IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 24, 2021, accepted June 29, 2021, date of publication July 5, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3094532

Adaptable Reduced-Complexity Approach Based on State Vector Machine for Identification of Criminal Activists on Social Media

IMRAN SHAFI[1], SADIA DIN[2], ZAHID HUSSAIN[1], IMRAN ASHRAF[2], AND GYU SANG CHOI[2]

[1]Department of Electrical and Mechanical Engineering, National University of Science and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38544, South Korea

Corresponding authors: Imran Ashraf (ashrafimran@live.com) and Gyu Sang Choi (castchoi@ynu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) by the Ministry of Education under Grant NRF-2021R1A6A1A03039493, and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program by the Institute for Information and Communications Technology Promotion (IITP) under Grant IITP-2021-2016-0-00313.

ABSTRACT Security agencies face an emerging challenge of identifying and counter the malicious contents spread on the social media by the terrorists. However, text classification techniques are limited by visualization, pre-processing, features extraction, and larger features space. Additionally, change in criminal content require the learning models to identify altered malicious textual contents which poses extra challenge. This study proposes simplified yet adaptable framework that uses a novel features extraction algorithm for extracting features from the textual part of social media contents. The feature extraction considers selective features from only 8 dimensions and follows a six step process. The extracted features are suitably used to train the state vector machine for the classification of the malicious content. The performance of the proposed method is evaluated against other popular feature selection/ extraction algorithms like term frequency-inverse document frequency, Gini Index (GI), Chi square statistics, and PCA. Additionally, machine learning classifiers like decision tree, random forest, and Naïve Bayes are also used for classification. Results suggest that the proposed approach consumes less energy on text visualization, pre-processing, and dimensionality reduction. It also reduces the time-space complexity of the features extraction process and is capable to steer according to the changing strategies of the active criminal groups. In addition, it can effectively analyze the propaganda material published by the extremists. It automatically identifies the radical text on social media platforms allowing understanding of the behaviors, characteristics and subsequent blockage of such content.

INDEX TERMS Criminal intent prediction, feature dimensionality reduction, machine learning, feature extraction.

I. INTRODUCTION

The presence of criminals and terrorists in the online world is not new, but the current trend of social media enhanced their interest in the online community. As the access to the mainstream media for such elements is expensive and difficult, so in this case, social media has become their favorite option. They use social media to propagate propaganda, recruitment campaigns, and twisting real-life events in their favor for provoking people for violence and countering their opponents. Several cases have been reported in the near past from different parts of the world in which innocent people are provoked and recruited for criminal deeds by different terrorist groups through their malicious contents present on social media.

Studies show that such criminal groups are found across the globe and have the capability to create unrest in society through their malicious campaign on social media. They use the social media in a wise and effective manner for achieving a range of objectives. They provoke the youngsters, defame the states and states' organizations, and justify through advertising their criminal activities. Many practical social hate campaigns have been observed recently in countries like US, UK, France, Egypt, Syria and Pakistan.

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

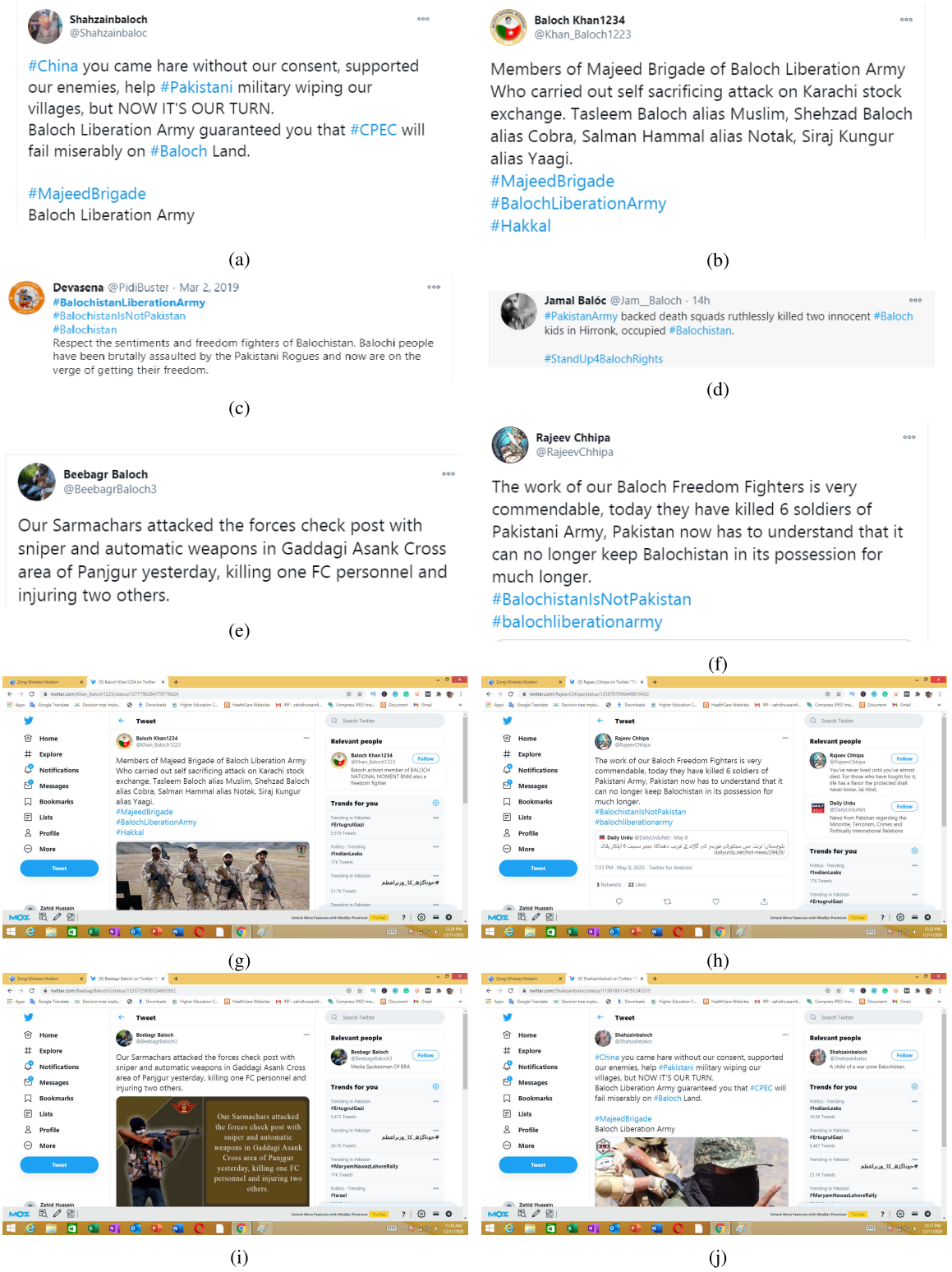

FIGURE 1. Social media contents of terrorist group active in Pakistan.

Similarly, people have been provoked by the criminal activists via social media against states to satisfy a specific agenda like blasphemy cases. Certain examples have been presented in Figs.1 and 2 where certain Twitter ids (now blocked by Twitter) are shown displaying malicious intent. Some ids are trying to provoke the locals against a development project, while some are justifying and advertising the terrorist attack on a stock exchange and few others are

FIGURE 2. Social media contents of terrorist activists active in Europe and the Middle East.

convincing to attack the other religions. Fig. 2 highlights the process of recruitment of young and innocents through social media in Europe/Middle East, use of threatening language to attack Twitter employee for suspension terrorists' account and provocation of the followers to attached US embassies.

