization and extract the n-grams, accommodating the text to the concept representations of SenticNet. Additionally, since not all characteristics in SenticNet are numerical, we represent the categorical characteristics using one-hot encoding. The transformed categories are used as the primary and secondary moods and the polarity labels. In addition, the semantic relations to other concepts are omitted. With this processing, each n-gram is represented by a vector of 22 dimensions that retains rich affective knowledge.

Thus, similar to before, for a document D composed of m n-grams, we extract the representations for each n-gram, obtaining a matrix with dimensions $m \times 22$. As in both [76] and [77], we then construct a feature vector by applying a statistical summary of the previous matrix. Following previous observations, we choose the average and maximum as the summary functions, which are computed columnwise. Finally, we obtain a feature vector with 44 dimensions, since we apply two summary functions.

Textual Feature Extraction Methods

As described above, we use generic textual representations that are later combined with the previous affect-driven features. The chosen domains (radical and hate speech detection) represent complex domains with their idiosyncrasies, and using a unified set of feature extraction methods does not always achieve good results. To avoid decreasing the quality of the representations, we use two methods that have been studied in these domains and have been demonstrated to provide good performance.

First, we use the SIMilarity-based sentiment projectiON (SIMON) model [7]. Although this method was initially proposed in the sentiment analysis context, we propose applying this method as a feature extractor in the target domains. This method uses a word embedding model and orients the extracted features to a particular domain utilizing a domain-centered lexicon. SIMON uses a pretrained word embedding model to compute the similarity between the analyzed text's words and a selection of domain-related words. Thus, the input text is measured against a domain vocabulary by computing a vector that encodes the similarity between the input text and the lexicon. This model can leverage the information contained in both a word embedding model and a domain lexicon. Additionally, the training process does not necessarily need large corpora and thus can be used in problems where annotated data are scarce. SIMON has been studied previously in radicalization detection [76], moral value estimation [77], and hate speech analysis [78].

Second, we use the TF-IDF method [79]. This kind of representation is robust, suitable for text classification, and adaptable to almost any domain. Such a method provides a reference point for the evaluation by offering a comparison baseline.

Evaluation

The different feature extraction methods were evaluated through several text categorization tasks, where the aim is to predict the associated label for a given document. In this

[4] We use the 5th version. A new version of the resource, SenticNet 6 by [12], has been recently released and could be used for this model.

Table 2 Statistics of the used datasets. Number of instances, number of classes, class balance (percentage), and average number of words per instance

Dataset	No. of instances	No. of classes	Class balance (%)	Avg. no. of words	Source	Domain
Pro-Neu	224	2	50/50	18,646	Twitter	Radicalization
Pro-Anti	1,132	2	50/50	36,352	Twitter	Radicalization
Magazines	468	2	68/32	950	Magazines	Radicalization
SemEval2019	10,000	2	58/42	26	Twitter	Hate speech
Davidson	24,783	3	77/17/6	17	Twitter	Hate speech

way, these tasks belong to two different research domains: radicalization and hate speech detection. Therefore, the proposed methods were validated using the materials listed in Sect. 4.1, and following the methodology described in Sect. 4.2. The obtained results of these experiments are shown in Sect. 4.3.

Materials

Table 2 presents the datasets used in the evaluation. In total, we use five datasets. The majority of these data were extracted from Twitter.

Pro-Neu. This dataset was generated by combining two different English datasets, which were collected by [34]. The first set consists of 17,350 tweets extracted from 112 different Twitter accounts that support ISIS. The list can be found online[5]. By means of a study that spanned three months, a collection of users was identified using a number of keywords (e.g., Wilayat, Amaq, and Dawla), and filtered according to how said users used images (e.g., ISIS flags and radical leaders images), as well as their follower network. The second set, which contains 122k tweets from more than 95k different accounts, has been utilized as a counterexample, offering a reference to the pro-ISIS instances. It contains ISIS-related messages that may be either neutral or anti-ISIS. This last set of tweets was obtained using ISIS-related keywords (e.g., ISIL, ISIS, IslamicState, Daesh, Mosul, and Raqqa). Filtering was performed using the original accounts, as performed by [34], retaining 112 users. This additional filtering was performed by removing accounts that were not recently active from the dataset. Therefore, this process ensured that the remaining accounts were not pro-ISIS. We made the same selection and split as in [34] and [76].

Pro-Anti was generated from 1,132 Twitter accounts and their timelines, which were collected by [36]. As before, this dataset is in English. [36] identified users as pro-ISIS by measuring their sharing activity of incitement material from known pro-ISIS users, as well as their use of extreme language. First, [36] identified 727 accounts, but 161 of these Twitter users were either hidden or suspended from public access. This situation prevents further attempts to access their profile information. Thus, these 161 accounts were removed, resulting in 566 pro-ISIS users in total. To balance the data, [36] added 566 anti-ISIS users. The annotation of anti-ISIS accounts was performed by observing the use of anti-ISIS language.

