Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 55 Issue: 09:2022
DOI10.17605/OSF.IO/7FW5Y

EXTREMISM CLASSIFICATION BASED ON TWITTER TEXT USING ENSEMBLE LEARNING

[1]MUHAMMAD SABIR, [2]DOST MUHAMMAD KHAN, [3]FAISAL SHAHZAD, [4]MUHAMMAD NAUMAN, and [5]ASAD ALI

[1, 2, 3]Department of Information Technology, Islamia University of Bahawalpur, Pakistan.
[4]Department of Artificial Intelligence, Islamia University of Bahawalpur, Pakistan.
[5]Department of Computer Science and Information Technology, National College of Business Administration and Economic Bahawalpur, Pakistan.

Abstract

The internet, especially social media networks, has changed the way that the criminal and militant groups influence and accelerate individuals. According to a recent report, the way these groups operate begins with exposing a large online audience to extreme content and moving to a more open online platform for further interest. Therefore, it is important to identify online extreme content to limit its spread and distribution. The purpose of this research is to categorize ways to detect extreme content on social media automatically. Identification of numerous signals included in the text, psychology, and behavior, which allow an extremist categorization from Twitter text messages particularly. The proposed approach is based on machine learning methods that help to achieve the goals to determine the extreme contents in the textual data. The dataset used in this research was extracted from twitter based on threatening, terrorism, and cyber-bullying that are classified as extreme content to analyze the psychological behavior of the people and to prevent the crime. The proposed research perform analysis based on machine learning models including naive bayes, decision tree, random forest, support vector classifier, logistic regression, and ensemble model based on logistic regression, support vector classifier, and decision tree that presents a state of art classification system using feature fusion and hyper parameter optimization techniques such as Grid search and A Fast and Lightweight Auto ML Library (FLMAL) techniques. The major contribution of the study is to build Extremism Classification based on Ensemble Optimized Feature Fusion (ECEOFF) system using K-fold cross validation. The evaluation parameters such as accuracy, recall, precision, and f1 score are used. The proposed approach gain 99.8% accuracy, 99% precision, 97.5% recall, and 99% f1 score results that outperforms the other machine learning model used in comparative analyses. Further, the comparative analysis with three different feature techniques and two hyper parameter tuning technique is carried out to evaluate the proposed research in scientific manners.

Index Terms: Extremism, Social profiling, Classification, Supervised Learning, Machine Learning

1) INTRODUCTION

Extremism and criminal activities affect directly human beings and also there is a number of other con- sequences that a nation faces which may vary from nation to nation. The world is facing security issues as well as economic issues and a large section of the population in the World is forced to live below the poverty line. Due to this, International organizations also imposed restrictions on different countries. There is a number of factors that help these extremist groups and criminal mindset etc [Khan et al.,2019], [Lee et al., 2022] in the form of finance and technical training. The digital world is part of our life, as every module around us is smart which deals with data, [Shoeibi et al., 2021], [Iqbal et al., 2021] so information delivery, as well as storage and its related linked processes, are commonly based on text. There are many ways to produce digital text, major ways to produce a larger amount of textual data may include blogs, documents, news, search queries, tweets, message conversations, tags, social network posts, [Chaparro et al., 2021], [Roy et al., 2020].

Machine learning is used to develop machines to deal with data and gets information more productively. Machine learning, as with other research fields and areas, is same popular while dealing with text mining and especially extreme sentiment analysis from text data on social media. There are many types of machine learning but the more popular are supervised machine learning, unsupervised, reinforcement learning and ensemble learning. Ensemble learning is equally important from radical mind detection from text data, as with other types of machine learning. In ensemble learning multiple models of machine learning are combined to solve computing problems and improve performance of the model. Ensemble learning is also famous for decision making, feature selection, error correction and feature fusion. [Mahesh, 2020], [Mashechkin et al., 2019] Understanding a huge amount of text is important and it is also a fundamental task involved in research contribution in the areas of text summarization, recommended systems, search engines, online marketing, and so on. However, text understanding for the computer is not an easy task, especially ambiguous and noisy text. It is nearly impossible to cover all topics in text analytics. The focus of this research is mainly on crime surveillance and extremism detection to design a digital profile of the suspects by using textual data in online tweets as an information carrier. Crime surveillance aims to prevent from occurrence of crimes and an effective action plan could be developed to minimize the ratio and effects of any crime [Lal et al., 2020].

Crime such as terrorist attacks, organized gang attacks, serial killings, etc. causes loss of human life and property and physical destruction around the world. In order to minimize the loss from crime or attack, an essential task is to efficiently analyze and understand crime reports updates [de Carvalho et al.,] which can be gathered and extracted from web- sites having news and reports that are related to the criminal activities. In recent years, social media got popular in public and had gained popularity due to its ease of use. Social platforms, such as Twitter reveals much about public taste and many studies focus on product promotion and sentiment analysis [Sangher and Singh, 2019]. All crimes are now covered by cybercrimes where the internet, computers, and mobile phones are the major facilitators. There is a number of reasons to limit the research on cyberbullying detection which may include unavailability of data sets, user privacy, the hidden identity of predicators, etc. [Yu et al., 2019].

