Computers, Materials & Continua
DOI:10.32604/cmc.2020.012770
Article

Tech Science Press

Understanding the Language of ISIS: An Empirical Approach to Detect Radical Content on Twitter Using Machine Learning

Zia Ul Rehman[1,2], Sagheer Abbas[1], Muhammad Adnan Khan[3,*], Ghulam Mustafa[2], Hira Fayyaz[4], Muhammad Hanif[1,2] and Muhammad Anwar Saeed[5]

[1]School of Computer Science, National College of Business Administration & Economics, Lahore, 54000, Pakistan
[2]Department of Computer Sciences, Bahria University, Lahore, 54000, Pakistan
[3]Department of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan
[4]School of Systems and Technology, University of Management and Technology, Lahore, 54000, Pakistan
[5]Department of CS & IT, Virtual University of Pakistan, Lahore, 54000, Pakistan
[*]Corresponding Author: Muhammad Adnan Khan. Email: madnankhan@lgu.edu.pk
Received: 12 July 2020; Accepted: 10 August 2020

Abstract: The internet, particularly online social networking platforms have revolutionized the way extremist groups are influencing and radicalizing individuals. Recent research reveals that the process initiates by exposing vast audiences to extremist content and then migrating potential victims to confined platforms for intensive radicalization. Consequently, social networks have evolved as a persuasive tool for extremism aiding as recruitment platform and psychological warfare. Thus, recognizing potential radical text or material is vital to restrict the circulation of the extremist chronicle. The aim of this research work is to identify radical text in social media. Our contributions are as follows: (i) A new dataset to be employed in radicalization detection; (ii) In depth analysis of new and previous datasets so that the variation in extremist group narrative could be identified; (iii) An approach to train classifier employing religious features along with radical features to detect radicalization; (iv) Observing the use of violent and bad words in radical, neutral and random groups by employing violent, terrorism and bad words dictionaries. Our research results clearly indicate that incorporating religious text in model training improves the accuracy, precision, recall, and F1-score of the classifiers. Secondly a variation in extremist narrative has been observed implying that usage of new dataset can have substantial effect on classifier performance. In addition to this, violence and bad words are creating a differentiating factor between radical and random users but for neutral (anti-ISIS) group it needs further investigation.

Keywords: Radicalization; extremism; machine learning; natural language processing; twitter; text mining

1 Introduction

Internet and social media networks may play a dynamic role in the violent radicalization process as it allows convenient spread of misinformation and propaganda. It also facilitates the identification and

engagement of selected audience who seem to be interested in radical messages. Online radicalization introduces an individual to ideology that encourages the drive to shift from mainstream belief to extremism. Precisely defining radicalism, it specifies embedding oneself into others opinion, belief, and knowledge to persuade the behavior of other individual [1]. It has been identified that radical individual control the minds of young generation by initiating radicalization process through social media [2]. Online radicalization of social network users is growing pain and concern in current digital age as it escalates negative activities. With the increase of terror attacks and extremism in religious context radicalization is receiving serious attention and research work [3]. Recent research proves that 78% terrorists are using internet for various activities and social media as a facilitative tool for online violent radicalization and for other extremist activities [4]. Secondly we also have some examples of people who ended up being radicalized through online media as they were actively involved with like-minded people ending up believing on their ideology [1]. Anders Breivik [5], an anti-Islam rebel was a normal social individual but somehow carried out Norway Bombing. Secondly, Zachary Chesser [6] a young enthusiastic student accepted Islam and speedily entered process of radicalization. He posted extremist views, comment, and videos and ended up joining Al Shabab.

This research work primarily concentrates on the online radicalization process of Islamic State of Iraq and Syria (ISIS). Facts depict that ISIS has a deep-rooted online radicalization network which is highly persuasive and operative [7,8]. One such illustration is the rise of ISIS propaganda on Twitter by 4% in the month of October 2019, which was the month of Baghdadi's death (leader of ISIS). On average 214 new ISIS accounts were activated in the month of September 2019 and October 2019 as reported by Ctrl-Sec (https://twitter.com/CtrlSec). These all statistics demonstrate that ISIS online network is working at its best to induce their rationale into common user and death of their leader has only contributed towards more publicity to ISIS agenda and thoughts. Previous research reveals that killing is ineffective counter terrorism strategy against separatist and religious organizations [9]. This has been proven as ISIS has announced their new leader just after a week of Baghdadi death [10]. All these sightings and statistics demonstrate that ISIS has intensive embedded wide network of militants who are aimed at radicalizing potential online users and we have to formulate a strategy where we can hinder and stop this process as just by killing their leaders would not make a difference. Secondly it is essential to note that ISIS has been active in all good and bad times of organization itself. Recently many operations are executed against ISIS, but organization is fulfilling its motive. In January 2020, suicide bombing was conducting in Quetta, Pakistan killing 15 people [11]. In same year in March 2020 an attack on Sikh complex was also executed by ISIS [12]. ISIS is also proving threat in recent Covid-19 pandemic, executing attacks in 10 countries including Syria, Iraq, first ever attack in Maldives and Philippine. After Baghdadi's death 25 terrorist attacks are executed by ISIS resulting in 491 fatalities till June 2020 [13]. This all strongly indicate that ISIS has penetrated various corners of world and online radicalization has been one of the most fertile platforms to do so.

Previous research has worked on Radicalization and have determined five Radical Groups. Doosje in their research [14] have discussed Separatist radical group which aims at securing a territory for their own group and religious group which aims at strict interpretation of their religion to justify violence against other religions. ISIS lies in both groups as they intent to control Iraq and Syria as their territory. Secondly ISIS are religiously motivated also as they use religious text to propagate their ideology. Additionally, ISIS followers normally support for violent actions and tend to use violent and bad words. This research work has mainly focused on identifying tweets which relate to these two characteristics (radical and religious). Based on all these findings this research effort investigates these main questions:

- Can incorporation of religious text into computational model enhance the performance of radicalization detection approaches? Previous research affirms that users when became activated as radical users, they started to use more religious terms while communicating on social media [15].

- Are there any changes in linguistic narrative and patterns of Pro-ISIS users? It is an important factor when detecting radicalization as basing on language we train our model. For answering this question, a comprehensive *corpus* analysis is conducted with previous and recent datasets.
- Is there any empirical evidence which proves that including violent and bad words in radicalization detection process is good measure? Secondly, we need to assess if Anti-ISIS users are also using it frequently or not.

The research work is organized as follows: Section 1 introduces the scope and need of research, Section 2 presents an insight of previous work and its short coming, Section 3 demonstrates detailed methodology describing each and every stage of working model, Section 4 illustrates experiments and results, further it also presents *corpus* analysis, and section 5 concludes the research work discussing future perspective as well.