Magazines. This dataset was presented by [76]. These data are from the Dabiq [80] and Rumiyah [81] online magazines, which are published by the Islamic State of Iraq and the Levant (ISIS) radical organization [82] and are written in English. Dabiq was released from July 2014 to July 2016 by a branch of ISIS's Ministry of Media. After producing fifteen issues of Dabiq, the same organization released the first issue of Rumiyah in September 2016. In total, thirteen issues of Rumiyah were released until September 2017. As a comparison point of the previous radical text, the dataset contains two online newspapers that address ISIS-related issues but are not radical sources: Cable News Network (CNN)[6] and The New York Times[7]. This content can be freely downloaded through the newspapers' APIs. These data were obtained using domain-based keywords (Daesh, ISIS, Islamic State, etc.). As part of the data processing, articles that were not related to the topic were manually filtered. Additionally, images, links, and other media were removed, leaving the text. In total, 129 articles were collected from CNN, and 23 articles were collected from The New York Times. For more information on this dataset, please read [76].

SemEval2019. This dataset, which is part of the international semantic evaluation SemEval 2019, has been obtained from the work of [83]. The data were extracted from Twitter and a previous dataset on misogyny identification [84]. Three different methods were used for the collection: potential hate victim monitorization, downloading identified haters histories, and filtering the Twitter stream using keywords. Some examples of these

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

[6] https://cnn.com

[7] https://nytimes.com

keywords are migrant, refugee, and bitch. Although the original dataset is multilingual, we selected the English data. Additionally, since it is not possible to access the test data, we used the training and development datasets and joined them. To provide a reference to the original challenge, the selected annotations were those of Task A, where the models were trained to predict the presence of hate speech in a binary task [83].

Davidson. The dataset was constructed by [60] for hate speech detection. It was extracted from Twitter, using a collection of hate-related words, which are compiled in hatebase.org. After obtaining a random subsample of the original data downloaded, the authors instructed crowd workers to annotate these data into three categories: hate speech, offensive speech, and neither hate nor offensive speech. The original work [60] reported that the percentage of tweets annotated as hate speech was unusually low, probably due to the criteria used to identify hate speech. The full dataset can be downloaded from GitHub[8].

In addition these datasets, other resources used in this work are the AffectiveSpace and SenticNet frameworks (see Sect. 3.1), which can be downloaded from sentic.net[9]. Finally, as described in Sect. 3.2, the SIMON method uses a word embedding model. Following the insights obtained in [76], we selected the FastText word embedding model, presented by [72].

Methods

The main objective of the experiment is to evaluate the different proposed feature extraction methods. Therefore, the classification step can be implemented with a variety of resources and methods. An extensive experimental setup was designed to thoroughly evaluate the effectiveness of the feature extraction methods. In all experiments, we use the weighted average of the F1 score as the performance metric. For each dataset, k-fold cross-validation is performed, where $k = 10$.

We study the different features and their combinations. Mainly, we aim to evaluate whether affect features can enhance the performance of the proposed tasks. To accomplish this, as described in Sect. 3.2, we also evaluate generic feature extraction methods that we combine with the proposed affect strategies. For the machine learning algorithms, we select logistic regression and SVM with a linear kernel since the main objective is to assess the effectiveness of the studied features. This finding is in line with previous research [76].

To further assess the impact of affect knowledge on the considered tasks, we add a variation of the SIMON method. As explained, SIMON utilizes a domain lexicon to represent a text given its similarities to the words of said lexicon. To explore the relevance of an affect vocabulary in this setting, we implement a SIMON variation that uses the AffectiveSpace and SenticNet vocabulary. In this paper, we denote this variation as SIMON SenticNet. The original method, which exploits a domain lexicon, is called SIMON domain. Such a domain lexicon has been obtained as in [76], using frequency-based filtering of the word occurrences in each dataset.

To encourage research, we have published the code for all the methods and experiments on GitHub[10].

Results

Table 3 shows the results of the full evaluation. As described above, we aim to assess the performance of the different proposed feature extraction methods.