Text mining and machine learning is useful to structure the human language's irregular patterns. Mostly people use text data to communicate with each other. To develop high-quality applications, the main goal of text mining is to get a system trained to seek information and process this information from human language. The art of sharing meaningful information is considered an advantage. Text mining techniques extract unstructured data and convert it to meaningful data that can be used for a specific purpose [Andleeb et al., 2019]. It is similar to text mining, which will incorporate the full NLP scheme into its system to effectively study human language and structure the unstructured data patterns accordingly. As advancement in technology day by day, the text mining system gets better and that's what everyone is looking for. The Twitter APIs can be used to extract data from Twitter [Salloum et al., 2017]. For the classification of tweets negative and positive classes can be created. Negative tweets mean the criminal and extremist tweets and positive means other than the negative [Yu et al., 2020], which can be classified as neutral tweets. Due to the presence of incorrect spellings and slang tweet text, the extraction of keywords from tweets is a difficult task [Fraiwan, 2022]. Before extracting the keywords from text data, for filtering out slang words and misspell words, some preprocessing techniques can be used. The slang and misspell words in twitter text data must be replaced with their nearest relevant words by using the slang word dictionary.

The focus of this research is to apply machine learning based classification on Twitter text data to determine extremism in tweets. It is necessary to use the internet, obtain intelligence and monitor important sites. Useful data for law enforcement purposes can be accessed through social networking sites to gain insight, identify threats, make predictions and perform many other analyses [Tundis et al., 2018]. For a better understanding of the people and making of decisions of any type, these types of analyses will provide new methods of investigation to authorities, law enforcement agencies, and private organizations. Mostly, these types of analyses use data that are publicly available on social media websites/apps and users always have the option to set visibility of data public or private. When an organization or researcher obtains data illegally, then privacy and trust get breached. The application of the proposed study is the collection of data from Twitter and the use of text mining approaches for digital profiling for digital policing, intelligence, and law enforcement agencies. Society is affected by extremism in the last two decades and faced a number of issues nationally and internationally including economic issues. There are extremist groups and individuals, who are benefited extraordinarily. There will be more opportunities to spread extremism by using the Internet and OSN platforms and to facilitate potential recruitment opportunities. There is not only religious extremism but also other factors that cause extremism which may include politics, socialists, and linguistics nationally and internationally. These goals and targets cannot be completed until the required technologies are made available to the countries facing high crime rates and extremism, for the development of new strategies and models to counter the security issues and malefactors. The research on text analysis using Twitter has been increased since 2006 and this field is expected to grow in the coming years [Karami et al., 2020]. Law and enforcement is a major issue of every society. Most of the time law enforcement agencies take action after adverse incidents occur. There may some traditional methods exist to profile a criminal that commits some crime. There should be a mechanism to minimize the occurrence of adverse events. Social media is the platform where people share their feelings.

People with an extreme criminal mindset need treatment, should be profiled digitally and given treatment by a psychologist. This might be done by the sentiment analysis from different types of data, especially text data shared on social media by using different types of given methods and develop new techniques. Data about the criminals and extremists is maintained in databases of law enforcement and regulatory agencies. Most of the information about criminals and extremist groups and individuals can also be discovered in unstructured text in different types of literature, newspapers, social media, emails, text messages, and blogs. In that regard few questions are rising such as how textual data on Twitter is useful to detect extremist minds? What characteristics are useful to classify online extreme content? And what kind of methods can be used and developed to avoid misleading outcomes in the classification of extremism from the Twitter text? The contributions of this research are:

1) The analysis of extreme and criminal information shared by criminal groups or individuals by using machine learning techniques and develop a new state of the art Extremism Classification based on Ensemble Optimized Feature Fusion (ECEOFF) system for the classification of extremism.
2) Evaluate these models on Twitter to automatically identify online criminal and extremist tweets for a digital profile.
3) The preprocessing steps have been designed to eliminate the unnecessary contents from the text data.
4) The three types of feature engineering techniques are used to perform experiment. The Term-Frequency (TF) is used with Term Frequency-Inverse Document Frequency (TF-IDF), and feature fusion of TF and TF- IDF.
5) The proposed hybrid model was created by combining three machine learning models such as decision tree, support vector classifier and logistic regression using fast and Lightweight Auto ML Library (FLMAL).
6) The K-fold cross validation technique is used for the dataset splition based on highest results achieved with final splition ratio. Lile 70% training and 30% is used for testing.
7) Different machine learning models such as naive bayes, decision tree, random forest, support vector classifier, logistic regression, and ensembled model are used to analyze the predicted results and compare them with the proposed hybrid model.
8) Proposed research is evaluated by using accuracy, re- call, precision and f1 score.

The rest of the article is divided into the following parts: Part II, this section presents a literature review and previous studies. Part III explains the structure and methods of the proposed approach. Part IV is the evaluative analysis of the proposed approach. Part V concludes this research.

2) LITERATURE REVIEW

There is lot of work is published by the researchers relevant to social data mining, text mining, sentiment analysis and digital profiling [Akhter et al., 2018], [Razzaq et al., 2019], [Mittal and Patidar, 2019], [Razzaq et al., 2019], [Feizollah et al., 2019], [Iqbal et al., 2019], [Ahmad et al., 2019], [Gaikwad et al., 2021b], [Aldera et al., 2021], [Gaikwad et al., 2021a], [Alghamdi and Selamat, 2022]. But there is always an improvement needed. Ensemble learning can be used for improvement of models by combining different machine learning models for extremism detection from text data especially social media text. Ensemble optimized feature fusion can be used for extremism detection [Mahesh, 2020].