The presence of criminals and terrorists in the online world has also drawn significant attention from multiple research disciplines, from computer sciences to social sciences, and to political sciences, especially in the recent inter-disciplinary fields of computational social science, computational socio-economics and computational socio-economics [1], [2].Also, other researchers from disciplines of psychology, social science, criminology, mental health, and political science are also actively involved in this domain. Kim and other fellow colleagues have predicted the mental health of twitter user [3]. Burnap *et al.* [4] measure the social disturbance from social media data. Jibril, [5], have used social media analytics for intelligent information gathering to combat terrorism. Williams and others identify the racial and religiously aggravated crime from anti-black and anti-social media posts with the help of machine learning techniques [6].

The dataset generated by such criminal activists on social media platforms consists of multiple categories covering recruitment, finances, communication, provocation, defamation, dis-information and propaganda. Resultantly, the identification/classification of different-natured textual malicious content comprising of various subcategories via text classification techniques is an extremely challenging task. It's important to highlight that the social media contents are both textual and visual. This research work focuses on the textual content of social media. The identification of malicious content within a text is a classification task, for which sentiment analysis and text classification are the favorite solutions. But multiple challenges are faced during sentiment analysis and text classification tasks. Sentiment polarity is used during sentiment analysis to classify the text and researchers have highlighted certain issues for the correct analysis. The most important requirement for accurate text categorization is the availability of multilingual sentiment lexicon [7]. To assign suitable weights to the ambiguous words is another challenge as certain words possess dual meanings which cause the ambiguity [8]. Also, the sentiment polarity is domain-dependent, e.g. a sentiment may be positive in one domain and negative in another domain. To accurately predict the sentiment polarity in a particular domain, sentiment analysis needs separate sentiment lexicon for each domain [9]. Moreover, sarcasm detection and negation handling [10] make sentiment analysis further challenging.

On the other hand, text classification requires fitting the text into a machine learning algorithm in the form of numbers. This text passes through several processes which increases the complexity of the text classification [11]. These processes include preprocessing of text, text visualization, features selection, and features extraction. During the text pre-processing, removal of stop words, stemming, lemmatization, and conversion of text into lowercase are carried out. Text pre-processing is expected to significantly improve the accuracy and reduce dimensionality of the features space [12]. Text visualization is the conversion of text into numbers; for this technique like a word cloud, bag of words, and term frequency are used. Text visualization converts the text into a high-dimensional feature space.

Dimensionality reduction is another challenging task because reducing dimensions generally lowers the accuracy. Moreover, textual data has data sparseness and semantic gap which requires semantic knowledge to be integrated with the textual data processing. Representing short text in features vector loses contextual information leading to synonymy and polysemy. To overcome, key indicators or meaningful concepts are identified in the text [13]. To reduce dimensions with minimum loss of accuracy, the feature set is reduced through feature selection and features extraction. Techniques like term frequency-inverse document frequency (TF-IDF), information gain, Gini index, and Chi-square statistics are used for feature selection. However, the above-mentioned techniques only count the occurrence of features in a document. For part of speech and features that are semantically important, techniques like Word2Vec are employed [14]. Once the features are selected, the feature extraction techniques are used to reduce the number of features to lower the complexity and improving computational efficiency of the classifiers in learning the right parameters. The features extraction techniques are powerful in reducing the dimensionality of feature space, the most common include principal component analysis (PCA) and latent semantic indexing (LSI).

The dynamic nature of the problem is also a big hurdle, i.e. the objectives and interests of the criminal groups may change with time causing changes in their design, conduct and style of campaign. They may also twist the real-life events for their interest. This results into multiple dictionaries of words for the same criminal group. The machine learning algorithms are generally unable to adapt and adjust to the dynamically changing situations dropping the accuracy of their results. To retain accuracy, the algorithms need to be retrained regularly, which is time-consuming.

In this paper, a novel framework based on a combination of feature extraction and classification algorithms is introduced for the classification of text. We propose a novel feature extraction algorithm (FEA) to extract suitable features used to train SVM for classification purposes. The FEA does not need lexicon definition, sarcasm detection, and negation handling and avoids text pre-processing. The proposed approach reduces the high dimensionality of features and complexity of text classification. It also avoids retraining through altering the contents of input arrays according to the changing interests and objectives of the criminals. The experiments are conducted on the data collected from the official social media accounts of criminal and terrorist groups, active in the Middle East, Africa, Afghanistan, and Pakistan. The data is labeled, as malicious and non-malicious, with the help of the Federal Investigation Agency's (FIA) cybercrime experts. The performance of the proposed method is evaluated against both the feature extraction approaches and some other popular machine learning algorithms like Decision Tree (DT), Random Forest (RF), and Naïve Bayes (NB) used for classification. The objectives of this research paper are to find answers to the following questions:

- How accurately the proposed method classifies the text?
- How much the FEA reduced the dimensionality of the features set?
- How much the FEA reduced the complexity of text classification?
- How successfully the proposed method adjust itself to the dynamic nature of the problem?

The rest of the paper is structured as follows. Section II discusses the research papers related to the current study. The proposed model and its working mechanism are described in Section III. Section IV contains the results and discussion and in the end, the conclusion is drawn in Section V.

II. RELATED WORK

In the last decade, text classification remained a hot topic for research. It is used for different purposes like sentiment analysis, brand analysis, opinion mining, and classification of news articles as sports, politics, and economics. In the same way, texts produced by social media are analyzed for emotion detection, sentiment analysis, investigation, and identification of malicious contents of criminals and child abusers.

The challenges related to the domains of sentiment analysis and text classification are widely studied in the literature [14], [15]. Shah et al create a multilingual sentiment lexicon with intensity weights to classify social media contents as high extreme, low extreme, moderate, and neutral. The accuracy up to 82% is achieved with machine learning classifiers using this approach [7]. Shedge and fellow workers present a review of special features extraction techniques for handling ambiguity of words [8]. To address the issue of contextual polarity i.e. a sentiment being positive in one domain and negative in another domain, Richard and co-authors adopt hierarchical classification in text mining. They use three filtering schemes to progressively classify text with respect to the contextual polarity and frequent term of documents [9]. Zhou Yao et al adopt a filter method of stop word list in text pre-processing for text classification. Three different filtering algorithms are designed for creating a customized stop word list based on the difference of text document's domain. The results show significant improvement in dimensionality reduction of feature space and accuracy of text classifiers [10].

There are some other interesting approaches found in the literature to solve above mentioned issues, like [13] uses hash tag as an indicator to identify meaningful concept in tweets. While authors in [12], suggest a text pre-processing framework for text mining on big data infrastructure to reduce the computation time. Vibha and colleagues present an overview of various features selection techniques like Gini index, information gain, Chi square, term frequency and inverse document frequency for text classification. The authors also discuss PCA and latent semantic index as suitable techniques for the dimensionality reduction. However, these methods do not consider the part of speech and semantic importance of a feature and only count the occurrence of a feature in a document. To resolve this issue, Tian and fellow researchers propose a much-improved vector representation model, word2vec, to consider the semantic meaning [14].