Focusing on the feature methods with no combinations, it can be seen that the TF-IDF and SIMON approaches obtain strong baseline results. This result agrees with the previous research. First, TF-IDF is a fundamental method used to represent text accounting for the internal frequencies of tokens on documents, and it consistently obtains strong performance. Second, the SIMON model has shown to achieve high-performance scores in diverse text categorization tasks, including radicalization detection [76], moral value estimation [77], and hate speech detection [78]. When incorporating the affect features, the overall performance is comparatively lower. This decrease is to be expected, as the proposed feature extraction methods that exploit the AffectiveSpace and SenticNet resources are limited. Additionally, as shown in previous research [76], a method that includes only affect information with no domain knowledge usually achieves lower performances. This finding indicates the importance of including domain knowledge for this kind of task. These observations are consistent across the two machine learning models evaluated.

Another aspect to be considered is the difference between the SIMON method when using a domain lexicon (SIMON domain) and when using an affect vocabulary (SIMON SenticNet). Following the above results, we observe (Table 3) that the domain variant (SIMON domain) obtains higher scores across all datasets. This, as explained, further suggests the importance of incorporating domain-oriented knowledge

[8] https://github.com/t-davidson/hate-speech-and-offensive-language

[9] https://sentic.net/downloads/

[10] https://github.com/gsi-upm/sentic-computing-radical-hate

Table 3 Averaged F1-scores for the studied features, using the logistic regression and linear SVM classifiers

LOGISTIC REGRESSION					
Features	Pro-Neu	Pro-Anti	Magazines	SemEval19	Davidson
TF-IDF	86.61	84.63	88.89	75.40	89.33
AffectiveSpace	87.05	72.53	67.52	64.49	81.45
SenticNet	70.98	71.64	70.30	65.15	80.61
SIMON domain	97.77	86.66	94.02	72.74	90.54
SIMON SenticNet	96.43	82.77	88.25	71.23	89.90
AffectiveSpace + SenticNet	70.98	72.00	70.30	67.31	83.19
TF-IDF + AffectiveSpace	87.05	84.81	89.10	75.65	89.48
TF-IDF + SenticNet	90.62	86.75	86.75	75.07	89.02
TF-IDF + AffectiveSpace + SenticNet	90.62	86.93	86.75	75.30	89.23
SIMON domain + AffectiveSpace	97.77	86.66	94.02	72.95	90.49
SIMON domain + SenticNet	97.32	86.66	94.23	72.86	90.51
SIMON domain + AffectiveSpace + SenticNet	97.32	86.66	94.23	72.92	90.52
SIMON SenticNet + AffectiveSpace	96.43	82.77	88.25	72.32	90.15
SIMON SenticNet + SenticNet	96.43	82.16	88.03	71.91	90.17
SIMON SenticNet + AffectiveSpace + SenticNet	96.43	82.16	88.03	72.40	90.19
SVM					
Features	Pro-Neu	Pro-Anti	Magazines	SemEval19	Davidson
TF-IDF	95.98	87.54	94.02	73.98	89.76
AffectiveSpace	86.61	72.88	67.74	65.29	82.22
SenticNet	83.48	72.26	70.51	65.10	80.65
SIMON domain	98.21	85.25	94.66	72.81	90.58
SIMON SenticNet	98.21	80.21	86.75	71.25	89.97
AffectiveSpace + SenticNet	85.71	73.67	69.66	67.60	83.69
TF-IDF + AffectiveSpace	95.98	88.43	94.44	73.93	89.74
TF-IDF + SenticNet	97.32	90.81	93.80	74.08	89.80
TF-IDF + AffectiveSpace + SenticNet	97.32	90.90	93.80	74.16	89.82
SIMON domain + AffectiveSpace	98.21	85.34	94.44	72.45	90.63
SIMON domain + SenticNet	98.21	85.69	94.44	72.85	90.62
SIMON domain + AffectiveSpace + SenticNet	98.21	85.51	94.44	72.69	90.69
SIMON SenticNet + AffectiveSpace	98.21	80.39	86.75	72.10	90.32
SIMON SenticNet + SenticNet	98.21	79.68	86.11	71.79	90.28
SIMON SenticNet + AffectiveSpace + SenticNet	98.21	79.95	86.11	72.48	90.28

into a learning system. This observation also arises in previous research [76, 78].

Next, we focus on the different combinations of the studied features. Given the complexity of assessing such comparisons across considerable numbers of datasets and methods, we conduct the Friedman statistical test [85]. This statistical test outputs a ranking of methods that aggregates their performance across all datasets. For this method, the lower the numerical ranking, the better a method performs compared to the rest. We conduct the test with $\alpha = 0.01$. Table 4 shows the ranking of the studied combinations considering the two classifiers.