2 Related Work

Researchers are actively working on the online radicalization phenomenon [16] and how extremist organizations are employing it to achieve their motive [17,18]. In this section we are summarizing established computational approaches developed to analyze, detect, and predict the processes of radicalization.

Starting off with analysis of radicalization, in 2015 Klausen [19] has explored the role of Twitter in the jihadist functioning policy in Iraq and Syria. Considering 55 manually identified Twitter accounts, author has found out that the flow of communication is from terrorist account to the fighters situated in rebellious region to the supporters located in west. Author also discovered that 38.54% tweets were related to religious guidance. In 2015, Chatfield [20] has also found strong evidence that ISIS supporters/members utilize multisided Twitter network for terrorism propaganda, communication, recruitment, and radicalization. Vergani [21] has studied the growth and progression of ISIS language by examining the text used in 11 issues of Dabiq which is ISIS' official English magazine. Findings are as follows: Firstly, using expression related to power and achievement, secondly using emotional words and thirdly frequent mentions of females, death, and Islam. Rowe [15] have examined user behavior at different stages of adopting pro-ISIS stance. Utilizing data mining approach author has defined the points of activation, portraying divergent behavior, and finally embracing pro-ISIS behavior. Bröckling [22] in their work assessed the contents of Rumiyah which is the second IS propaganda magazine. Comprehensive analysis of all the issues of Rumiyah reveal that content instigate in readers that Caliphate is longer the objective rather they should start battle from their homelands. Secondly the magazine clearly narrates enemy images thus providing a basis for radicalization of youth. In 2019, Torregrosa [23] has conducted a study to compare the features of the rhetoric utilized by pro-ISIS and general Twitter users. Author investigated that usage of words with six or more letters relate with emotionality in negative sense and is an indicator towards psychological distancing. Previous research [24,25] are also reporting same.

Elaborating the considerable efforts regarding detection of radicalization, we will first talk about the research of Berger et al. [26]. In their work they had formulated a scoring mechanism to point out social media accounts which were most influential or had the tendency of being influenced within an extremist sphere. In [27] Berger focused at identifying ISIS supporting accounts from profile descriptions which comprised of words such as linger, succession, Caliph state, Islamic State, Iraq, and many more with the accuracy of 94%. The limitation of the study is that only 70% accounts had the profile information. Agarwal [28] has employed single class KNN and SVM algorithms for automatic detection of hate/extremism promoting tweets. The research concluded that the presence of war related term, religious term, negative and offensive words clearly indicate a tweet to be hate promoting. The model achieved F-Score 83%. Ashcroft [29] utilizes sentiment, time, and stylo-metric features to automatically detect text distributed from jihadist groups on Twitter. Author concludes that proposed technique can only assist humans in detection and is not capable of replacing completely. Saif [30] tested the efficacy of semantic

features in comparison to sentiment features, unigram features, network features, and topic features for categorizing anti-ISIS and pro-ISIS users on Twitter. Experiments demonstrate that classifiers trained by semantic features perform well by 7.8% on F1-measure when compared with other baseline features. Lara-Cabrera [31] has established a set of Radicalization indicators based on belief, religion, personality from social sciences to evaluate if user has signs towards radicalization. In an extension to their work [32] author has mapped a set of radicalization indicators by social sciences (discrimination, introversion, frustration) to set of computational features (keywords) which could be extracted automatically from data. Results show this mapping is effective but has a limitation as it based on keywords only. In 2018 Smedt [33] developed a system that automatically detects online jihadist hate speech with F-score of 80%. The tweets were collected from October 2014 to December 2016. The research effort of Nouh [34] demonstrate that radical users (focused group was ISIS) particularly from Twitter show some textual, behavioral, and psychological characteristics. Secondly psychological are the most distinctive ones. Utilizing textual features author achieved F-Score of 80% with word2vec.

In context of prediction, Ferrara [35] proposes a technique to predict if a Twitter user will embrace extremist content and reciprocates it. The results divulge that number of hashtags promoted, proportion of retweeting to tweeting and total number of tweets are all determining factors. In 2018, Fernandez [36] predicts and detects the radicalization influence a specific user has based on roots of radicalization extracted from social science models [37]. In 2019, Waqas [38] proposed an approach to identify extreme behavior of Twitter users (focused extremist group was Taliban). TF-IDF features were extracted from n-gram terms in the tweets and best classifier (SVM) achieved average F-score of 84%.

In this section we have conducted an exhaustive survey on types of models developed for analyzing, detecting, and predicting radicalization. Some of the significant findings are as follows:

- Most of the approaches are categorizing users based on few pieces of user content (new comments, recent posts), very few works consider complete history of user for instance their complete timelines.
- Most of the approaches are considering ISIS as only separatist group (aims at securing a territory) hence are extracting features based on radical language only.
- According to our knowledge very few approaches are focusing on violent and bad words. Some research efforts indicate that they are contributing towards radicalization process [28,34].
- Most approaches are not considering recent ISIS datasets which includes tweets from 2019.
- Some of the studies [14,15,19] particularly which are analyzing radicalization process have clearly hinted towards the presence of religious terms in radical content.

We are aiming at providing a step ahead in relevance to the previous research efforts. We intend to introduce an approach which is considering ISIS as both Religious and Separatist group, hence incorporating religious text in addition to radical language for feature extraction. In contrast to previous studies our technique considers complete timeline of Twitter accounts to detect radicalization stance. In addition, we are analyzing violent and bad words usage pattern for radical and neutral users. It will allow us to illuminate the effect of violent and bad words in the process of radicalization as it is observed that all radical groups share common characteristics of showing violent and bad languages. Finally, we are employing our own new dataset which is collected from September 2019 to November 2019 so that results are based on the recent activity of ISIS.

3 Methodology

This research effort aims at developing an efficient framework for detecting radicalization content in potential Twitter accounts. The proposed framework outlines some major steps as illustrated in Fig. 1. The major steps are data collection, data pre-processing, feature engineering, model generation, and evaluation of results.