[Iqbal et al., 2019] described an approach to improve accuracy and scalability that is a framework to minimize the gap between machine learning and lexicon approach. New algorithm is proposed in this study to reduce the features set and address the issues of scalability that grow when feature set grow. This proposed study used a hybrid approach to reduce up to 42% the size of feature set without compromising accuracy. After the comparison of proposed model with different feature reduction methods base on PCS and LSA, this is about 15% more accurate than the PCA and about 40% more better than LSA. While evaluating the sentiment analysis framework on other metrics, these are also improved.

[Lal et al., 2020] discussed a method to analyze the twitter text data for identify tweets with criminal text data, which requires more attention of the police to take action against such criminal and extremists, was presented. This research is used to classify 369 tweets into two classes such as criminal tweets and normal tweets by using text mining approach. Different classifier are used for the classification of tweets that includes Naive Bayesian, Random Forest, J48 and ZeroR. The Random Forest classifier is the best classifier that offer higher accuracy as comparison of the four classifiers. Shakeel et al. [Ahmad et al., 2019] proposed a framework for analyzing terrorism-related content that focuses on classifying tweets into radical and non- radical classes. Based on user-generated social media posts on Twitter, we are developing a tweet classification system using deep learning-based emotion analysis techniques to classify tweets as radical or non-radical.

[Akhter et al., 2018] aimed to identify activity patterns in several places indicates of future events. For example, the arrest of a gang leader can lead to an attack by a nearby rival gang, resulting in murder. Access point mapping is one of the most popular approaches for mapping crime- prone areas. [Meliana et al., 2019] presented the phrase “Tiger’s thumbs off" contains inappropriate comments on social media. Take us to the criminal section and if the words on social networks can be legally justified, an example is threatening. Yes, and threatening is one of the articles of the ITE law. The threat has been removed from Twitter’s social media and you can see an example on Twitter. Many of these threats, when recovering data on social networks, several methods are used to recover it. The method used is a clustering method or a data aggregation method. [Kathuria et al., 2019] said that tweet collection was retrieved using the Twitter Streaming API. The extracted data as text, so we had to clean it up, remove redundancy and use our computing power to convert it into a format that allowed us to know the feelings of the tweet. We used Word2Vec to do this, and then we used a deep learning system to identify Tweets and categorize them as having positive or negative emotions. In the future, we suggest extending this system to social media sites such as Facebook to form an online interface.

[Sharif et al., 2019] discussed the analysis of extreme content form the internet especially from social media is very challenging and new research area due to noisy, short, dynamic and context dependent nature of data. In this research tweets were scanned and labeled into two classes that were extreme and non-extreme and the extreme class has two sub classes: pro-Afghan and pro-Taliban. Principal Component Analysis (PCA) was used for the exploratory data analysis. TF-IDF used for feature selection from the huge data space into low data space. The classification algorithms Naïve Bays, KNN, SVM and hybrid methods were used in this research with PCS-based feature reduction.

[Araque and Iglesias, 2020] discussed in this research two types of information sources were exploited by using the proposed machine learning model: first emotions and second embedding-based se- mantic similarity features. The first approach is designed to detect radical text in the existing emotion dictionary to generate features that can be used. This approach generates features by calculating a statistical summary of emotions in the text. And secondly SIMON model is used to extract radical text features. The method for domain word collection is proposed (FreqSelect) in this research which improves the comparison with the existing collections. The magazine data set is used that expands the scope of the research. A new dataset is used with two existing datasets that were previously studied.

[Asif et al., 2020] described that there are number of reasons for extremism among the community that may include political, religious and social uncertainty etc. and people share their sentiments on social media and different microblogging websites. English is most common language to share sentiments of internet especially on social media, but different social media platforms and websites are used to share sentiments in different local languages. Study focuses on four categories of extremism: High, low, moderate and neutral. Multilanguage textual data was used to analyze the sentiments of extremism in this research. Multilingual lexicon was developed and verified by domain experts. It attained 88% accuracy for validation. Multinomial naïve bays and linear support vector classifier algorithms were applied for the classification and attained 82% accuracy in results.

Tundis et al., 2019 described feature engineering method rotates into three phases that are used for the suspicious user analysis on social media networks that are linked to TN and OC. By combining the different techniques from well-known groups and association rules, it is based on improvement and expansion of different approaches on social media networks and other online sites. The major purpose of the research article is to identify criminal activities on social media and relationships between the activities and sort out the groups involved in some illegal activities including drugs, human trafficking, weapons, etc.

[Torregrosa et al., 2021] described that NLP is more powerful field to understand the sentiments of human. The use of language by different extremist groups is com- pared with non-extremists groups. This research discussed the analysis of different techniques, tools, approaches and datasets used for extremism classification. This study is more helpful for the researchers to find out the direction, trends and challenges in extremism analysis in text data.