First, it can be seen that the combination of SIMON using a domain vocabulary with the features extracted by our SenticNet method using the SVM classifier achieves the best ranking. A similar phenomenon can be observed when assessing the logistic regression, where, in this instance, the combination with our AffectiveSpace feature extraction provides better performance compared to the SIMON domain. Indeed, when assessing the TF-IDF comparison, the Friedman test indicates that combining the affect features with SenticNet and AffectiveSpace improves the performance compared to just using TF-IDF. Again, this finding is

Table 4 Friedman rank for all the proposed methods. LogR is logistic regression, and LinSVM is SVM with linear kernel

Approach	Friedman Rank
LinSVM SIMON domain + SenticNet	6.8
LinSVM SIMON domain + AffectiveSpace + SenticNet	7.2
LinSVM SIMON domain	7.3
LinSVM SIMON domain + AffectiveSpace	8.0
LogR SIMON domain + AffectiveSpace + SenticNet	8.7
LogR SIMON domain + AffectiveSpace	8.8
LogR SIMON domain + SenticNet	9.1
LogR SIMON domain	9.2
LinSVM TF-IDF + AffectiveSpace + SenticNet	9.4
LinSVM TF-IDF + SenticNet	10.0
LinSVM TF-IDF + AffectiveSpace	10.8
LinSVM TF-IDF	11.7
LogR TF-IDF + AffectiveSpace + SenticNet	14.6
LogR TF-IDF + AffectiveSpace	14.7
LinSVM SIMON SenticNet + AffectiveSpace	15.0
LogR TF-IDF + SenticNet	15.2
LinSVM SIMON SenticNet + AffectiveSpace + SenticNet	15.5
LogR TF-IDF	15.9
LogR SIMON SenticNet + AffectiveSpace	16.5
LogR SIMON SenticNet + AffectiveSpace + SenticNet	16.7
LinSVM SIMON SenticNet + SenticNet	16.9
LinSVM SIMON SenticNet	17.0
LogR SIMON SenticNet + SenticNet	17.5
LogR SIMON SenticNet	17.9
LinSVM AffectiveSpace + SenticNet	26.0
LinSVM AffectiveSpace	26.9
LogR AffectiveSpace + SenticNet	27.4
LogR AffectiveSpace	27.7
LinSVM SenticNet	27.8
LogR SenticNet	28.8

consistent across the two classifiers. In light of these results, we conclude that adding affect knowledge through the SenticNet resources effectively improves the classification performance on the tasks at hand. This result, as indicated before, is in line with previous research works.

In this work, we use the Friedman test as a statistical measure of the validity of our results and as a method of aggregating the numerous scores of the experiments. However, the results in Table 3 are also interesting. When the SIMON domain and TF-IDF methods are compared, the Friedman rank indicates their separation. However, the performance increases of SIMON compared with TF-IDF are not very large when both are combined with SenticNet features (Table 3). For example, when SIMON domain + SenticNet is compared with TF-IDF + AffectiveSpace + SenticNet (the best combination of SIMON and TF-IDF), the performance scores are similar except for the Pro-Anti dataset.

This last observation requires a different comparison that considers the execution complexity of the proposed methods. Although a particular approach outperforms another in terms of accuracy scores, it is interesting to study the complexity of these methods. In this paper, we evaluate the complexity by measuring the execution times of our implementation. The results are shown in Table 5.

As seen, this table includes three types of execution times: loading time, which represents the time necessary to load necessary resources; preprocessing, which summaries the specific computations needed by some of the proposed methods; and feature extraction, which aggregates all the operations needed to extract the features for each dataset. Importantly, since the SIMON method uses a word embedding model, it is necessary to load the word vectors (loading time and word embeddings). This model can later be reutilized by all SIMON instances in our implementation; thus, the loading time occurs only once. Additionally, the combination time is negligible and is thus not included in the study. All computations were performed using equipment with 20 CPUs, 120GB of RAM, and an SSD drive.

Following this comparison methodology, we can compare the SIMON-based methods with the other methods, and the results show that their execution times when performing feature extraction are higher. This effect is evident with the Pro-Anti dataset due to its larger number of words per instance (Table 2). This difference is approximately one order of magnitude. The same effect can be seen for the Pro-Neu dataset.

After considering these observations, it is safe to conclude that although combinations with SenticNet obtain the best features according to the performance metrics with the SIMON model, the feature complexity of TF-IDF combined with SenticNet is lower. However, selecting a definitive model is impossible as users may have different considerations for these two measurement strategies during model selection. In general, when computational resources are abundant, the default model supported by the evaluation is the SIMON domain + SenticNet. In contrast, when computational resources are scarce, a good alternative is the TF-IDF + AffectiveSpace + SenticNet model.