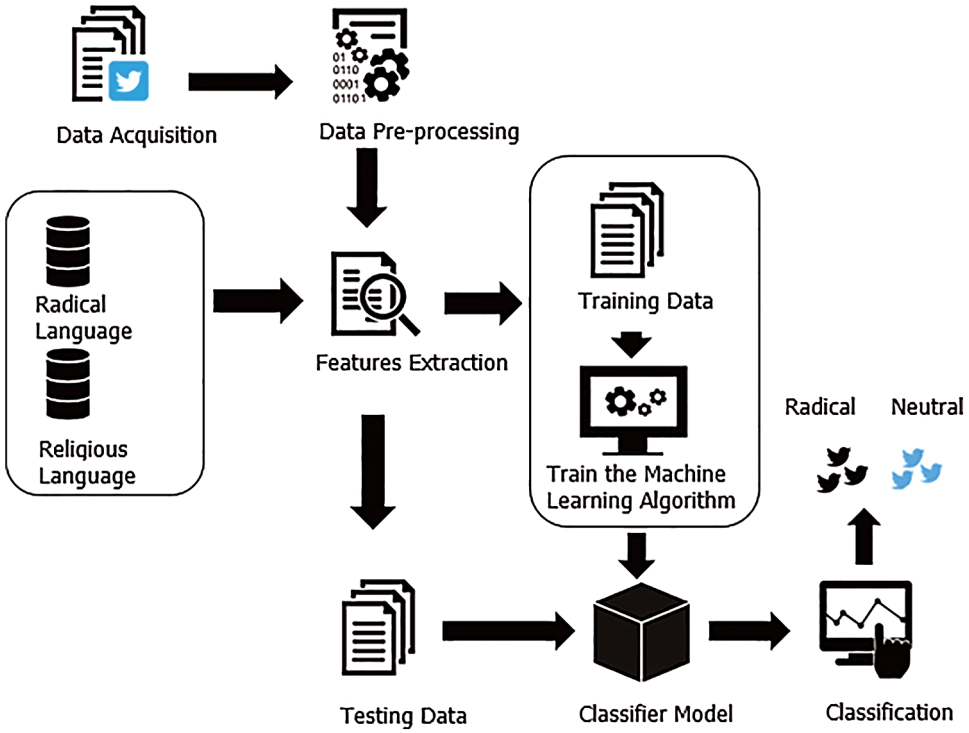

Figure 1: Proposed approach where radical features extracted from pro-ISIS users' tweets, violent/bad words, and Arabic terms from dabiq and inspire magazines. Religious features extracted from religious text used in dabiq and rumiyah magazines

3.1 Data Collection

Starting from step 1 which is data acquisition we have utilized five distinct data sets in this study. In the following we will elaborate each dataset in detail.

3.1.1 Radical Corpus (D1)

This data set is named as D1 and represents Radical *Corpus*. It consists of 17,350 Twitter user tweets. It is collected from 112 user accounts claimed to be pro-ISIS. It ranges on the period of three months. Users are identified basing on keywords such as Amaq, Dawla and Wilyat. Next, users are filtered basing on their usage of images (images of radical leadership, ISIS flags, Anwar Awlaki). It is publicly available on Kaggle data science community (https://www.kaggle.com/fifthtribe/how-isis-uses-twitter).

3.1.2 Neutral Corpus (D2)

This dataset is named as D2 and represents Neutral *Corpus*. It has been extracted as counterpoise of dataset D1. It consists of 122k tweets all collected from 95,725 anti-ISIS users on two days 7-11-2016 and 7-4-2016. Tweet are gathered basing on keywords such as isil, isis, Islamicstate, Mosul, raqqa, 'islamic state'. It is publicly available on Kaggle (https://www.kaggle.com/activegalaxy/isis-related-tweets).

3.1.3 Religious Corpus (D3)

This dataset is named as D3 and represents Religious *Corpus*. This dataset is vital in research as religious text plays a fundamental role in promoting and publicizing ISIS ideology and recruitment. It contains all the ideological and religious texts (Jewish, Muslim, Christian) utilized in ISIS English magazines. It analyzes 15 issues of Dabiq (June 2014–July 2016) and 9 issues of Rumiyah (September 2016–May 2017)

generating 2,685 texts. It is publicly available on Kaggle data science community (https://www.kaggle.com/fifthtribe/isis-religious-texts). There are total 8 features in the dataset named as Magazine (Dabiq, Rumiyah), Issue #, Date, Type (Quran, Hadith, Jihadist, Islamist, 20 others), Source (Sahi Muslim, Bukhari, 1155 others), Quote (the actual text), Purpose, Article Name. We filter the dataset on basis of type selecting Jihadist, Islamist, and others as sources. We have excluded Quran, Hadith, and Bible as it does not represent extremist content. After filtration 936 quotes are used for the training purpose in our experiments. Most of these are the quotes from head of extremist groups for example Abu Mus'ab az-Zarqawi, Abu Bakr al-Baghdadi, and Osama Bin Laden.

3.1.4 New Dataset (D4)

This is newly collected dataset for this research work. It comprises of both radical and neutral tweets. To generate this dataset, we have extracted relevant data from Twitter employing Twitter Streaming API. It consists of tweets from user previously recognized as Pro-ISIS and suspended by Twitter. Ctrl-Sec is responsible for reporting these accounts. It is a non-profit organization which voluntarily reports the ISIS propaganda on Twitter. Earlier research [36] asserts that Ctrl-Sec is an efficient resource in generating dataset which assist in identifying radical users. Ctrl-Sec claims to be responsible for suspension of more than 350,000 reported Twitter accounts since it started working 2015. These reported accounts are immediately suspended hindering the collection of data. Consequently, researchers are unable to use most relevant tweets for further advance research endeavors. For D4 dataset we crawled the tweets (complete timeline) of latest reported accounts by Ctrl-Sec before they are suspended by the Twitter. The 9000 Tweets are collected by 13 reported accounts and their followers. The dataset is collected between September 2019 and November 2019. The ISIS propaganda increased in September 2019 after the releasing of new recording of their leader Abu Bakr al-Baghdadi in which he urges followers to continue attacks. We also manually review the tweets from each account to make sure it is promoting ISIS ideology or are involved to support their propaganda. While crawling the tweets we also found the English translation of the latest speech of the Abu Bakr al-Baghdadi shared by one of the reported accounts. The speech text is also incorporated in D4. These 13 accounts are suspended and no longer available. This dataset also comprises tweets from neutral users. For neutral class we collected tweets from 28 October 2019 to 30 October 2019 after Baghdadi's death. We used the hashtags #Al-baghdadi, #baghdadikilled #abubakaralbaghdadi #ISIS to gather these tweets. These tweets are also analyzed manually to make sure it does not contain pro-ISIS contents.

3.1.5 Random Dataset (D5)

We have also collected tweets by random users discussing general topics. The dataset is collected by using Twitter API. The tweets are collected from 15 October 2019 to 20 October 2019. We manually observe the tweets to make sure it does not contain pro-ISIS or anti-ISIS contents. There are total 7000 tweets that are discussing current affairs, sports, and general topics. We have used this dataset in experiments to investigate our third research question which is to examine the usage of violence and bad words by radical users.