In another work [4], tweets are analyzed to detect spikes in social tension triggered by a racial abusing accident that occurred in a football match. In this work, Pete have proved that using a combination of conversation analysis and syntactic and lexicon-based text mining rules can beat several sentiment analysis techniques and machine learning approaches. While in [16] investigate a chat log for evidence using a social graph-based text mining framework. They find key terms, key users, and key sessions in a group chat. The authors in [3] present an approach to identify depressive users on Twitter. They use learning-based text analysis, a word-based emoticon analysis, and SVM based image classifier to extract mood from the text, emoticon, and images, respectively, and finally, aggregate the extracted mood. [17], [18] [19] studied rumor propagation and suggested multiple rumor and anti-rumor models.

Similarly, Govin et al use the AFINN lexicon database for assigning impact factor to every word and machine learning algorithms SVM, KNN, and NB for classification of tweets. Vikas and co-workers [20], suggest two hypotheses for detecting cyber-aggressive comments on social media

along with supervised machine learning. Study [21] accurately predict the author of a text message from the style, habit, and other peculiarities of writers with the help of supervised machine learning. Similarly [22] use external lexicons and TF-IDF features extraction for finding a correlation between Twitter sentiment and events that have occurred. [23] identifies malicious users on social media who propagate false opinions and distort the general perception, using KNN, Naïve Bayes, and decision tree machine learning algorithms. They also determine terror awareness level on social media using a scalable framework. The authors in [24] worked on the cross-domain sentiment analysis and prove that a multinomial Naïve Bayes classifier trained on the dataset of one domain (tweets) could classify the sentiment in others domain (reviews) successfully.

People also share their emotions on social media about different events that occur around them. So, this opens a new challenge of emotion detection from the text. For this, [25] presents a two-stage emotion detection mechanism. In the first stage emotion bearing tweets are collected, and in the second stage, tweets are classified according to emotion, i.e., love, anger, hate, etc. Similarly, [26], [27] detects emotions from the text and classifies the content according to the emotions detected.

In another study [5], open-source social media analytics are recommended such as Gephi, igraph, NetworkX, and some commercially available social media analysis tools such as i2 analyst, and sentinel visualizer. These tools help to improve intelligence gathering for the security agencies for analyses of the activities of the terrorist's organization on social media and to track them down. To trace the geospatial location of the activists and track them properly is a new methodology adopted by the researchers, where NLP and fuzzy logic is applied on the public posts and locations of the social media users for finding suspects [28].

An interesting application in the related area is discussed in [29] where forensics assessment of social media platforms is performed after a critical incident by closely monitoring the social reaction. To deal with this critical issue, machine-learning techniques, particularly natural language processing and latent dirichlet allocation (LDA) topic modeling is used to create an unsupervised text reduction method that is used to study social reactions in the aftermath of the 2017 Manchester Arena bombing. The database is comprised of millions of messages posted on Twitter within the first 24 hours after the attack. This study resulted in tracking different types of social reactions over time and identifying sub-events that have a significant impact on public perceptions [29].

In addition, the digital practices and discourses of the Islamic State when exploiting social media communication environments to propagate their Jihad ideology and mobilize specific audiences have been examined by a group of authors [30]. To counter and examine the Islamic State of Iraq and Syria (ISIS) recruitment behavior through social media, [31] suggest text mining methods be deployed, that categorize text document into specific categories and help to make tracing network more efficient. Both the machine-based text classification and ontology-based text classification are compared for the best results.

Apart from terrorism, the criminal activists go well beyond the racist and Islamophobic content being shared to target a community or group of society. Matthew and fellow researchers [6] discuss the use of computational criminology that draws on data science methods, to link police crime, census, and Twitter data to establish a temporal and spatial association between online hate speech that targets race and religion, and offline racially and religiously aggravated crimes. Similarly, another group of authors present an automated tool to detect and classify Islamophobic hate speech robustly and at scale, by a quantitative analysis of large textual datasets collected from social media platforms [32]. This research draws towards an in-depth conceptual work to build a programmable software tool that differentiates and classifies between non-Islamophobic, weak Islamophobic, and strong Islamophobic content.

A related approach based on ant colony optimization is presented by the researchers for threatening account detection on Twitter based on social honeypot database [33]. The strong connection between the different Twitter users is determined by the pheromone substance of high quality secreted by ants on the edges of the path traveled. The characterization, classification, and meta-analysis for detecting radical and extremist groups on social media platforms are provided in a research work [34]. This study unfolds the statistics on the data source, features, geo-location, language, machine learning techniques, and tools that have been applied to detect cyber-extremist activists.

To understand the radical mindset and behaviors characteristics on social media platforms Mariam and colleagues have analyzed propaganda material published by extremist groups and created a contextual text-based model of radical content. They present a model of psychological properties inferred from these material and evaluated these models on Twitter to automatically identify online radical tweets. The results exhibit the presence of distinguishable textual, psychological, and behavioral properties in the radical users. The most distinguishing features come out to be psychological properties. The results indicate that the textual models using vector embedding features significantly improve the detection over TF-IDF features [35].

Study [15] highlights the displacement effects that result from the automated removal and blocking of terrorist content and suggest that regard must be implemented to the whole social-media ecology covering Jihadi groups other than the so-called Islamic State and other forms of violent extremist organizations. Since rule by law is only a necessary and not a sufficient, condition for compliance with rule-of-law values, the study examines two further sets of issues including the clarity with which social media companies define terrorist content and the adequacy of the processes by which a user

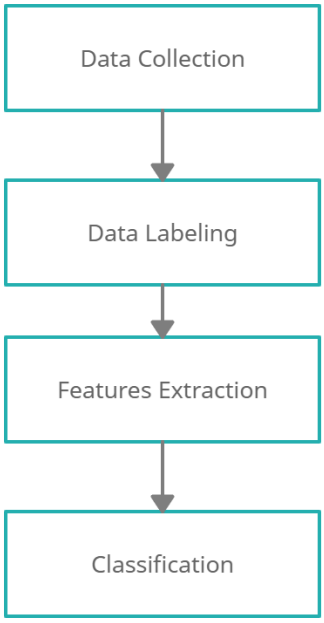

FIGURE 3. The flow chart of the proposed methodology.

may appeal against an account suspension or the blocking or removal of content.

In a study by authors, a new set of methodologies is presented that is designed to allow for efficient data mining and information fusion from social media and of the new applications and frameworks that are currently appearing under the umbrella of the social networks, social media and big data paradigms [36]. A new intelligent forensics approach that incorporates the advantages of artificial intelligence and machine learning theory to automatically flag unseen criminal media is investigated by [37]. The research is extensively focused on the law enforcement cybercrime specialists from different countries and law enforcing agencies. The approach has been implemented into the iCOP toolkit, a software package that is designed to perform live forensic analysis on a peer-to-peer network environment. The system offers multiple secondary features.

III. MATERIALS AND METHODS

This section explains the methodology covering data collection, features extraction, and classification, as shown in the following figure.