Table 5 Measures of execution times (in seconds)

	Loading time (s)				
AffectiveSpace	9.3				
SenticNet	2.3				
Word Embeddings	590				
	Preprocessing				
	Pro-Neu	Pro-Anti	Magazines	SemEval19	Davidson
N-gram computation	17.3	182	1.7	1.3	2
SIMON domain	0.9	10.2	0.1	0.1	0.1
SIMON SenticNet	9.9	22	9.5	10	8.9
	Feature Extraction				
	Pro-Neu	Pro-Anti	Magazines	SemEval19	Davidson
AffectiveSpace	2.4	20.9	0.5	0.7	1.3
SenticNet	2.1	18	0.5	0.7	1.3
TF-IDF	2.7	26.5	0.3	0.3	0.5
SIMON domain	29	288.9	3.2	3.3	6.7
SIMON SenticNet	29	295.2	3.2	3.3	6.8

Conclusions

This paper proposes a machine learning framework that exploits the textual and affect features extracted from text to conduct radical and hate speech detection. In particular, we design two methods for extracting affect features that exploit the sentic resources AffectiveSpace and SenticNet, which constitute rich sources of information. In addition, these methods are combined with domain textual representations in an effort to enhance the overall performance on the tasks at hand. The evaluation results show that such a combination is effective and that adding affect information through the mentioned resources does increase the classification performance. As an additional verification, we performed a statistical test that further supports the obtained results.

Previously, this paper raised two research questions related to this research work. First, RQ1 asked whether affect-aware systems and, more concretely, sentic computing resources can improve performance in NLP application domains. Here, the experiments show that adding the affect features improves the prediction performance. This improvement can be seen in the Friedman test results, which organize the different feature combinations in a unified ranking. Indeed, when assessing the SIMON and TF-IDF feature methods, one can observe that adding the affect-aware features improves the ranking and thus the performance on the studied tasks.

Next, RQ2 addresses a criterion that allows us to select among the variety of studied combination methods while considering both the classification performance and the computational complexity of a model. Therefore, this paper has presented a detailed study of the models' complexity by assessing the execution times across all datasets. This experiment shows that although using the SIMON model in combination with affect features achieves the best performance scores, the complexity of this model is higher than that of the alternative TF-IDF representations. To address this limitation, we delineate a possible criterion considering computational resources (such as computation time and memory) and the overall classification performance. In any case, this selection consistently supports the combination with affect features.

The obtained results indicate a path for incorporating affect-oriented knowledge into learning systems for NLP application domains. However, we believe that this work also presents new challenges that can be addressed by future work. One of these challenges is to improve the combination method. In this paper, we selected vector concatenation, but more sophisticated methods, including neural operations, could be explored. In another line of work, we used SenticNet resources, which contain a rich network of concepts. This characteristic could be included in a system similar to ours to exploit an additional source of knowledge that could improve the robustness of the system. Additionally, the 6th version of the SenticNet resource was recently released. Future work should assess whether this new version produces improved results.

Acknowledgements The authors would like to thank Miriam Fernandez and Harith Alani for sharing part of the data used in this research.

Funding This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under project Participation (grant agreement no. SEP-210655026) and by the Spanish

Ministry of Science and Innovation through project COGNOS (PID2019-105484RB-I00).