3.2 Data Pre-Processing

Moving towards Step 2 which is data pre-processing, we clean the acquired data (D1, D2, D3, D4, and D5). The extracted set of tweets are in raw state and contains undesired data and noise. This reduces the performance of classifiers. Consequently, we have performed a series of pre-processing tasks to clean data and make it appropriate for feature extraction.

The steps are as follows; (i) **Tokenization** transform each tweet into word segments, it is necessary as we have to remove undesired term and construct word vector, (ii) It is very likely that token are repetitive with uppercase and lowercase. Character **transformation** aims at converting each single token into a specific

format to control repetitive features, so here we are converting all tokens to lowercase, (iii) **Irregular terms** such as tagged URL links or images with combined text decreases the performance of model, hence all such term are also removed from tweets, (iv) In this step we **remove all non-English** tweets. The approach is to tokenize each tweet into a list of words and compare each word against an English dictionary then calculate a score between 0 and 1 measuring the amount of known English words in the tweet. We used Microsoft dictionary and remove the tweet if it contains 50% of non-English words. We kept the threshold value at 50% because we do not want to remove the tweet containing some non-English words such as Kafir, Muwahhid and Taghut because these are the radical words and will help us to develop radical dictionary at feature engineering stage, (v) **Stop words** increases the dimensionality of features and are of no use. All the stop words (the, is, an, am, e.g., to, how and, punctuation marks) are removed from the dataset, (vi) **Stemming** is performed to eradicate various inflected form of words to a common base from, storing the result in word vector.

3.3 Feature Engineering

In the next step, feature engineering was performed. We have considered two categories to determine appropriate features to detect radical text. In the following we will explain these two categories of textual features.

3.3.1 Radical Language

In this category we had to comprehend the construction and usage of radical terms. For this we extracted textual features from the D1 dataset which contains the tweets from radical users. For understanding the construction of radical messages, we have extracted radical language used by the pro-ISIS users. We have also incorporated existing dictionaries with the aim of providing a wider set of terms representing radicalization terminology. The integrated dictionaries are:

- The Top 50 terms used by Pro-ISIS users as mentioned in [30,31,39]. Some of these terms are Caliphate, Crusaders, Apostate, assad, war, enemy, isis, Islamic state, Syria.
- Arabic terms glossary as referred in [40]. These all terms are extracted from ISIS official magazine Dabiq (issue July 2014–July 2016) and Al-Qaida magazine Inspire (issue June 2010–May 2016). Some of the terms are not in standard format with respect to their usage and spelling so these terms are considered as jihadist slang. This glossary includes 300 distinct terms, but we have only used the ones which clearly speak for ISIS propaganda. Some sample terms are: al-wala wal bara, Alawite, dar al-kufr, fajir, Cihad (Turkish form of Jihad), Madkhalis, manhaj, murtad, Muwahhid, salaf, shirk, takfir, takhmis, wilayah, zindiq, zani, Hijrah, Bay'ah, Murtaddin, Mushrikin, Taghut.
- In addition, we also incorporate vocabularies of violence, terrorism related words [41], and curse/bad words [42]. We combine all these vocabularies. After removing duplicate words there are 2375 unique words in total. These vocabularies are also used in previous studies [34] for radicalization detection.

3.3.2 Religious Text

In this category we had to understand the construction and usage of religious language. As mentioned earlier ISIS falls into the category of religious radical group and uses religious text to promote their ideology and for propaganda contents. For this we extracted textual features from D3 dataset which contains the religious text used in ISIS propaganda magazines Dabiq and Rumiyah.

We have used TF-IDF technique of extraction of radical terms from the radical *corpus* (D1) and the religious terms from the religious *corpus* (D3). In Term Frequency-Inverse Document Frequency (TF-IDF) the values of each words within these datasets are calculated. Unigrams and bigrams are applied to identify the terms and the perspective in which these phrases are being used. Top-scoring

grams are then selected to be employed as features for the language modelling. Previous literature proves that N-gram have been efficient in detecting extremist content and hate speech [38].

3.4 Classification Model

Moving ahead next step is to train machine learning algorithm. Choosing an efficient classifier is an imperative task to construct a progressive predictive capability. In this study we have employed naïve Bayes, Random Forest, and Support Vector Machines. Previous studies [34,38] investigating similar problems proves that these classifiers are performed better. The efficacy of these algorithms is discussed in the coming sections which elaborates experiments and results.

3.5 Evaluation Metrics

The last step is to evaluate our results. The standard evaluation metrics are used to measure the effectiveness of the classifiers. These are precision, recall, accuracy, and F-score.

4 Experimental Setup

In our experimental setup we have taken following steps; (i) Dataset collection and generation, (ii) Exploratory Data Analysis, (iii) Applying classification algorithms on developed classification model and, (iv) Evaluating results. We have already described datasets D1, D2, D3, D4, and D5 in Section 3. An elaborated insight about the purpose of each dataset is also given. It all accumulates Step 1, data collection and generation. In the next sub-section, we will demonstrate exploratory data analysis.

4.1 Corpus Analysis

Corpus analysis is a textual analysis which facilitates comparison between textual object at a larger scale. It is an effective approach for in-depth demographic analysis. It enhances the understanding of dataset as it represents data with improved visualization techniques. In this sub-section we will present exploration of radical tweets from *corpus* D4 (testing *corpus*) and D3 (Religious *Corpus*). D1 (Radical *Corpus*) is used in previous studies [23,31,34,36] and it is already established that it contains pro-ISIS tweets. To validate that our newly collected dataset (D4) is also aligned with this we performed this investigation. We investigated the new dataset (D4) with two different tools and techniques to explore varying patterns and properties of the datasets. We have conducted (i) Linguistic analysis by employing LIWC, (ii) Scatter-text. Following sections explain them in detail.

4.1.1 Linguistic Analysis

Linguistic analysis is a study which aims at analyzing the distinct pattern in language employed by a group. In our research effort we have employed the Linguistic Inquiry Word Count (LIWC) software to observe and understand terminologies contained in our dataset D4. As elaborated earlier D4 is one of the pivotal datasets which effects results immensely. It consists tweets from pro-ISIS accounts which are reported by ctrl-sec. It is a novel dataset, so we have explored the characteristics it possesses.