A. DATA COLLECTION

Two different datasets, i.e., dataset A and dataset B, are collected from the famous social media site Facebook. Dataset A is collected from the social media accounts of criminal activists of terrorist groups active in the Middle East, Africa, and Baluchistan province of Pakistan. About 8,043 posts are collected from their Facebook accounts using Ncapture and Nvivo software. Out of 8043, about 4000 posts are labeled as negative, and the rest are labeled as positive with the help of Federal Investigation Agency (FIA) experts. About 70% of dataset A was used for training, and 30% was used for testing the machine learning model.

TABLE 1. Description of dataset A and Bused for experiments.

Dataset	Type of post	Number of posts
A	Positive	4,043
	Negative	4,000
	Total	8,043
B	Positive	1,000
	Negative	1,000
	Total	2,000

While in dataset B, about 2000 posts are collected from Facebook. Out of which 1000 are from the criminal activists' campaign, in which they support another terrorist group active in Afghanistan, and Khyber Pakhtoon Khwa province of Pakistan, which are labeled as negative while the rest 1000 posts are from the other non-criminal users' accounts which are labeled as positive. This labeling is carried out with the help of experts at FIA.

Along with this data is collected for the following lists, which are later used in the feature extraction process, which will be updated according to the expected change in the interest of the terrorist groups.

1) **Hashtags list:**
   As nowadays, the hashtag is an integral part of every social media site, which is used to deliver a message on a specific topic, draw the attention of users, and get things viral on social media. So, in the same way, these criminal activists use hashtags related to real-life events, which they can twist for their interest or make a trend of their message on social media that will support their cause. In this list, expected hashtags are collected that could be used by them.
2) **URL list:**
   Criminal activists use social media in an organized way; not everyone creates their malicious content for social media. They adopt a centralized content generation approach. A team of them creates malicious and provoking content and place it on specific websites or a YouTube channel, or a particular social media account. In contrast, the rest of the team members drag the social media users to such content by sharing the URL where it is placed. In this URL list, known URLs are collected where they placed their malicious contents.
3) **Membership Category 1:**
   MC1 is the list of names of all those entities for which they want to create sympathy and support in the community.
4) **Category of Attributes 1:**
   CA1 is the list of all those attributes they use for the members of MC1.
5) **Membership Category 2:**
   MC2 is the list of names of all those entities about which they want to spread hate and blame them for decreasing their support in the community.

6) **Category of Attributes 2:**
CA2 is the list of all those attributes by which they want to target the members of MC2.

B. DATA LABELING

The data have been labeled with the help of an expert of the cybercrime unit of the Federal Investigation Agency of Pakistan. The collected Posts are labeled as either positive or negative.

1) NEGATIVE POSTS

A negative post means a post that contains hate, blame, and propaganda materials against the state or state department. Or contain sympathy, advertising, or provoking material of criminal groups. The posts that contain some contents that are supporting the agenda and narrative of the targeted criminal groups are labeled negative. And assigned -1 as a class attribute.

Example - Video of burning Bugti Baloch homes in Dera Bugti Similarly, the Pakistan Army is burning fire and houses in Balochistan. And Balochistan FC is bombarding the people.

2) POSITIVE POSTS

And the posts other than negative are labeled as positive. And assigned 1 as a class attribute.

Example - Our vision is building a peaceful and prosperous society through social media based on the active participation of youth emphasizing equity, equality, and & social justice.

C. FEATURES EXTRACTION

Text classification is a computationally expensive task. The features extraction algorithm, abbreviated as FEA, is the newly proposed algorithm for extracting features from textual posts. The proposed FEA extracts a feature vector of eight Boolean values per textual content in six steps. The algorithm does not need pre-processing of data and reduces the time and space complexity of the features extraction process. The proposed FEA reduces the computational overhead of text classification task by avoiding the pre-processing of text being not needed. It also restricts a two-stage features extraction process to a single stage by extracting a low dimensional feature set from the text, which does not need any statistical method for dimensionality reduction. Besides this, the extracted features are Boolean instead of decimal.

Terrorist groups across the world are using social media in an organized and strategic way. The FEA captures their whole strategy in a feature vector of eight dimensions. The criminal activists drag the users from the social media sites to their official websites, online video channels, and official social media accounts and groups by sharing its URL on social media. Also, these criminals do not generate malicious content individually. Some of them post malicious content on a central social media account, while their followers share the URL of these contents from their own accounts. The proposed FEA monitors and extracts the information about the presence of such URLs in the social media textual contents. The second tactic of these criminal activists is the use of hashtags on social media. They use special hashtags to help the audience find their message easily on social media. They make the trend of these hashtags by frequently using them in their contents. The proposed FEA has also been configured to extract information about the presence of such hashtags in the textual contents. During their social media campaign, these criminal activists try to create positive sentiment and sympathy, in the online community, for their colleagues and justify their criminal deeds. In order to track such activities, the FEA works to extract the features related to such a campaign from the textual contents. They also target and blame their opponents on social media and try to defame and spread misinformation about the state and state organizations fighting against them. Finally, the newly proposed algorithm FEA computes information from textual social media content about such a campaign.

If n is the total number of textual posts and L is the average number of words per post and b is the average length of lists of URLs list, hashtag list, $MC1$, $MC2$, $CA1$, and $CA2$. Then the computational overhead of each step of FEA can be given as below.

- The combined computational overhead of step 1 and step 2 is $2(nL + nb)$
- Wile that of setp 3-6 is $4(nLb)$

So, the net time complexity function of FEA can be calculated as below

$$T_@(n, L, b) = 4nLb + 3nL + 2nb \tag{1}$$

Considering the growth of the above function and representing it in big O notation, the lower order terms and constants will be ignored as given below.

$$T_2(n, L, b) = O(nLb) \tag{2}$$

As b is the average length of lists used in FEA and the current experiment, b is equal to 98. These lists are seldomly updated during the feature extraction process, so by assuming b as constant, the growth of the T_2 function can be reduced to the following.

$$T_2(n, L, b) = O(nL) \tag{3}$$

When L is relatively small (as usual, it is), then the above function T_2 will behave as a linear function, which will be its best case, while in the worse situation when L will be nearly equal to n, the function T_2 will behave as quadratic.

Comparing both functions T_1 and T_2 show that FEA has enormously reduced the computational overhead of features extraction because T_1 is cubic in its growth while T_2 is quadratic in the worst case (when L is assumed to be equal to n) and linear in the best case (when L is assumed constant).

Similarly, considering the space complexity of both feature extraction methods (TF-IDF + PCA and FEA), it is evident that TF-IDF+PCA requires more memory than FEA.

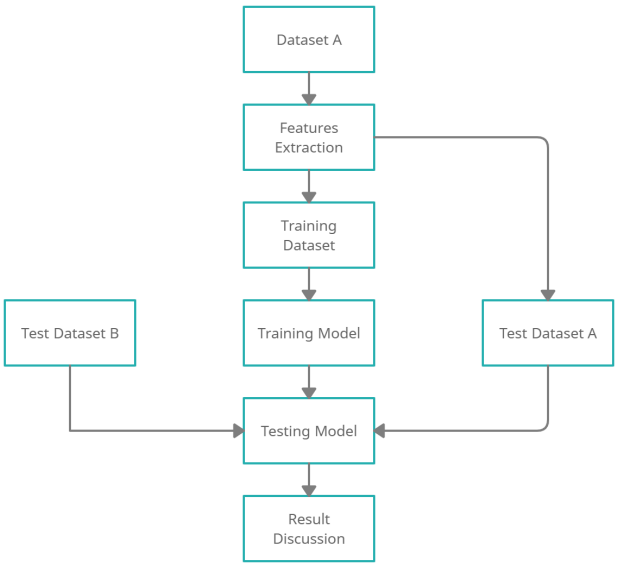

FIGURE 4. Flow chart of the experiment methodology.