[Araque and Iglesias, 2022] discussed the growth of online data attracted the researcher to get knowledge from different data sources and use this knowledge for different applications. This research focuses on the NLP that is used for the emotion detection from the text. Two new feature extraction methods AffectiveSpace and SenticNet are used. This research also discussed well known feature extraction methods TF-IDF and SIMilarity-based sentiment projectiON (SIMON). The proposed techniques are validated on different datasets that were designed to cover radical text and hate speech. The results, in this research shows the improvement in performance. The performance based criteria for method selection from given methods is also discussed for classification performance and complexity.

[Mouhssine and Khalid, 2018] described in this research that the focus of proposed framework and methods with the existing Natural processing language and machine learning methods and algorithms for text mining and sentiment analysis is to prevent innocent people from wolf extremist attacks that harms them. For the protection of people, the proposed framework is can be implemented by the government agencies. The proposed framework will be used as a social surveillance and monitoring tool as a next generation. Data set is extracted from the Facebook through API for the implementation purpose. Text mining is used for web mapping that contains extreme and violent blogs.

Table 1: Literature Review

Technique(s)	Strength(s)	Weakness(s)
This model represented a novel technique that cyber influence on different campaign this model used to trained data set for twitter data and analyze various cyber influence scenarios. This model reflected real life scenarios very effectively. Proposed system used terse triple language. [Johnson et al., 2020]	A novel idea to tackle cyber impact campaign. A very important network to detect the misinformation. Proposed system used terse triple language.	Proposed system represented samples of data analysis that could get with cismod model. Additional areas should be covered to enhance current limited scenario.
The Proposed methodology used tweet to act, also used word mover distance. This technique helped to recognized tweet and group them into similar identification of event. It used large data set and used recurrent neural network	Accuracy is 96 % and score is 86.2 % was achieved on real people data set. This method achieved an improvement for the detection of terrorist attack more than any previous model.	It based only on text data, Images and videos could also be added to extract more information about various event.

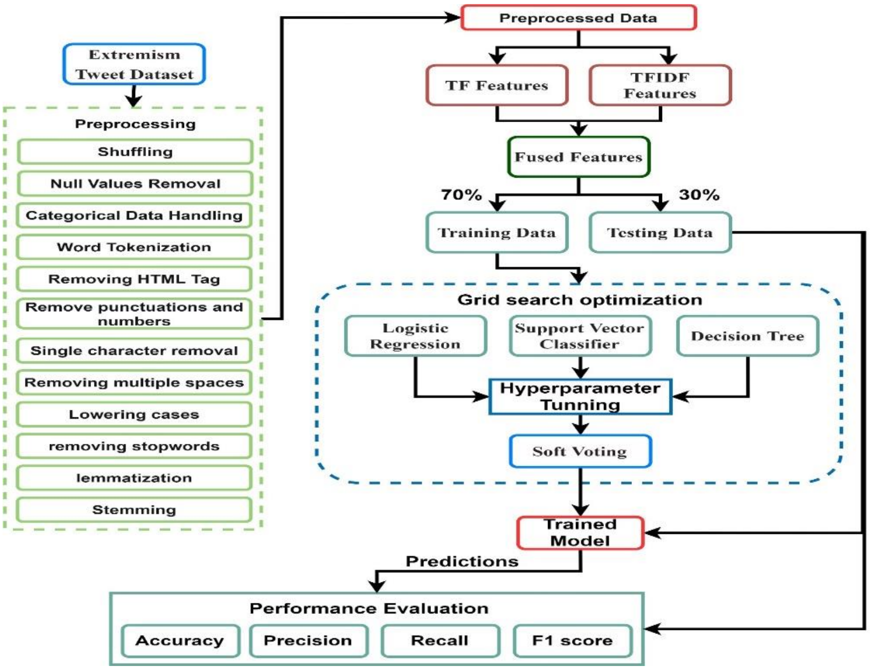
to extract the valuable information. [Iqbal et al., 2021]		
This proposed system used deep convolutional neural network. Text data was extracted with the help of convolutional network researcher will able to achieve a good performance to tackle hate speech in various scenarios. [Roy et al., 2020]	Various algorithms and models were used to detect hate speech works. A satisfactory accuracy was achieved on trained data set.	Languages other than English could not be detected. Limitation occurred on large data set.
A trained model was used. The model use more than seventeen lacs tweet in one year. The trained model showed 21% of the score and 40% of the accuracy. [Chaparro et al., 2021]	This proposed method helped in achieving not only the tone of the text but also helped in calculation of people perception. It helped in exposing main crime like robbery and how people effect by them.	The data set was limited which should be skill able.
The technique of stacked ensemble model was used to detect racism in text posted on twitter the model was able to detect 97 % of data that add racist content in them. [Lee et al., 2022]	Various deep learning technique were used together for proposed system. This model per- formed batter than previous model. 97 % of accuracy was achieved.	Lack of audio and video data limited the overall the performance of deep neural network techniques
Classification of algorithms ware used with predefined data-set random forest and bert model were used. [Aldera et al., 2021]	The use of Arabic data set other than English has opened new base to tackle online extremism. Various natural language processor has also been used.	Limitation occurred because people could use different vocabulary or code word to express their sentimental. Different combination for various features can be used will improve performance of current model.
The pattern of user behavior was analyzed by using meta data and behavior features of user were also investigated and abnormal behavior was extracted using natural language processing method and other algorithms. [Shoeibi et al., 2021]	Many artificial intelligence techniques were used to detect crime and hate speech on twitter. A network to analyze communication between users were also proposed. By using this method illegal activities can easily detect.	Other social media platform can also use this technique. Enhance algorithms can be designed to understand words and to determine hidden messages in the text.
A novel dataset was used. Based on human emotions that has not been addressed before. Specialized machine learning model has been used. [Araque and Iglesias, 2020]	This proposed methodology differentiates be- tween good content and hatred in magazines and social media by using emotions dictionary and word selection.	Text and words can also be used other than emotions. Limitation because languages al- ways change over time.
Various classification algorithms such as J48, ZeroR, Random Forest and Naive Bayesian have been used and compared [Lal et al., 2020].	RF produce the best performance while achieving 91% accuracy.	ZeroR algorithm was found to be ineffective while classifying the crime tweets.