Previously many research attempts have been made which deeply analyzed the language patterns of jihadi groups. Some substantial findings are; i) Jihadi groups use more words related to anger, emotionality, religion and death [43], ii) Increased usage of words related to females so that women get attracted to their motive [21], iii) Very less usage of words depicting friendship while significant use of words with negative tone, certainty and power [44]. It is also stated that pro-ISIS users use third person plural pronouns and words with six or more letters while interacting as compared to singular pronoun or other words [24]. Torregrosa in his research effort [23] has conducted a comprehensive linguistic analysis of pro-ISIS users' dataset which in our case is D1. Some common characteristics are defined for the

extremist contents. Following we assess our dataset D4 based on these characteristics of extremist data. We will discuss each property in relevance to D4.

- ***Supportive comments towards jihadi/extremist groups and their ideologies***. In D4 the tweets support the extremist groups such as ISIS and Taliban and express sentiments by using words like Mujahedeen, Muwahideen, Fighters, and Taliban.
- ***Usage of extremist terminologies to criticize the groups and people for instance "Kafir" and "Kuffar."*** D4 do possess the attribute of criticizing other groups by using violence and aggression related words such as "Kill them all," Taghut, Pigs, and Kafir.
- ***Expression of support and dedication towards jihadi leaders (for example, Al-Baghdadi).*** In D4 "Ameer ul momnin" and "Shykh" are used to show respect and support towards leaders.
- ***Intensive sharing of interviews, links to interviews of jihadi figure and of individuals who show support towards ISIS.*** In case of D4, Twitter removed the video links containing extremist contents, but we found statements, quotes, and speeches of ISIS leaders. We also found the English translation of the Al-Baghdadi's speech of September 2019, which is the last speech before his death. This testify that intensive sharing has been done.
- ***Strong condemnation of countries which are involved in US-led coalition in Iraq and Syria***. In our dataset (D4) strong opposition to US coalition by using extremist terminologies is observed. Presence of US-led coalition in Afghanistan is also criticized in addition to Syria and Iraq.

Above we have manually observed the characteristics of our dataset D4. The characteristics demonstrated by D4 are similar with the characteristics of D1. This proves that D4 is promoting radical propaganda especially the motive of ISIS. This analysis will support us in our experiments as we aim to train and test our model on D4. Secondly, this dataset can be used in further research. Now we will assess the psychological patterns of the tweets shared by the radical users in D4. For the purpose we employ LIWC default dictionaries which will measure the occurrence of word from each dictionary. We compare our results with [23] and found our results are in-line with them. Tab. 1 is showing the average words count for each category. From Tab. 1 it is evident that emotional tone is relatively high as compared to previous analysis. First person plural and third person plural values are also higher in our case. As expected, they use more words expressing anger, certainty, power, religion, and death. These results are in-line with previous analysis [23], the only contradiction witnessed is for the results of emotional tone and first person singular.

4.1.2 Scatter-Text Analysis

Scatter-text has the competency of comparing the frequency of two terms and label them in scatterplot. In our study we have employed scatter-text; an open source tool to visualize linguistic dissimilarity between document groups. It presents a scatterplot where each axis corresponds to the frequency of a term occurring in document groups. Through state of art technology this tool can exhibit thousands of terms represented in point form and finds space to legibly label at least hundreds of them. We have used scatter-text for data visualization and identify differences in word usage between radical and neutral tweets. Fig. 2 demonstrate the graph generated by scatter-text for D4.

In the graph, x-axis represents radical group with red color and y-axis represents neutral group with blue color. The total documents analyzed are as follows; (1) Radical Documents Count: 6,922 and Word Count: 59,215, (2) Neutral Document Count: 1,898 and Word Count: 17,749. Point placement and intelligent word labelling makes analysis extremely competent and resourceful. As the graph illustrates some of the top frequent terms used by radical users are *Taliban, allah, muslim, afghanistan, islam, war and ibn*. In the same manner top frequent terms used by neutral users are *raid, baghdadi, trump, terrorist, dog*, and *isis* leader. Characteristics shows the unordered list of the frequent used words by both radical and neutral.

One noteworthy observation is that even non-English terms and abbreviation which are frequently used are highlighted in this analysis, for instance *khawaarij, kuffar, kufr, ummah* and *dunya*. The scatterplot is highly interactive, as you click on any specific term its TF-IDF score will be displayed. For instance, considering the top frequent term *terrorist* we can see the frequency and TF-IDF score as "173.13 per 25k words, score: 0.59990." It is quite useful information regarding hidden pattern of our new dataset D4.

Table 1: Average no. of words used by radical users of *corpus* D4 in comparison to users of *corpus* D1

Category	[23] D1	Ours D4
Emotional tone	13.35	21.67
Words with six letters or more	24.00	**22.07**
First person singular pronouns	0.47	1.49
First person plural pronouns	0.52	1.02
Second person pronouns	0.52	2.02
Third person singular pronouns	0.43	1.06
Third person plural pronouns	0.70	1.46
Positive emotions	1.53	3.01
Negative emotions	2.55	**3.28**
Anger	1.71	**1.61**
Certainty	0.58	**1.72**
Power	2.97	**3.78**
Religion	2.05	**2.79**
Death	0.72	**0.79**