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

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

References

1. Hendler J, Shadbolt N, Hall W, Berners-Lee T, Weitzner D. Web science: an interdisciplinary approach to understanding the web. Commun ACM. 2008;51(7):60–9. https://doi.org/10.1145/1364782.1364798.
2. Cambria E, White B. Jumping NLP curves: A review of natural language processing research. IEEE Comput Intell Mag. 2014;9(2):48–57. https://doi.org/10.1109/MCI.2014.2307227.
3. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, Zhou Q. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput. 2016;8(4):757–71. https://doi.org/10.1007/s12559-016-9421-9.
4. Tao J, Tan T. Affective computing: A review. In International Conference on Affective computing and intelligent interaction. Springer, 2005. pp. 981–95. https://doi.org/10.1007/11573548_125.
5. Crowston K, Allen EE, Heckman R. Using natural language processing technology for qualitative data analysis. Int J Soc Res Methodol. 2012;15(6):523–43. https://doi.org/10.1080/13645579.2011.625764.
6. Cambria E, Hussain A. Sentic computing: A common-sense-based framework for concept-level sentiment analysis. Cogn Comput. 2015;7:183–5. https://doi.org/10.1007/s12559-015-9325-0.
7. Araque O, Zhu G, Iglesias CA. A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowl-Based Syst. 2019;165:346–59. https://doi.org/10.1016/j.knosys.2019.105184 http://www.sciencedirect.com/science/article/pii/S095070511930526X.
8. Cambria E, Fu J, Bisio F, Poria S. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2015. pp. 508–14.
9. Cambria E, Poria S, Hazarika D, Kwok K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Thirty-Second AAAI Conference on Artificial Intelligence. 2018. pp. 1795–802. https://doi.org/10.1109/MIS.2017.4531228.
10. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. IEEE Intell Syst. 2017;32(6):74–80. https://doi.org/10.1109/MIS.2017.4531228.
11. Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–7. https://doi.org/10.1109/MIS.2016.31.
12. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. SenticNet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. CIKM'20, Oct 20-24. 2020. pp. 105–14. https://doi.org/10.1145/3340531.3412003.
13. Dragoni M, Poria S, Cambria E. Ontosenticnet: A commonsense ontology for sentiment analysis. IEEE Intell Syst. 2018;33(3):77–85. https://doi.org/10.1109/MIS.2018.033001419.
14. Weichselbraun A, Gindl S, Fischer F, Vakulenko S, Scharl A. Aspect-based extraction and analysis of affective knowledge from social media streams. IEEE Intell Syst. 2017;32(3):80–8. https://doi.org/10.1109/MIS.2017.57.
15. Chen M, Wang S, Liang PP, Baltrušaitis T, Zadeh A, Morency LP. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction. 2017. pp. 163–71. https://doi.org/10.1145/3136755.3136801.
16. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, sep 2017. Association for Computational Linguistics. pp. 1103–14. https://doi.org/10.18653/v1/D17-1115. https://www.aclweb.org/anthology/D17-1115.
17. Chen X, Sun Y, Athiwaratkun B, Cardie C, Weinberger K. Adversarial deep averaging networks for cross-lingual sentiment classification. Transactions of the Association for Computational Linguistics. 2018;6:557–70. https://doi.org/10.1162/tacl_a_00039.
18. Esuli A, Moreo A, Sebastiani F. Cross-lingual sentiment quantification. IEEE Intell Syst. 2020;35(3):106–14. https://doi.org/10.1109/MIS.2020.2979203.
19. Liu R, Shi Y, Ji C, Jia M. A survey of sentiment analysis based on transfer learning. IEEE Access. 2019;7:85401–12. https://doi.org/10.1109/ACCESS.2019.2925059.
20. Hussain A, Cambria E. Semi-supervised learning for big social data analysis. Neurocomputing. 2018;275:1662–73. https://doi.org/10.1016/j.neucom.2017.10.010 http://www.sciencedirect.com/science/article/pii/S0925231217316363.
21. Park S, Lee J, Kim K. Semi-supervised distributed representations of documents for sentiment analysis. Neural Netw. 2019;119:139–50. https://doi.org/10.1016/j.neunet.2019.08.001 http://www.sciencedirect.com/science/article/pii/S0893608019302187.
22. Lo SL, Cambria E, Chiong R, Cornforth D. A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. Knowl-Based Syst. 2016;105:236–47. https://doi.org/10.1016/j.knosys.2016.04.024 http://www.sciencedirect.com/science/article/pii/S0950705116300764.
23. Xia Y, Cambria E, Hussain A, Zhao H. Word polarity disambiguation using bayesian model and opinion-level features. Cogn Comput. 2015;7(3):369–80. https://doi.org/10.1007/s12559-014-9298-4.
24. Vechtomova O. Disambiguating context-dependent polarity of words: An information retrieval approach. Inf Process Manag. 2017;53(5):1062–79. https://doi.org/10.1016/j.ipm.2017.03.007 http://www.sciencedirect.com/science/article/pii/S0306457316305416.
25. Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications. 2017;77:236–46. https://doi.org/10.1016/j.eswa.2017.02.002. http://www.sciencedirect.com/science/article/pii/S0957417417300751.
26. Emre Isik Y, Görmez Y, Kaynar O, Aydin Z. Nsem: Novel stacked ensemble method for sentiment analysis. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). 2018. pp. 1–4. https://doi.org/10.1109/IDAP.2018.8620913.
27. Akhtar MS, Ekbal A, Cambria E. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble. IEEE Comput Intell Mag. 2020;15(1):64–75.
28. Al-Azani S, El-Alfy ESM. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. In ANT/SEIT. 2017. pp. 359–366. https://doi.org/10.1016/j.procs.2017.05.365
29. Sarkar K. A stacked ensemble approach to bengali sentiment analysis. In: Tiwary US, Chaudhury S, editors. Intelligent