If n is the number of documents, L is the average number of words per documents and p is the number of words in a word dictionary, then TF-IDF will need $O(n * p)$ memory for holding each TF, IDF, and TF-IDF values in memory and $O(nL)$ memory for preprocessing. While an additional memory of $O(np + p2)$ will be required for the execution of PCA. As compared to this, FEA extracts eight one-bit features per features vector, which will need only $O(n * 8)$ memory to hold the features set in the main memory.

FEA takes the textual post,p and the hashtags list, URLs list, MC1, MC2, CA1, and CA2 as input and returns a vector of features containing 8 Boolean features. FEA aims to reduce the complexity of feature extraction in the text classification, which the FEA has achieved successfully. The detailed algorithmic presentation of FEA is given below Algorithm 1.

D. CLASSIFICATION

The four famous machine learning classifiers, the Decision Tree, Random Forest, Support Vector Machine, and Multinomial NB of Naïve Bayes family, are used for classification and comparison.

IV. RESULTS AND DISCUSSIONS

A. EXPERIMENT SETUP

In this work, two parallel experiments are performed. In both the experiments, 70% of dataset A is used for training algorithms, and dual tests are performed, one with 30% of dataset A and the other with 100% dataset B. The structure of the experiment is given in Fig. 4.

In the first set of experiments, features are extracted from the text using the FEA. Then the mentioned classifiers DT, RF, SVM, and NB are trained and tested on the extracted features according to the said structure. In the second set of experiments, a novel method of text classification is implemented. Data is pre-processed, then features are selected from the data using various approaches like TF-IDF,Information Gain (IG), Gini Index (GI), and Chi Square statistics. After

Algorithm 1 Feature Extraction Algorithm

1. Check the presence of URL in p and store as url
if url>1 **then**
  $ft[1] = 1$
  if url is subset of URLLIST **then**
    $ft[2] = 1$
  else
    $ft[2] = 0$
  end if
else
  $ft[1] = 0$ and $ft[2] = 0$
end if
2. Check the presence of hash tag in p and store as htl
if htl>1 **then**
  $ft[3] = 1$
  if htl is subset of TAGLIST **then**
    $ft[4] = 1$
  else
    $ft[4] = 0$
  end if
else
  $ft[3] = 0$ and $ft[4] = 0$
end if
3. Check the occurrence of instances in p of membership category1 $mc1$ from $MC1$
if mc1>1 **then**
  $ft[5] = 1$
else
  $ft[5] = 0$
end if
4. Check the occurrence of instances in p of category attributes $ca1$ from $CA1$
if ca1>1 **then**
  $ft[6] = 1$
else
  $ft[6] = 0$
end if
5. Check the occurrence of instances in p of membership category2 $mc2$ from $MC2$
if mc2>1 **then**
  $ft[7] = 1$
else
  $ft[7] = 0$
end if
6. Check the occurrence of instances in p of category attributes $ca2$ from $CA2$
if ca2>1 **then**
  $ft[8] = 1$
else
  $ft[8] = 0$
end if
7. Return feature set $ft[]$

feature selection, the dimensionality of the feature set is further reduced by using the Principal Component Analysis (PCA). Finally, the machine learning classifiers DT, RF, SVM, and NB are trained and tested on the reduced feature set according to the said structure.

B. RESULTS USING TF-IDF+PCA VS FEA

Finding information of interest in large textual data is a difficult task. Different methods are used for the extraction of such information from text such as TF-IDF, GI, IG, and Chi Square Statistics. The combination of TF-IDF and PCA is found to be one of the most efficient text feature selection and extraction methods. However, its implementation is a lengthy process. Textual data passes through four different phases during features extraction via TF-IDF+PCA. In the first phase text is pre-processed that includes removal of stop words, lower casing, stemming, and lemmatization. In the second phase text is visualized by conversion of pre-processed text into the proper numerical presentation. Then TF-IDF features selection method selects the important words in the third phase. The importance of a word is considered based on the occurrence of that word in the document and its relative occurrence in the whole corpus of documents. TF-IDF extracts a very high dimensional feature set. The dimensionality of the feature set is reduced, with minimum loss of accuracy, in the fourth phase via a famous mathematical technique called PCA.

The proposed FEA is simpler in implementation as it does not need pre-processing, text visualization, and dimensionality reduction. It extracts only eight Boolean features from each textual content. The input arrays make the FEA more flexible and dynamic. The algorithm can be adjusted according to the changing interest of information by updating the input arrays according to the new situation. The goal of the criminal activist is to create unrest in society. For which they support both religious and secular extremism at a time. So, the classifier trained, with TF-IDF/GI/IG/Chi Square statistics followed by the PCA, for identification of one type of malicious content could not identify other types of malicious content. In this situation, the classifier needs to be retrained. While the proposed FEA can identify both types of malicious contents without re-training. To display this flexibility, the FEA requires to update the input arrays according to the expected change in contents.

The analysis of the results shows that the desired objectives have been achieved. The FEA improved classifiers' performance, reduced the time and space complexity of the features extraction process, and adjust according to the dynamic nature of the problem. The same is discussed in detail as under.

1) PERFORMANCE IMPROVEMENT

The results show that classifiers' performance is more stable and improved with the FEA used for features extraction instead of using the combination of feature selection (like TF-IDF/GI/IG/Chi Square Statistics) and feature extraction

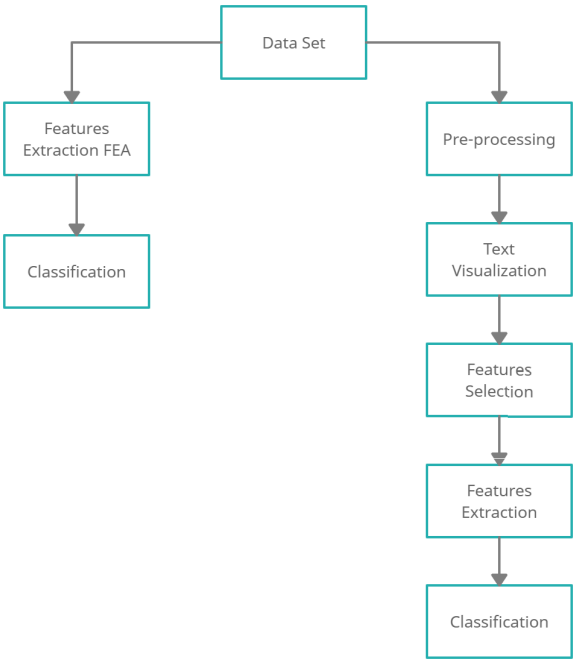

FIGURE 5. Flow chart for FEA vs TF-IDF+PCA.

TABLE 2. Performance evaluation of different feature extraction methods.