Deep learning techniques along with CNN, LSTM, and RNN are applied to classify radical and non-radical tweets [Ahmad et al., 2019].	For a noteworthy margin improving precision, accuracy, recall, and f-measure emotion analysis technique is used to overtake standard methods.	The model shows exclusion of context- features and no automation for data cleaning. Need more deep learning methods to investigate radicalism for robust outcomes
Proposed a novel multi-level multi-task frame- work to identify crime patterns [Akhter et al., 2018].	Results show MLMT Model can predict future crime gang-homicides and homicidal-violence effectively	The study is based upon the gangs having local police record areas. Area-specific datasets results may vary prominently when applied to different scenarios.
A 3-step SAP solution is proposed to analyze text sentiment using KNN and related to older algorithms [Razzaq et al., 2019].	Sentiment Analysis and Prediction of the text show reasonable enhancement when compared to prevailing solutions	KNN may have problems regarding precision and large data as it can slow down the prediction phase.
Proposed a novel method to classify opinion mining of Client's sentiment using SVM, Naïve Bayes, and Decision Tree and compared the results [Khan et al., 2019].	State-of-the-Art methodology shows Support Vector Machine gives an accuracy of 90.30 which is very supportive for strategy makers, service providers and scholars.	The suggested model has not yet been tested on varied and big datasets because when talking about big data, results can be comparatively different.

3) METHODOLOGY

The classification of extreme content on social media using tweets is proposed in this research based on the designed Extremism Classification based on Ensemble Optimized Feature Fusion (ECEOFF) system structure. Opinion Mining is a subfield of text mining. The purpose of social text analysis is to extract text data and identify the opinions of the users. By using text mining techniques, the leading objective is to estimate attitude, sentiments and evaluate the emotions of the writer. There is a number of libraries and tools that are available for text mining as well as for classification. The Scikit or Sklearn [Pedregosa et al., 2011], NLTK [Loper and Bird, 2002], matplotlib [Hunter, 2007], seaborn [Bisong, 2019], and numpy [Oliphant, 2006] libraries based on python language implemented by using anaconda platform. These tools are used for preprocessing and classification of the Twitter text data as shown in Figure 1.

Figure 1: The proposed methodology for the Extremism tweets



A. Dataset

There is a number of datasets available publically that are extracted from Twitter and other online forums and social media networks. There are many other sources available including magazines that are listed as extremist magazines. The classification system proposed for the extreme content by using the dataset available on [A.R.Zaidi, 2021], that consists of about sixty thousand tweets labeled as 0 (Extremism) and 1(Normal). Figure 1 shows that about 10% tweets are labeled as extreme and rest of about 90% tweets are labeled as normal or neutral. Figure 2 shows the word cloud which shows that the highest number of frequency and lowest frequency in the data set.

Figure 2: The Extremism dataset presenting extreme and normal

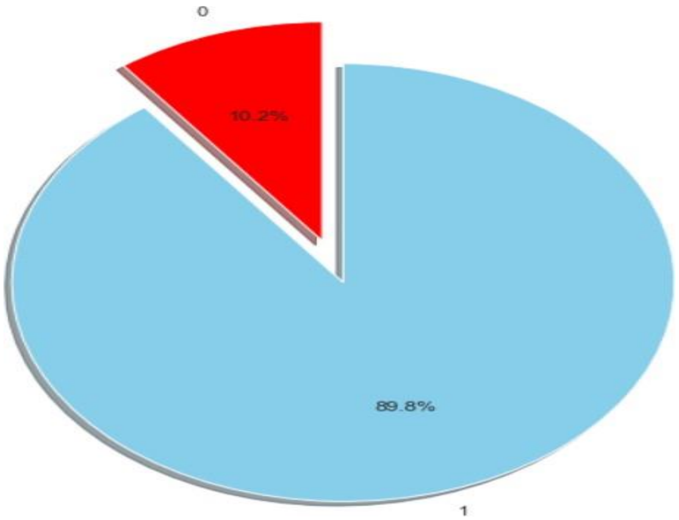


Table 2: Kaggle Data set to detect suspicious communication

Tweet type	No. of tweets	Label	Range of Class	Percentage
Extreme/Radical	6,145	0	0.0-0.05	10.2%
Non-Extreme/Neutral	53,855	1	0.95-1.0	89.8%
Total Tweets	59703			

B. Preprocessing of Twitter Text

Text preprocessing simply meant that change your text into a format that will be useful in analyzing and predicting a task. There are many types of text preprocessing which may include text lowercasing, categorical data handling, shuffling, and removal of null values, HTML tags, punctuations, multiple spaces applying stemming, lemmatization, stop- word removal, text normalization, noise removal, and text augmentation. Lowercasing is one of the simplest forms of preprocessing and this will help to get consistent output. Categorical data handling converts the labels into numerical form for processing easily. Shuffling mixes the data according to labels for better training and to gain accurate prediction rates.