Human Computer Interaction., ppCham: Springer International Publishing; 2020. p. 102–111.
30. Oussous A, Lahcen AA, Belfkih S. Improving sentiment analysis of moroccan tweets using ensemble learning. In International Conference on Big Data, Cloud and Applications. Springer, 2018. pp. 91–104. https://doi.org/10.1007/978-3-319-96292-4_8.
31. Bandhakavi A, Wiratunga N, Massie S, Padmanabhan D. Lexicon generation for emotion detection from text. IEEE Intell Syst. 2017;32(1):102–8.
32. Araque O, Gatti L, Staiano J, Guerini M. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. IEEE Trans Affect Comput. 2019. pp. 17877–91. https://doi.org/10.1109/TAFFC.2019.2934444.
33. Correa D, Sureka A. Solutions to detect and analyze online radicalization: a survey. arXiv preprint 2013. arXiv:1301.4916.
34. Fernandez M, Asif M, Alani H. Understanding the roots of radicalisation on Twitter. In Proceedings of the 10th ACM Conference on Web Science, WebSci '18, pp. 1–10, New York, NY, USA, 2018. ACM. http://doi.acm.org/10.1145/3201064.3201082.
35. Agarwal S, Sureka A. Topic-specific youtube crawling to detect online radicalization. In International Workshop on Databases in Networked Information Systems. Springer, 2015. pp. 133–51. https://doi.org/10.1007/978-3-319-16313-0_10.
36. Rowe M, Saif H. Mining pro-isis radicalisation signals from social media users. In Proceedings of the tenth international AAAI conference on web and social media (ICWSM 2016). pp. 329–38.
37. Ferrara E, Wang WQ, Varol O, Flammini A, Galstyan A. Predicting online extremism, content adopters, and interaction reciprocity. In International conference on social informatics. Springer, 2016. pp. 22–39. https://doi.org/10.1007/978-3-319-47874-6_3.
38. Agarwal S, Sureka A. Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. arXiv preprint 2015. arXiv:1511.06858.
39. López-Sáncez D, Revuelta J, de la Prieta F, Corchado JM. Towards the automatic identification and monitoring of radicalization activities in twitter. In International Conference on Knowledge Management in Organizations. Springer, 2018. pp. 589–99. https://doi.org/10.1007/978-3-319-95204-8_49.
40. Abbasi A, Chen H. Affect intensity analysis of dark web forums. In 2007 IEEE Intelligence and Security Informatics. IEEE, 2007. pp. 282–8. https://doi.org/10.1109/ISI.2007.379486.
41. Chalothorn T, Ellman J. Affect analysis of radical contents on web forums using sentiwordnet. International Journal of Innovation Management and Technology. 2013;4(1):122–4.
42. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates, 71(2001):2001.
43. Vergani M, Bliuc A-M. The evolution of the ISIS language: a quantitative analysis of the language of the first year of Dabiq magazine. Sicurezza, Terrorismo e Società Security, Terrorism and Society. 2015;2(2):7–20.
44. Ghajar-Khosravi S, Kwantes P, Derbentseva N, Huey L. Quantifying salient concepts discussed in social media content: A case study using twitter content written by radicalized youth. Journal of Terrorism Research. 2016;7(2):79–90. https://doi.org/10.15664/jtr.1241.
45. Jurek A, Mulvenna MD, Bi Y. Improved lexicon-based sentiment analysis for social media analytics. Security Informatics. 2015;4(1):1–13. https://doi.org/10.1186/s13388-015-0024-x.
46. Saif H, Dickinson T, Kastler L, Fernandez M, Alani H. A semantic graph-based approach for radicalisation detection on social media. In European Semantic Web Conference. Springer, 2017. pp. 571–87. https://doi.org/10.1007/978-3-319-58068-5_35.
47. Dewan P, Suri A, Bharadhwaj V, Mithal A, Kumaraguru P. Towards understanding crisis events on online social networks through pictures. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2017. pp. 439–46. https://doi.org/10.1145/3110025.3110062.
48. Bermingham A, Conway M, McInerney L, O'Hare N, Smeaton AF. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in. IEEE, 2009. pp. 231–6. https://doi.org/10.1109/ASONAM.2009.31.
49. Agarwal S, Sureka A. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology. Springer, 2015. pp. 431–42. https://doi.org/10.1007/978-3-319-14977-6_47.
50. Ashcroft M, Fisher A, Kaati L, Omer E, Prucha N. Detecting jihadist messages on twitter. In Intelligence and Security Informatics Conference (EISIC), 2015 European, IEEE, 2015. pp. 161–4. https://doi.org/10.1109/EISIC.2015.27.
51. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Comput Surv. 2018;51(4):7. https://doi.org/10.1145/3232676.
52. Dadvar M, Jong FD, Ordelman R, Trieschnigg D. Improved cyberbullying detection using gender information. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). University of Ghent, 2012. pp. 23–5.
53. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. In Fifth International AAAI Conference on Weblogs and Social Media. 2011. https://ojs.aaai.org/index.php/ICWSM/article/view/14209.
54. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web. 2016. pp. 145–53. https://doi.org/10.1145/2872427.2883062.
55. Nandhini BS, Sheeba J. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). pp. 1–5. https://doi.org/10.1145/2743065.2743085.
56. Burnap P, Williams ML. Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data Science. 2016;5(1):11. https://doi.org/10.1140/epjds/s13688-016-0072-6.
57. Greevy E, Smeaton AF. Classifying racist texts using a support vector machine. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004. pp. 468–9. https://doi.org/10.1145/1008992.1009074.
58. Kwok I, Wang Y. Locate the hate: Detecting tweets against blacks. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press, 2013. p. 1621–2.
59. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion. 2017. pp. 759–60. https://doi.org/10.1145/3041021.3054223.
60. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM. 2017. pp. 512–5.
61. Liu S, Forss T. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In KDIR. 2014. pp. 530–7. https://doi.org/10.5220/0005170305300537.
62. Mehdad Y, Tetreault J. Do characters abuse more than words? In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2016. pp. 299–303. https://doi.org/10.18653/v1/W16-3638.
63. Burnap P, Williams ML. Cyber hate speech on twitter: An application of machine classification and statistical modeling for