Classifier	Features method	Accuracy	Precision	Recall	F1-Score
SVM	FEA	**0.964**	**0.952**	**0.965**	**0.957**
	TF-IDF+PCA	0.928	0.926	0.925	0.929
	Chi-Squ +PCA	0.801	0.824	0.784	0.813
	GI+PCA	0.720	0.743	0.710	0.700
	IG +PCA	0.683	0.695	0.697	0.696
DT	FEA	**0.956**	**0.951**	**0.943**	**0.955**
	TF-IDF+PCA	0.822	0.851	0.801	0.810
	Chi-Squ +PCA	0.701	0.726	0.742	0.714
	GI+PCA	0.730	0.735	0.720	0.749
	IG+PCA	0.742	0.775	0.726	0.701
RF	FEA	**0.952**	**0.956**	**0.951**	**0.950**
	TF-IDF+PCA	0.805	0.843	0.779	0.787
	Chi-Squ +PCA	0.866	0.829	0.791	0.824
	GI+PCA	0.740	0.764	0.735	0.782
	IG +PCA	0.739	0.742	0.761	0.728
MNB	FEA	**0.959**	**0.959**	**0.959**	**0.9591**
	TF-IDF+PCA	0.901	0.899	0.898	0.894
	Chi-Squ +PCA	0.791	0.804	0.795	0.804
	GI+PCA	0.753	0.768	0.735	0.751
	IG +PCA	0.670	0.689	0.670	0.681

techniques (like PCA). The performance comparison of the classifiers is provided in Table 2. The table describes the computed values of performance evaluation parameters including accuracy, precision, recall and F1-Score. The accuracy is used as a measures of how close the result is to the standard one, precision is the measure of false-positive, i.e., the data point classified as positive that should not be positive. The Recall value is the measure of false-negative, i.e., miss-classification in classifying the data point as negative. The F1-Score is the weighted average of precision and recall. It takes both false positives and false negatives into account.

Clearly, the statistics of Table 2 show that the FEA equally improves the performance of all the classifiers with regards to the above-mentioned bench marks.

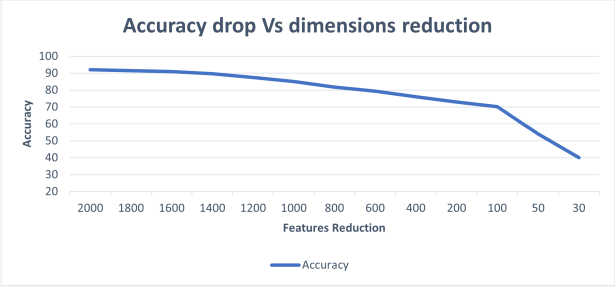

FIGURE 6. Comparison of accuracy vs dimension reduction.

TABLE 3. Comparison of dimensionality for feature extraction methods.

Classifier	Features extraction method	No. of dimensions per feature vector	Accuracy
SVM	FEA	8	**0.964**
	TF-IDF+PCA	2,000	0.928
	Chi-Squ +PCA	2,000	0.801
	GI+PCA	2,000	0.720
	IG+PCA	2,000	0.683
DT	FEA	8	**0.956**
	TF-IDF+PCA	2,000	0.822
	Chi-Squ +PCA	2,000	0.701
	GI+PCA	2,000	0.730
	IG+PCA	2,000	0.742
RF	FEA	8	**0.952**
	TF-IDF+PCA	2,000	0.805
	Chi-Squ +PCA	2,000	0.866
	GI+PCA	2,000	0.740
	IG+PCA	2,000	0.739
MNB	FEA	8	**0.959**
	TF-IDF+PCA	2,000	0.901
	Chi-Squ +PCA	2,000	0.791
	GI+PCA	2,000	0.753
	IG+PCA	2,000	0.670

2) DIMENSIONALITY REDUCTION

The high dimensionality of the features set is a curse in the field of text classification. Because dealing with high dimensional features increases the computational overhead of the process. TF-IDF features extraction method generates a very high dimensional features set; reducing its dimensionality downgrades the performance of classifiers, as can be seen in Figure 6. Using a statistical method like PCA and requires additional time and space.

In contrast, the FEA generates very low dimensional feature set that reduces the time and space complexity of the process and enhances the performance of classifiers. The following table, Table 3, presents the statistical comparison of both features' extraction methods with regards to dimensions per feature vector and accuracy.

The statistics in Table 3 show that using FEA, the classifiers give better results with eight dimensions per features vector than with features set of 2000 dimensions extracted via other features extraction methods.

3) REDUCTION IN TIME AND SPACE COMPLEXITY

Text classification is a highly computational complex task as it consumes a lot of time in pre-processing, feature selection, feature extraction, and classification. One of the objective this work is to reduce the execution time that is achieved through the proposed FEA. The proposed approach avoids the pre-processing of text and reduces the time complexity of feature selection, features extraction, and classification. A comparative analysis of FEA versus features extraction method (TF-IDF + PCA) is given in this section.

The text classification based on TF-IDF + PCA features extraction method pre-processes the text and then feeds the pre-processed text into TF-IDF for features selection and then to PCA for features extraction. The time complexity of pre-processing is

$$T_p(n, L) = O(nL) \tag{4}$$

The features are selected from the pre-processed text via TF-IDF whose time complexity is

$$T_{tfidf}(n, L) = O(nL.lognL) \tag{5}$$

where n is the total number of textual posts in the dataset and L is the average number of words in a post. The selected features set is then passed to PCA for extraction of features set to reduced dimensions. As PCA comprises two sub-processes, the covariance matrix computation whose computational cost is $O(p^2n)$ and single value decomposition whose computational cost is $O(p^3)$. So, the net computational overhead of PCA becomes

$$T_{pca}(n, p) = O(p^2n + p^3) \tag{6}$$

where n is the total number of data points in the dataset and p is the number of dimensions per data point. PCA is usually implemented in a situation where the numbers of dimensions per data point are nearly equal to the number of data points, and the same situation is in this case. If it is assumed that $n = p$, then the computational complexity of PCA will be

$$T_{pca}(n, p) = O(n^3) \tag{7}$$

The net computational overhead of the whole process of features extraction is the sum of the computational overhead of preprocessing, TF-IDF, and PCA, as given below

$$T_1(n, L, p) = O(nL) + O(nL.lognL) + O(n^3) \tag{8}$$

$$T_1(n, L, p) = O(n^3 + nL.longnL + nL) \tag{9}$$

which is a combination of cubic, linear algorithmic, and linear function.

This shows that feature extraction from the text is a highly computational task. Ignoring the lower order terms, the growth of the above function still remains cubic.

$$T_1(n, L, p) = O(n^3) \tag{10}$$

On the other hand, the FEA reduces the computational overhead of this highly computational task by avoiding the preprocessing of text because the FEA does not need pre-processed text. The proposed FEA also restricts a two-stage features extraction process to just a single stage by extracting a low dimensional feature set from the text, which does not need any statistical method for dimensionality reduction. Besides this, FEA features are Boolean instead of decimal.

TABLE 4. Comparison of time and space complexity for feature extraction methods.