Removing the HTML tags, because it is not important feature in classification of text based on sentiment values. Stemming is the process of removing inflection of the words into their root form and the resultant word may not be a real root word but it might be a canonical root form of that word. Lemmatization is similar to stemming but it converts the words into their original root form. Removal of stop words means that the extra words (is, am, are, etc.) will be removed and the focus will remain on the original words.

Text normalization means transforming the text into its standard form and noise removal on the other hand is to remove digits and extra text from the text data that interfere with the text analysis. Augmentation or enrichment adds the information to the text data that is not previously included and introduces more semantics in the text to improve prediction and the depth of the analysis.

C. Feature Extraction

Feature engineering is one of the most crucial parts of the classification process that converts the text data into numerical vectors which facilitates the computational training processing of the machine learning models. The Term Frequency (TF), Term-Frequency over Inverse-Document- Frequency (TF-IDF) and the fusion of TF and TF-IDF features are adopted to accomplish the goals of this research.

$$\mathrm{tf}(\mathrm{t},\mathrm{d}) = (\text{count of t in d})/(\text{number of words in d}) \qquad (1)$$

$$\mathrm{d\,f\,(t)} = \text{occurrence of t in documents} \qquad (2)$$

$$\mathrm{idf\,(t)} = \log \mathrm{N}/((\mathrm{df}+1)) \qquad (3)$$

Where; t — term (word)

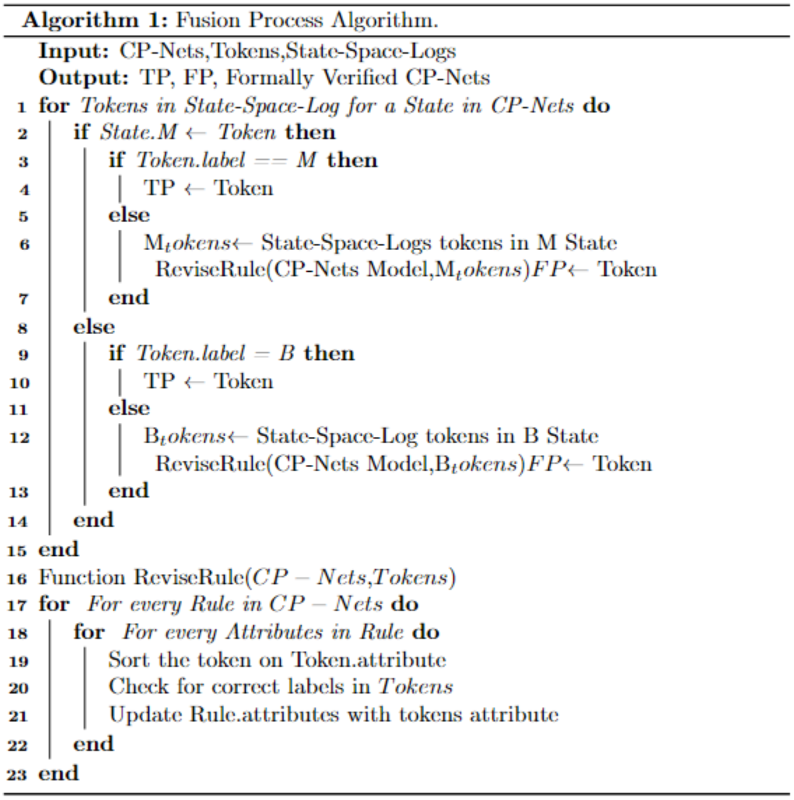
d — document (set of words)

N — count of corpusf — frequency

The TF is the simplest and strong feature engineering technique for text data that calculates the frequency of the words individually and document-wise creates a matrix of vectors. Where TF-IDF is the weighted feature engineering technique in which the data vectors are created based on the ratio of the TF and IDF.

$$\mathrm{t\,f} - \mathrm{id\,f\,(t,d)} = \mathrm{t\,f\,(t,d)} * \mathrm{id\,f\,(t)} \qquad (4)$$

The TF-IDF describes the importance of the word by the ratio of the occurrence of the word in the single document with all other documents.



Algorithm 1: Fusion Process Algorithm.

```
Input: CP-Nets,Tokens,State-Space-Logs
Output: TP, FP, Formally Verified CP-Nets
for Tokens in State-Space-Log for a State in CP-Nets do
    if State.M ← Token then
        if Token.label == M then
            TP ← Token
        else
            M_tokens← State-Space-Logs tokens in M State
             ReviseRule(CP-Nets Model,M_tokens)FP← Token
        end
    else
        if Token.label = B then
            TP ← Token
        else
            B_tokens← State-Space-Log tokens in B State
             ReviseRule(CP-Nets Model,B_tokens)FP← Token
        end
    end
end
Function ReviseRule(CP − Nets,Tokens)
for  For every Rule in CP − Nets do
    for  For every Attributes in Rule do
        Sort the token on Token.attribute
        Check for correct labels in Tokens
        Update Rule.attributes with tokens attribute
    end
end
```

D. Classification

In text mining, text classification is a process in which the text is assigned labels or categories according to its content. This is one of the basic tasks of text mining with wide applications, such as tracking extremist messages, sentiment analysis, topic labeling, and finding goals. Text data in unstructured format is everywhere: websites, emails, chats, social media, support tickets, survey responses, and more. Data in text form is a very rich source of knowledge, but due to its unstructured nature, extracting information can be time-consuming and difficult.