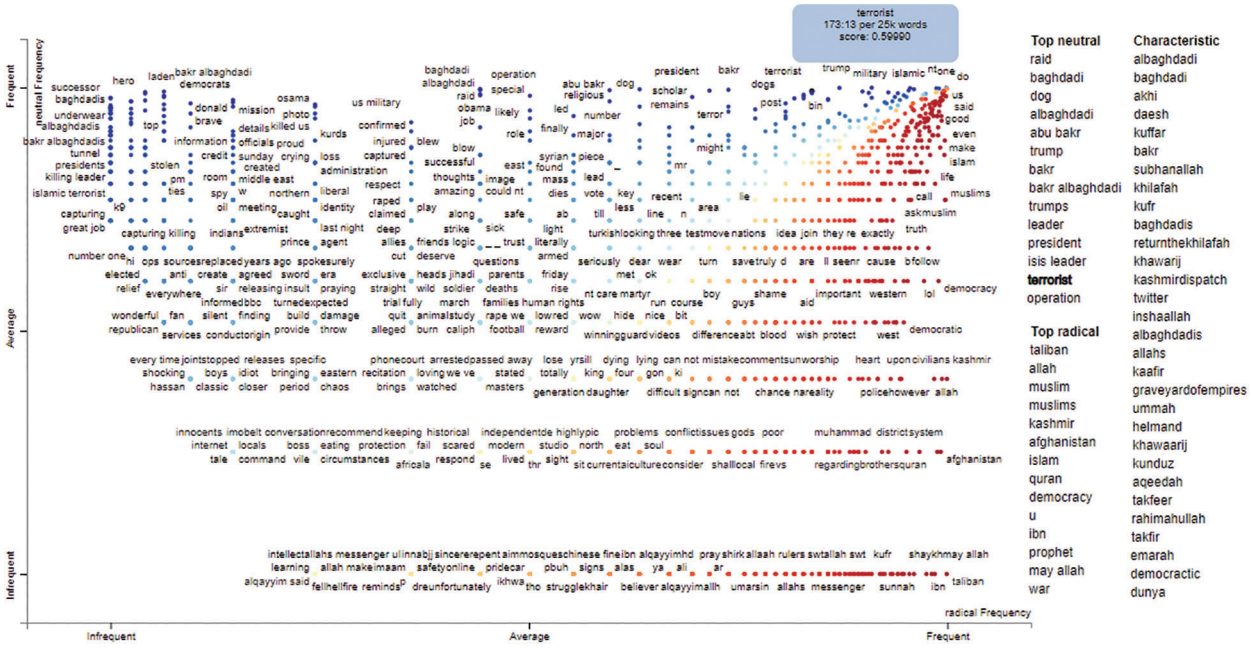

Figure 2: X-axis shows the frequency of words used by radical users and y-axis shows the frequency of words used by neutral users. On the right, top 10 neutral and top 10 radical, and an unordered list of words (characteristic) used by both users are listed

4.1.3 Results and Findings of Corpus Analysis

This *corpus* analysis is quite fruitful as it has rendered some noteworthy observations. We have figured out some similarities and differences in previous and new dataset. According to our analysis, both datasets possess following similarities.

- Negative emotions occur with high frequency which means the ratio of negative words is higher in the tweets.
- Power, Certainty, Anger, Religion and Death related words are sighted with higher value. These observations are in-line with the previous studies [38].
- Religious terms such as Allah, Muslim, People, and Islam occur frequently in both datasets.
- Violence and bad words vocabulary such as Killed, Attack, Kafir, and Enemy are also giving higher values.

Coming towards the contradicting properties we had following notes:

- Both datasets are showing different political terms. D1 is showing Syria, Iraq, Israel, Aleppo, and Ramadi with high frequency. On the other hand, D4 is frequently showing the terms related to Afghanistan, Kashmir, and Taliban. This shows the change in political narrative of ISIS.
- Terms related to Women are occurring more frequently in D4 as compared with previous dataset D1. This represent the ISIS agenda to motivate women to promote their ideology.

To sum up, it is reasonable to say that our new dataset is representing the current narrative of ISIS which has changed from 2015 to 2020. According to our understanding the proposed dataset D4 if included in future research will deliver promising results.

4.2 Experiments

For the evaluation of our research objectives, we have setup an experiment. Our first research objective states that supporters of radical group particularly ISIS may exhibit similar textual properties as presented by religious *corpus* D3 (extracted from propaganda magazines) when communicating on social media. We have put an effort to identify if religious text contributes positively towards detecting radical content.

4.2.1 Experiment 1

The first experiment is divided into two cycles. In the first cycle we trained the classifiers by employing radical features only as described in Section 3. We used our new dataset D4 for model training and testing by applying 10-fold cross-validation. A specific tweet is labeled as radical if it shows radical language, propagates violence, or encourages violence. In the second cycle we again trained our classifier and this time by selecting both radical and religious features so a clear comparison could be established. We compared the results from both experiments and found out that the performance of model improves by employing religious features in classifier training. The results can be seen in Tab. 2. To get better predictive competency, we have applied various learning algorithm such as naïve Bayes, SVM (linearSVC) and Random Forest (n-estimators = 100 with gini impurity for split). This experiment is performed on the complete set of features mined that comprises of TF-IDF feature set. We used scikitlearn's TfidfVectorizer.

F-score, accuracy, precision and recall with each classifier are recorded. With all three classifiers the values of accuracy, precision, recall, and F-score increases as we incorporate religious features in classifier training. This is evident from results that religious terms are vital in detecting radicalization content. This is clearly in-line with our research objective.

Table 2: Classifiers scores of evaluation metrics

Textual Features	Classifiers	Accuracy	Precision	Recall	F-Score
Radical Features only	NB	0.77	0.78	0.77	0.77
	SVM	0.81	0.82	0.81	0.81
	RF	0.79	0.80	0.79	0.79
Combining Radical and Religious Features	NB	0.81	0.81	0.81	**0.81**
	SVM	0.87	0.88	0.87	**0.87**
	RF	0.84	0.84	0.85	**0.84**

4.2.2 Experiment 2

Our second experiment intends to observe the usage of violent and bad words by extremist group supporters. We also need to look if it can aid in detection of radical content. For this we utilized LIWC to acquire the average word count (bad and violent words) from radical and neutral users' tweets. Dictionaries of violence, terrorism, and bad words are employed. These dictionaries are described in Section 3. First, we compare radical user's tweets from D4 with random tweets dataset D5. The results demonstrate that radical users on average use more words related to violence as compared with general Twitter users. The values for both users can be visualize in Fig. 3.

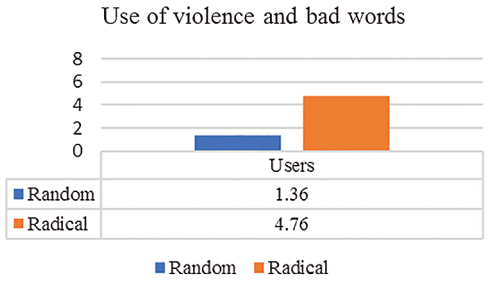

Figure 3: Average No. of violence/bad words used by radical (pro-ISIS) users and random (normal) users

After that we compare the radical and neutral users of our recent proposed dataset (D4) and already available datasets (D1, D2). The results demonstrate clearly that both classes radical and neutral are using violent and bad words while communicating on social media. Fig. 4 presents results in detail. For D4, the word count is higher for neutral class and same goes for D2. D1 is radical *corpus* but it is showing a lesser value. At this stage further research is required to measure the extent by which the usage of violent and bad words of radical group differ from neutral group.

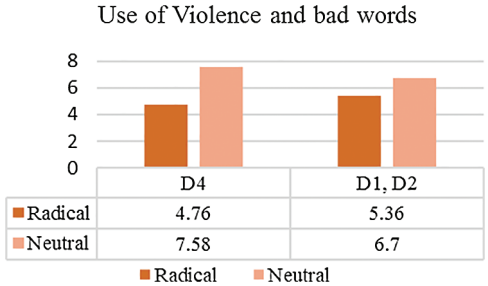

Figure 4: Average No. of violence/bad words used by radical (pro-ISIS) users and neutral (anti-ISIS) users

4.3 Analysis of Results and Discussion

In this section, we are deeply analyzing the results we achieved from our experiments. We also compare our results with previous related studies so that a clear comparison of proposed methodology could be established. Previous work used the lexicon-based techniques [36] and machine learning [28,33,34,38] to identify the radical content on Twitter. In our research we used machine learning to classify the tweets as radical or neutral. Our approach is different from [33,34,38] in terms of extracting features. We combine radical and religious terms in feature extraction. Experiment 1 studies the role of religious terms. The results we have achieved are better as compared with previous related studies that investigates radical or extremist contents on social media. We have presented a short summary of previous substantial research efforts and our research attempt in Tab. 3. We have included the focus group of study, extracted textual features and the achieved F-score with proposed methodology.