Features extraction method	Time
Preprocessing+TF-IDF+PCA	$T_1(n, L, p) = O(n^3 + nL.lognL + nL)$
FEA	$T_2(n, L, b) = O(nLb)$
	Memory
Preprocessing+TF-IDF+PCA	$O(nL) + O(p * n) + O(p * n + p^2)$
FEA	$O(8 * n)$

The net time complexity function of the FEA calculated in the featured extraction section Equ 1 is given below

$$T_2(n, L, b) = O(nL) \tag{11}$$

When L is relatively small (as usual, it is), then the above function T_2 will behave as a linear function, which will be its best case, while in the worse situation when L will be nearly equal to n, the function T_2 will behave as quadratic.

Comparing both functions T_1 and T_2 show that the FEA has enormously reduced the computational overhead of features extraction because T_1 is cubic in its growth while T_2 is quadratic in the worst case (when L is assumed to be equal to n) and linear in the best case (when L is assumed constant).

Similarly, considering the space complexity of both feature extraction methods (TF-IDF + PCA and FEA), it is evident that TF-IDF+PCA requires more memory than FEA. If n is the number of documents, L is the average number of words per documents and p is the number of words in a word dictionary, then TF-IDF will need $O(n * p)$ memory for holding each TF, IDF, and TF-IDF values in memory and $O(nL)$ memory for preprocessing. While an additional memory of $O(np + p2)$ will be required for the execution of PCA. As compared to this, FEA extracts eight one-bit features per features vector, which will need only $O(n * 8)$ memory to hold the features set in the main memory. Comparative presentation of each method is given in the following table, Table 4.

4) FLEXIBILITY OF FEA

Flexibility means the measure of adjustment of FEA to the dynamic nature of the problem. As to identify criminal activists' contents on social media in real-time require retraining of classifiers to keep the acceptable level of performance because the agenda of criminal groups changes with time, due to which the classifiers (using features set extracted via TF-IDF+PCA) lose their performance in identifying these new social media contents. While the classifiers with FEA perform well without retraining as evident in the following table, Table 5.

The statistics given in Table 5 are the results of classifiers on dataset B, which contains textual social media contents of a new campaign of criminal groups. The results show that when social media criminal activists change their strategies, then the classifiers using TF-IDF + PCA for features extraction are unable to identify malicious contents successfully until retrained according to the criminal's new strategies. While the classifiers using FEA successfully classifies the contents of dataset B (with 97% accuracy) without retraining.

TABLE 5. Performance evaluation results on dataset B.

Classifier	Features method	Accuracy	Precision	Recall	F1-Score
SVM	FEA	**0.972**	**0.979**	**0.973**	**0.974**
	TF-IDF+PCA	0.676	0.677	0.673	0.674
	Chi-Squ +PCA	0.480	0.315	0.488	0.334
	GI+PCA	0.421	0.364	0.402	0.370
	IG +PCA	0.392	0.380	0.399	0.391
DT	FEA	**0.961**	**0.967**	**0.961**	**0.963**
	TF-IDF+PCA	0.489	0.318	0.487	0.330
	Chi-Squ +PCA	0.318	0.306	0.318	0.324
	GI+PCA	0.345	0.336	0.351	0.321
	IG +PCA	0.340	0.346	0.343	0.347
RF	FEA	**0.970**	**0.971**	**0.970**	**0.970**
	TF-IDF+PCA	0.487	0.365	0.480	0.341
	Chi-Squ +PCA	0.418	0.404	0.399	0.410
	GI+PCA	0.371	0.361	0.378	0.371
	IG +PCA	0.369	0.367	0.361	0.360
MNB	FEA	**0.968**	**0.967**	**0.966**	**0.969**
	TF-IDF+PCA	0.661	0.660	0.668	0.656
	Chi-Squ +PCA	0.620	0.621	0.624	0.623
	GI+PCA	0.638	0.634	0.639	0.632
	IG +PCA	0.641	0.640	0.645	0.639

Criminals use social media in a very dynamic way; they daily change their strategies and highlight daily life in different ways to gain their objectives, due to which the themes of their content on social media change regularly. To adjust to such a situation, the classifiers with TF-IDF+PCA will need to be retrained on a dataset of newly changed contents. While the classifiers with FEA adjust themselves, without retraining, just by updating the input lists of FEA.

V. CONCLUSION AND FUTURE WORK

Experimental results reveal that the proposed approach in the current study performs well on the collected dataset with the extra characteristics of better flexibility and adaptability to the targeted criminal group's changing interests and strategies. The proposed method can identify the textual malicious contents on social media with high accuracy and less computational complexity. The new features extraction method adapts satisfactorily to the changing datasets with updates according to the expected strategy of the concerned criminal group. To limit the scope, only the textual part of social media content is considered, and images and videos are not taken into account. Future work can be undertaken considering the images and videos from social media by suitably extracting features and identification of criminal content.

ACKNOWLEDGMENT

(Imran Shafi and Sadia Din are co-first authors.)

REFERENCES

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, and T. Jebara, "Social science. Computational social science," *Sci. New York, NY*, vol. 323, no. 5915, pp. 721–723, 2009.

[2] J. Gao, Y.-C. Zhang, and T. Zhou, "Computational socioeconomics," *Phys. Rep.*, vol. 817, pp. 1–104, Jul. 2019.

[3] K. Kang, C. Yoon, and E. Yi Kim, "Identifying depressive users in Twitter using multimodal analysis," in *Proc. Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2016, pp. 231–238.

[4] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan, "Detecting tension in online communities with computational Twitter analysis," *Technol. Forecasting Social Change*, vol. 95, pp. 96–108, Jun. 2015.

[5] M. L. Jibril, I. A. Mohammed, and A. Yakubu, "Social media analytics driven counterterrorism tool to improve intelligence gathering towards combating terrorism in Nigeria," *Int. J. Adv. Sci. Technol.*, vol. 107, pp. 33–42, Oct. 2017.

[6] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, "Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime," *Brit. J. Criminol.*, vol. 60, pp. 93–117, Jul. 2019.

[7] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telematics Informat.*, vol. 48, May 2020, Art. no. 101345.

[8] H.-W. Zhang, J.-F. Cao, and S.-Q. Feng, "An improved text feature selection method based on key words," in *Proc. IEEE Int. Conf. Intell. Comput. Intell. Syst.*, Oct. 2010, pp. 293–297.

[9] V. Ikoro, M. Sharmina, K. Malik, and R. Batista-Navarro, "Analyzing sentiments expressed on Twitter by UK energy company consumers," in *Proc. 5th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Oct. 2018, pp. 95–98.

[10] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[11] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 2264–2269.

[12] Z. Yao and C. Ze-wen, "Research on the construction and filter method of stop-word list in text preprocessing," in *Proc. 4th Int. Conf. Intell. Comput. Technol. Autom.*, Mar. 2011, pp. 217–221.

[13] X. Wang, R. Chen, Y. Jia, and B. Zhou, "Short text classification using Wikipedia concept based document representation," in *Proc. Int. Conf. Inf. Technol. Appl.*, Nov. 2013, pp. 471–475.

[14] W. Tian, J. Li, and H. Li, "A method of feature selection based on Word2 Vec in text categorization," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9452–9455.

[15] S. Macdonald, S. G. Correia, and A.-L. Watkin, "Regulating terrorist content on social media: Automation and the rule of law," *Int. J. Law Context*, vol. 15, no. 2, pp. 183–197, Jun. 2019.