Researchers are quickly turning to text classification for text structuring to improve automated processes and decision- making. The machine learning models such as decision tree classifier, support vector classifier, multinomial naïve- bayes, logistic-regression, random-forest classifier and a proposed hybrid classifier based on the combination of logistic-regression, decision tree classifier, and support- vector-machine are used in this research for the detection of extremism contents on social media.

Previous research suggests that the performance of Random Forest (RF) is better in similar problems. RF generally per- forms very well because of its scalability and robustness against outliers. Random Forest (RF) normally has multiple trees and is based on a hierarchical structure, so it out- performs decision trees. Due to these properties, the RF is allowed to model non-linear decision boundaries. However, a very large amount of data is required for training. The decision tree model has based on the tree-based architecture that consists of the nodes and links in between them. The node represents the feature and the link represents the condition or relation between the parent and child nodes. The place of the feature in between nodes can be calculated based on gini or entropy, and leaf nodes represent the classes. Decision trees are best for binary, discrete, and less complex classification. The support vector machine is a hyper-plane-based model that converts the data into two-dimensional space and then

D. Evaluation Parameters

The machine learning model needs to perform prediction based on training and dataset quality. Several evaluation parameters are used to evaluate the model's performance, such as precision, f1-score, accuracy, and recall.

1. Accuracy

Accuracy is evaluation parameter that describes the total number of accurate classifications. The accuracy shows the ratio between the total number of predictions and correct predictions that present the model performance. Equations 6 and 7 present the formulation of the accuracy. Equation 7 presents the accuracy in terms of True Negative rate (TN), true positive rate (TP), False Negative rate (FN), and False Positive rate (FP). The total number of predictions is equal to TN + TP + FP + FN, and the correctly predicted values made by the model are TP + TN.

$$\mathrm{N} = \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} \qquad (5)$$

Accuracy (ACC): ACC specifies the accuracy of the classification model such that:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/\mathrm{N} \qquad (6)$$

Sensitivity (TPR): TPR specifies the correct classification rate of positive instances such as:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \qquad (7)$$

Specificity (TNR): TNR specifies the correct classification rate of negative instances such as:

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \qquad (8)$$

2. Recall

The recall is also called the sensitivity of the model. The recall is the true positive rate in the total of true positive + False Negative. Equation 8 presents the total number of extreme contents from the total number of predictions in actual or false predictions. If the comment has extreme content and the model also predicted it extreme as well, then it is considered a correct prediction. If the comment is not extreme, and the model predicts it extreme, it is regarded as a false prediction. So, recall describing the percentage of true prediction of extreme contents in Twitter comments.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{9}$$

3. Precision

The precision is also called the true positive rate. The precision is the ratio between the TP over TP + FP, as shown in equation 9 that presents the true positive rate in the total number of true predictions.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{10}$$

4. F1-Score

F1-score, as shown in equation 10, presents the balance between precision and recall by calculating the harmonic mean. Precision and recall are the class-wise prediction results in the comparison of actual and prediction outcomes. F1 score originates the results based on the precision and recall displays that how much precision and recall outputs are balanced and accurate.

$$\text{F1 Score} = \ \text{TP}/(\text{TP} + \ 1/2(\text{FP} + \text{FN})) \tag{11}$$

4) RESULTS

The identification of the extreme contents that would be the prevention of crime and other illegal activities, proposed in this research by using machine learning methods. The results and evaluation are discussed here. The proposed structure of methodology is based on the hybrid model implemented with the help of grid search and fused feature that improves the accuracy and prediction rate to detect extremism. The experiments are conducted using Dell Precision T3400 computer consists of 8 GB RAM and a 3.0 GHz processor based on two cores.

Table 3: Machine learning model results based on TF features

Machine learning model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.98	0.96	0.91	0.93
Logistic Regression	0.97	0.96	0.86	0.90
Multinomial Naïve Bayes	0.92	0.84	0.68	0.73
Support Vector Classifier	0.95	0.96	0.76	0.82
Decision Tree	0.96	0.88	0.91	0.89
Ada-Boost Classifier	0.97	0.96	0.85	0.90
Proposed ECEOFF	0.97	0.96	0.88	0.92

Table 4: Machine learning model results based on TF-IDF features

Machine learning model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.98	0.96	0.91	0.93
Logistic Regression	0.94	0.96	0.74	0.80
Multinomial Naïve Bayes	0.90	0.95	0.52	0.51
Support Vector Classifier	0.96	0.96	0.81	0.87
Decision Tree	0.97	0.94	0.91	0.93
Ada-Boost Classifier	0.97	0.96	0.85	0.90
Proposed ECEOFF	0.97	0.94	0.91	0.92