policy and decision making. Policy Internet. 2015;7(2):223–42. https://doi.org/10.1002/poi3.85.
64. Warner W, Hirschberg J. Detecting hate speech on the world wide web. In Proceedings of the second workshop on language in social media. Association for Computational Linguistics, 2012. pp. 19–26.
65. Agarwal S, Sureka A. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr microblogging website. arXiv preprint 2017. arXiv:1701.04931.
66. Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth International AAAI Conference on Weblogs and Social Media, 2014.
67. Del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). 2017 pp. 86–95.
68. Gitari ND, Zuping Z, Damien H, Long J. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering. 2015;10(4):215–30. https://doi.org/10.14257/ijmue.2015.10.4.21.
69. Thelwall M. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In Cyberemotions. Springer, 2017. pp. 119–34. https://doi.org/10.1007/978-3-319-43639-5_7.
70. Djuric N, Zhou J, Morris R, Grbovic M, Radosavljevic V, Bhamidipati N. Hate speech detection with comment embeddings. In Proceedings of the 24th international conference on world wide web. 2015. pp. 29–30. https://doi.org/10.1145/2740908.2742760.
71. Le Q, Mikolov T. Distributed representations of sentences and documents. In International Conference on Machine Learning. 2014. pp. 1188–96.
72. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;5:135–46. https://doi.org/10.1162/tacl_a_00051.
73. Khatua A, Cambria E, Khatua A. Sounds of silence breakers: exploring sexual violence on twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018. pp. 397–400. https://doi.org/10.1109/ASONAM.2018.8508576.
74. Zhang Z, Luo L. Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web. 2019;10(5):925–45. https://doi.org/10.3233/SW-180338.
75. Mathew B, Dutt R, Goyal P, Mukherjee A. Spread of hate speech in online social media. In Proceedings of the 10th ACM Conference on Web Science. 2019. pp. 173–82. https://doi.org/10.1145/3292522.3326034.
76. Araque O, Iglesias CA. An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity. IEEE Access. 2020;8:17877–91. https://doi.org/10.1109/ACCESS.2020.2967219.
77. Araque O, Gatti L, Kalimeri K. MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. Knowl-Based Syst. 2019;105184:11. https://doi.org/10.1016/j.knosys.2019.105184.
78. Benito D, Araque O, Iglesias CA. 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 Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. pp. 396–403. https://doi.org/10.18653/v1/S19-2070. https://www.aclweb.org/anthology/S19-2070.
79. Baeza-Yates R, Ribeiro-Neto B et al. Modern information retrieval, volume 463. ACM press New York, 1999.
80. Gambhir HK. Dabiq: The strategic messaging of the islamic state. Institute for the Study of War, 15, 2014.
81. Mahzam R. Rumiyah: Jihadist propaganda and information warfare in cyberspace. Counter Terrorist Trends and Analyses. 2017;9(3):8–14. http://www.jstor.org/stable/26351502.
82. Azman NA. Islamic state (is) propaganda: Dabiq and future directions of islamic state. Counter Terrorist Trends and Analyses. 2016;8(10):3–8. https://doi.org/10.1145/3041021.3054223.
83. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, Rosso P, Sanguinetti M. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation. 2019. pp. 54–63. https://doi.org/10.18653/v1/S19-2007.
84. Fersini E, Nozza D, Rosso P. Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian. 2018;12:59. https://doi.org/10.4000/books.aaccademia.4497.
85. Demšar J. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7(Jan):1–30, 2006.