Table 3: F-score in comparison to previous related studies

Reference	Focused Group	Extracted Textual Features	F-Score
Ours	ISIS	Radical terms, religious terms, violent and bad words	0.87
2019 [38]	Taliban	Sentiment based pro-Taliban/pro-Afghan features	0.84
2015 [28]	Jihadist	War terms, Bad/offensive words, internet slangs, negative emotions	0.83
2019 [34]	ISIS	Radical terms, violent and bad words	0.80
2018 [33]	Jihadist	Jihadist terms	0.79

Earlier, we have put up three research questions that we intended to investigate (Section 1). Research Question 1 focused on the religious language and how it can be employed for radicalization detection. For this purpose, an empirical approach which incorporates religious features for the detection is proposed. Experiments illustrate that integration of religious features yields higher value for F-score, Precision, Accuracy, and Recall. Consequently, it is reasonable to merge religious features along with other features for radicalization detection. In reference to Research Question 2 we anticipated to observe the narrative of radical (pro-ISIS) users over the time. We sought to establish a comparison of linguistic patterns for previous and new datasets by using different state of the art tools and techniques. The objective is to observe the difference or change in use of common terms/expressions, emotions and behavioral or psychological patterns. As we found some common terms or expressions in our new dataset that are also appeared in previous dataset, but we also observed some changes in narrative. The religious terms (Allah, Muslim, People, Prophet, Islam) used by pro-ISIS users previously are still used frequently in new dataset. Some political terms that are used frequently before such as Syria, Iraq, Aleppo, and Ramadi are not used frequently in our latest dataset. In contrast the words of Afghanistan, Kashmir and Taliban are used frequently in recent data. This is consistent with language of Rumiyah. The content of Rumiyah encourage readers that caliphate is no longer the objective rather they should start battle from their homelands. The recent attacks of ISIS in Afghanistan, Philippine and Maldives seem to relate with the content of Rumiyah magazine. In relevance to research Question 2, it is justifiable to say that new dataset (D4) can increase the accuracy of model as it represents the current narrative of ISIS.

Finally, coming to Research Question 3 which raises concerns about violent and bad words. We explored the use of bad and violent words by radical and neutral group. We employed LIWC to analyze the tweet text of radical and random users. Our results show that radical users are using more words related to violence and bad words as compared to random users. We also compared the tweets text of radical and neutral users (tweets that are showing Anti-ISIS sentiments). Our results show that violent and bad words are

frequently used by neutral classes as well when communicating on social media. This question needs further intensive research to see if bad and violent words could help in detection process.

5 Conclusion and Future Perspectives

This research attempt proposes an effective approach employing machine learning which aims at detection of radical content on social media. Its major focus is to take advantage of religious language used by extremist group particularly on Twitter as some evidence supports that religious language also contributes positively towards radicalization detection. Religious features along with radical features are employed. Various machine learning algorithms are trained to evaluate the results. Secondly, we have also conducted an in-depth analysis of datasets. For this purpose, *corpus* analysis was performed to understand the difference in language narrative of ISIS over time. Thirdly, violence and bad words are also considered. Our experiments produced positive results supporting our research objective, but we believe in future enhancements. First, we like to add psychological and contextual dimension to the proposed model and validate if its efficacy has increased over multiple datasets. Furthermore, most of the accounts reported by Ctrl-Sec are from Afghanistan and they tweet in Pashto language. We intend to expand this research with other languages particularly Pashto language. Collecting empirical data in terrorism is a grueling process, but in future we should aim at experimenting with larger samples.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] International Association of Chiefs of Police. *Online Radicalization to Violent Extremism Awareness Brief,* 2014. [Online]. Available: https://www.theiacp.org/sites/default/files/2018-07/RadicalizationtoViolentExtremismAwarenessBrief.pdf.

[2] M. Conway, M. Khawaja, S. Lakhani, J. Reffin, A. Robertson *et al.,* "Disrupting daesh: measuring takedown of online terrorist material and its impacts," *Studies in Conflict & Terrorism*, vol. 42, no. 1–2, pp. 141–160, 2018.

[3] R. Scrivens and G. Davies, "Identifying radical content online," *voxpol*, 2018. [Online]. Available: https://www.voxpol.eu/identifying-radical-content-online/.

[4] P. Gill, E. Corner, M. Conway, A. Thornton, M. Bloom *et al.,* "Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes," *Criminology & Public Policy*, vol. 16, no. 1, pp. 99–117, 2017.

[5] H. Pidd, "Anders behring breivik spent years training and plotting for massacre," *The Guardian*, 2012. [Online]. Available: https://www.theguardian.com/world/2012/aug/24/anders-behring-breivik-profile-oslo.

[6] Z. Chesser, "A case study in online islamist radicalization and its meaning for the threat of homegrown terrorism," Washington, DC: Senate Committee on Homeland Security and Governmental Affairs, 2012. [Online]. Available: www.hsgac.senate.gov/imo/media/doc/CHESSER FINAL REPORT%281%29.pdf.

[7] B. I. Koerner, "Why ISIS is winning the social media war?," *Wired*, vol. 24, no. 4, pp. 76–83, 2016.

[8] W. Magdy, K. Darwish and I. Weber, "Failedrevolutions: Using twitter to study the antecedents of ISIS support," 2015. https://arxiv.org/pdf/1503.02401v1.pdf.

[9] J. Jordan, *Leadership Decapitation: Strategic Targeting of Terrorist Organizations*. Stanford University Press, Redwood City, California, 2019.

[10] R. Callimachi and E. Schmitt, "ISIS names new leader and confirms al-baghdadi's death," *The New York Times*, 2019. [Online]. Available: https://www.nytimes.com/2019/10/31/world/middleeast/isis-al-baghdadi-dead.html.

[11] G. Yousafzai, "Death toll in pakistan mosque suicide bombing rises to 15," *Reuters*, 2020. [Online]. Available: https://www.reuters.com/article/us-pakistan-blast/death-toll-in-pakistan-mosque-suicide-bombing-rises-to-15-idUSKBN1ZA0GM.