[16] T. Anwar and M. Abulaish, "A social graph based text mining framework for chat log investigation," *Digit. Invest.*, vol. 11, no. 4, pp. 349–362, Dec. 2014.

[17] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, "Rumor propagation dynamic model based on evolutionary game and anti-rumor," *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, Jan. 2019.

[18] Y. Xiao, Q. Yang, C. Sang, and Y. Liu, "Rumor diffusion model based on representation learning and anti-rumor," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 3, pp. 1910–1923, Sep. 2020.

[19] Q. Q. L. H. X. Xiao and W. Li, "A rumor & anti-rumor propagation model based on data enhancement and evolutionary game," *IEEE Trans. Emerg. Topics Comput.*, early access, Oct. 27, 2020, doi: 10.1109/TETC.2020.3034188.

[20] V. S. Chavan and S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 2354–2358.

[21] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 5–33, Jan. 2017.

[22] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on Twitter and correlation between events and sentiment," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2016, pp. 52–57.

[23] B. Mutlu, M. Mutlu, K. Oztoprak, and E. Dogdu, "Identifying trolls and determining terror awareness level in social networks using a scalable framework," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 1792–1798.

[24] B. Heredia, T. M. Khoshgoftaar, J. Prusa, and M. Crawford, "Cross-domain sentiment analysis: An empirical investigation," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 160–165.

[25] J. E. The, A. F. Wicaksono, and M. Adriani, "A two-stage emotion detection on Indonesian tweets," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Oct. 2015, pp. 143–146.

[26] C. Paris, H. Christensen, P. Batterham, and B. O'Dea, "Exploring emotions in social media," in *Proc. IEEE Conf. Collaboration Internet Comput. (CIC)*, Oct. 2015, pp. 54–61.

[27] C. A. Steed, M. Drouhard, J. Beaver, J. Pyle, and P. L. Bogen, "Matisse: A visual analytics system for exploring emotion trends in social media text streams," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 807–814.

[28] S. Pallavi, R. K. Bedi, and S. K. Gupta, "Geo-spatial social media analytics for counter-terrorism," *Adv. Math., Sci. J.*, vol. 9, no. 6, pp. 3813–3820, 2020.

[29] M. Bérubé, T.-U. Tang, F. Fortin, S. Ozalp, M. L. Williams, and P. Burnap, "Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 Manchester arena terrorist attack," *Forensic Sci. Int.*, vol. 313, Aug. 2020, Art. no. 110364.

[30] M. KhosraviNik and M. Amer, "Social media and terrorism discourse: The islamic state's (IS) social media discursive content and practices," *Crit. Discourse Stud.*, pp. 1–20, Nov. 2020.

[31] T. M. Kibtiah, E. Miranda, Y. Fernando, and M. Aryuni, "Terrorism, social media and text mining technique: Review of six years past studies," in *Proc. Int. Conf. Inf. Manage. Technol. (ICIMTech)*, Aug. 2020, pp. 571–576.

[32] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *J. Inf. Technol. Politics*, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[33] A. Kumari, "Detection of threatening user accounts on Twitter social media database," *Int. J. Intell. Eng. Inform.*, vol. 7, no. 5, pp. 457–489, 2019.

[34] R. T. Adek and M. Ula, "Systematics review on the application of social media analytics for detecting radical and extremist group," in *Proc. IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1071. Bristol, U.K.: IOP Publishing, 2021, Art. no. 012029.

[35] M. Nouh, J. R. C. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on Twitter," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 98–103.

[36] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[37] C. Peersman, C. Schulze, A. Rashid, M. Brennan, and C. Fischer, "ICOP: Live forensics to reveal previously unknown criminal media on P2P networks," *Digit. Invest.*, vol. 18, pp. 50–64, Sep. 2016.

IMRAN SHAFI received the bachelor's degree (Hons.) in aeronautical engineering from the College of Aeronautical Engineering, National University of Sciences and Technology (NUST), Pakistan, and the M.S. and Ph.D. degrees (Hons.) in computer engineering from the Centre for Advanced Studies in Engineering, Islamabad, Pakistan, in 2003 and 2009, respectively. He did advanced certificates in disciplines of computer networks and databases at Stanford University, CA, USA, in 2011 and 2012, respectively. He is currently a Professor and the Chief Executive of the Vehicle Design and Manufacturing Laboratory, NUST College of Electrical and Mechanical Engineering. He works on high resolution time-frequency signal processing and supervised neural networks. He has authored various book chapters, numerous articles in international peer-reviewed scientific journals and numerous papers in conferences of impact factor. His interests include artificial intelligence, vehicular design, cognitive communication, and digital signal processing. He is serving as a reviewer for prestigious IEEE, IET, Elsevier, Springer, and Hindawi journals and conferences.

SADIA DIN received the master's degree in computer science from Abasyn University, Islamabad, Pakistan, in 2015, and the Ph.D. degree in data science and from Kyungpook National University, South Korea, in 2020. In 2015, she was a Visiting Researcher at CCMP Lab, Kyungpook National University, where she was working on big data and the Internet of Things. She was working as a Postdoctoral Researcher at Kyungpook National University, from March 2020 to August 2020. She is currently working as an Assistant Professor with the Department of

Information and Communication Engineering, Yeungnam University, South Korea. During her Ph.D. degree, she was working on various projects, including Demosaicking and Denoising using machine/deep learning and artificial learning. Furthermore, she extended her research toward the Internet of Things, 5G, and big data analytics. At the beginning of her research career, she has published highly more than 60 journals and conferences, including IEEE Internet of Things Journal (IoT), IEEE Transactions on Industrial Informatics (TII), IEEE Wireless Communications, IEEE GLOBECOM, IEEE LCN, and IEEE Infocom. In addition, she was a recipient of two Korean patents in 2019 and 2020. Her research interests include demosaicking and denoising using machine/deep learning, artificial learning, big data analytics, 5G, and the IoT. She was a recipient of two international awards, including the Research Internship at CCMP Research Lab, Kyungpook National University, in June 2015, and the CSE Best Research Award at Kyungpook National University, in October 2019. She was the Chair for the IEEE International Conference on Local Computer Networks (LCN'18). In IEEE LCN 2017 in Singapore, she has chair couple of sessions. She is a Guest Editor in journal of Wiley, including *Big Data* and *Microprocessors and Microsystems*.

ZAHID HUSSAIN received the M.S. degree in computer science from Abasyn University, Islamabad. He is currently pursuing career in the health sector. His interest includes deep learning tools implementation towards real-world problems.

IMRAN ASHRAF received the M.S. degree in computer science from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2010, and the Ph.D. degree in information and communication engineering from Yeungnam University, Gyeongsan, South Korea, in 2018. He worked as a Postdoctoral Fellow at Yeungnam University. He is currently working as an Assistant Professor with the Information and Communication Engineering Department, Yeungnam University. His research interests include indoor positioning and localization with 5G and beyond, advanced location-based services in wireless communication, and data analytics using machine and deep learning approaches.

GYU SANG CHOI received the Ph.D. degree from the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA, in 2005. He was a Research Staff Member at the Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, from 2006 to 2009. Since 2009, he has been a Faculty Member with the Department of Information and Communication, Yeungnam University, South Korea. His research interests include non-volatile memory and storage systems.