[12] C. Dwyer, "At least 25 people dead after hours-long attack on sikh complex in kabul," *NPR.org*, 2020. [Online]. Available: https://www.npr.org/2020/03/25/821428292/at-least-25-people-dead-after-hours-long-attack-on-sikh-complex-in-kabul.

[13] ISIS, "Terrorist attacks: a map of terrorist attacks, according to Wikipedia," *A Story Maps Labs Project*. [Online]. 2019. Available: https://storymaps.esri.com/stories/terrorist-attacks/.

[14] B. Doosje, F. M. Moghaddam, A. W. Kruglanski, A. De Wolf, L. Mann *et al.,* "Terrorism, radicalization and de-radicalization," *Current Opinion in Psychology*, vol. 11, no. 1, pp. 79–84, 2016.

[15] M. Rowe and H. Saif, "Mining pro-isis radicalisation signals from social media users," in *10th Int. AAAI Conf. on Web and Social Media*, Germany, 2016.

[16] O. Araque and C. A. Iglesias, "An approach for radicalization detection based on emotion signals and semantic similarity," *IEEE Access*, vol. 8, no. 1, pp. 17877–17891, 2020.

[17] M. Howell, *Fighting Extremism: Efforts to Defeat Online ISIS Recruitment Methods*. Oxford, UK: Honors Theses, 2017.

[18] M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on twitter," in *Semantic Web for Social Good Workshop (SW4SG) at Int. Semantic Web Conf.*, CEUR, 2018.

[19] J. Klausen, "Tweeting the Jihad: Social media networks of western foreign fighters in Syria and Iraq," *Studies in Conflict & Terrorism*, vol. 38, no. 1, pp. 1–22, 2015.

[20] A. T. Chatfield, C. G. Reddick and U. Brajawidagda, "Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks," in *Proc. of the 16th Annual Int. Conf. on Digital Government Research*, pp. 239–249, 2015.

[21] 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, no. 1, pp. 7–20, 2015.

[22] M. Bröckling, C. Fritsch, M. Haider and T. Yalman, "Kill them wherever you find them'-radicalizing narratives of the 'so-called' islamic state via the online magazine rumiyah," *Journal for Deradicalization*, vol. 17, pp. 240–294, 2018.

[23] J. Torregrosa, J. Thorburn, R. Lara-Cabrera, D. Camacho and H. M. Trujillo, "Linguistic analysis of pro-isis users on twitter," *Behavioral Science of Terrorism and Political Aggression*, vol. 12, no. 3, pp. 1–15, 2019.

[24] L. Kaati, A. Shrestha and K. Cohen, "Linguistic analysis of lone offender manifestos," in *IEEE Int. Conf. on Cybercrime and Computer Forensic*, pp. 1–8, 2016.

[25] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, 2011.

[26] J. M. Berger and B. Strathearn, "Who matters online: measuring influence, evaluating content and countering violent exremism in online social networks," *Int. Centre for the Study of Radicalisation and Political Violence*, vol. 41, pp. 1–41, 2013.

[27] J. M. Berger and J. Morgan, *The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter*. Washington DC USA: Brookings Institution Center for Middle East Policy, 2015.

[28] S. Agarwal and A. Sureka, "Using KNN and SVM based one-class classifier for detecting online radicalization on twitter," in *Int. Conf. on Distributed Computing and Internet Technology*, Bhubaneswar, pp. 431–442, 2015.

[29] M. Ashcroft, A. Fisher, L. Kaati, E. Omer and N. Prucha, "Detecting jihadist messages on twitter," in *European Intelligence and Security Informatics Conf.*, Manchester, UK, pp. 161–164, 2015.

[30] H. Saif, T. Dickinson, L. Kastler, M. Fernandez and H. Alani, "A semantic graph-based approach for radicalisation detection on social media," *European Semantic Web Conf.*, vol. 10249, pp. 571–587, 2017.

[31] R. Lara-Cabrera, A. Gonzalez Pardo, K. Benouaret, N. Faci, D. Benslimane *et al.,* "Measuring the radicalisation risk in social networks," *IEEE Access*, vol. 5, no. 1, pp. 10892–10900, 2017.

[32] R. Lara-Cabrera, A. Gonzalez-Pardo and D. Camacho, "Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in twitter," *Future Generation Computer Systems*, vol. 93, no. 1, pp. 971–978, 2019.

[33] T. De Smedt, G. De Pauw and P. Van Ostaeyen. "Automatic detection of online Jihadist hate speech," *CLiPS Technical Report Series*, Computational Linguistics & Psycholinguistics Research Center, 2018.

[34] M. Nouh, R. C. J. Nurse and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on twitter," in *IEEE Int. Conf. on Intelligence and Security Informatics*, Shenzhen, China, pp. 98–103, 2019.

[35] E. Ferrara, W. Q. Wang, O. Varol, A. Flammini and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Int. Conf. on Social Informatics*, UK, pp. 22–39, 2016.

[36] M. Fernandez, M. Asif and H. Alani, "Understanding the roots of radicalisation on twitter," in *10th ACM Conf. on Web Science*, Netherlands, pp. 1–10, 2018.

[37] A. Maskaliūnaitė, "Exploring the theories of radicalization," *International Journal of Politics, Culture, and Society*, vol. 17, no. 1, pp. 9–26, 2015.

[38] W. Sharif, "An empirical approach for extreme behavior identification through tweets using machine learning," *Applied Sciences*, vol. 9, no. 18, pp. 3723–3742, 2019.

[39] W. Afzal, "A study of the informational properties of the ISIS's content," in *Proc. of the American Society for Information Science and Technology*, 53, no. 1, pp. 1–6, 2016.

[40] R. J. Bunker and P. L. Bunker, "Radical Islamist English-language online magazines: Research guide, strategic insights, and policy response," *Army War College Carlisle Barracks Pa Carlisle Barracks*, USA, 2018.

[41] Myvocabularycom, "Violence vocabulary." 2020. [Online]. Available: https://myvocabulary.com/word-list/violence-vocabulary/.

[42] L. V. Ahn, "Bad words vocabulary," Carnegie Mellon University. 2016. [Online]. Available: https://www.cs.cmu.edu/~biglou/resources/bad-words.txt.

[43] K. Cohen, F. Johansson, L. Kaati and J. C. Mork, "Detecting linguistic markers for radical violence in social media," *Terrorism and Political Violence*, vol. 26, no. 1, pp. 246–256, 2013.

[44] C. M. Udani, "A content analysis of jihadist magazines: theoretical perspectives," Ph.D. dissertation, University of Central Florida, USA, 2018.