Table 5: Machine learning model results based on features fusion

Machine learning model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.98	0.97	0.97	0.94
Logistic Regression	0.97	0.87	0.95	0.90
Multinomial Naïve Bayes	0.91	0.58	0.89	0.62
Support Vector Classifier	0.97	0.94	0.88	0.87
Decision Tree	0.97	0.92	0.95	0.93
Ada-Boost Classifier	0.97	0.96	0.85	0.90
Proposed ECEOFF	0.99	0.98	0.97	0.98

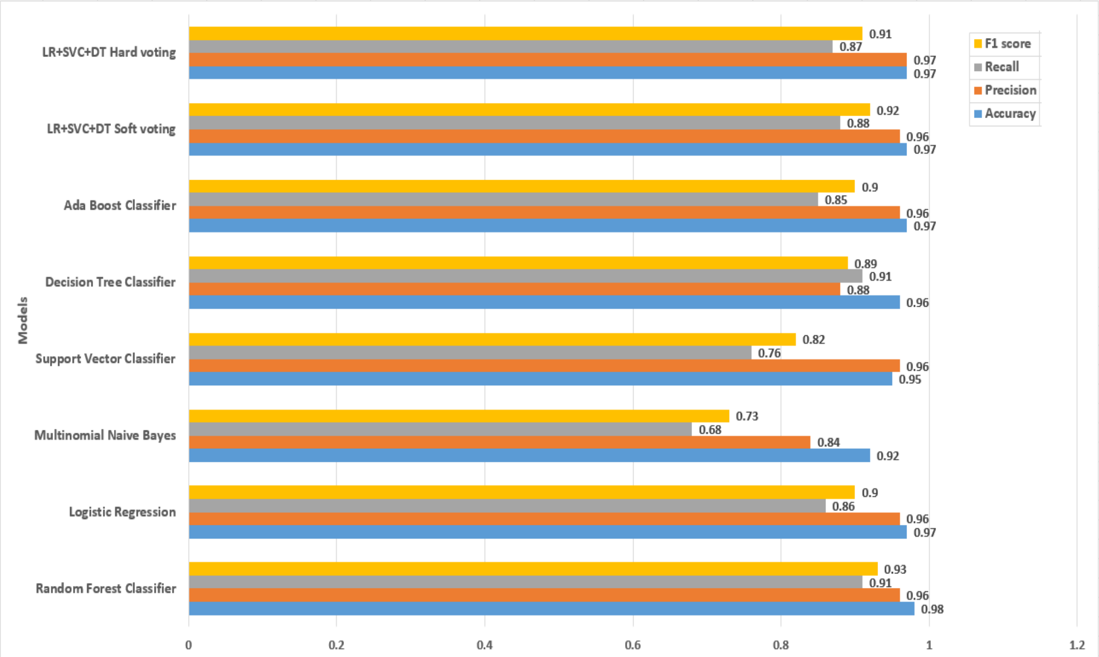
Table 6: Comparative analyses of Machine learning model results based on multiple feature engineering techniques

	Feature Fusion				TF-IDF Features				TF Features			
Machine learning model	Acc	Pre	Rec	F1-Sco	Acc	Pre	Rec	F1-Sco	Acc	Pre	Rec	F1-Sco
Random Forest	0.98	0.97	0.97	0.94	0.98	0.96	0.91	0.93	0.98	0.96	0.91	0.93
Logistic Regression	0.97	0.87	0.95	0.90	0.94	0.96	0.74	0.80	0.97	0.96	0.86	0.90
Multinomial Naïve Bayes	0.91	0.58	0.89	0.62	0.90	0.95	0.52	0.51	0.92	0.84	0.68	0.73
Support Vector Classifier	0.97	0.94	0.88	0.87	0.96	0.96	0.81	0.87	0.95	0.96	0.76	0.82
Decision Tree	0.97	0.92	0.95	0.93	0.97	0.94	0.91	0.93	0.96	0.88	0.91	0.89
Ada-Boost Classifier	0.97	0.96	0.85	0.90	0.97	0.96	0.85	0.90	0.97	0.96	0.85	0.90
Proposed ECEOFF	**0.99**	**0.98**	**0.97**	**0.98**	**0.97**	**0.94**	**0.91**	**0.92**	**0.97**	**0.96**	**0.88**	**0.92**

Table 7: The comparative analyses of hyper parameter tuning techniques based on machine learning models

Hyper parameter Tuning	Ada boost				Decision Tree			
	Acc.	Pre.	Rec.	F1-score	Acc.	Pre.	Rec.	F1-score
Grid Search	0.97	0.96	0.85	0.90	0.97	0.92	0.95	0.93
FLMAL	0.975	0.96	0.90	0.89	0.99	0.93	0.98	0.96
	Random Forest				Support Vector Machine			
Grid Search	0.98	0.97	0.97	0.94	0.97	0.94	0.88	0.87
FLMAL	0.99	0.96	0.98	0.96	0.98	0.95	0.88	0.88
	Logistic Regression				ECEOFF based on Soft Voting			
Grid Search	0.97	0.87	0.95	0.90	0.990	0.98	0.97	0.98
FLMAL	0.98	0.92	0.96	0.91	**0.998**	**0.99**	**0.975**	**0.99**
	Multinomial Naïve Bayes				ECEOFF based on Hard Voting			
Grid Search	0.91	0.58	0.89	0.62	0.98	0.94	0.91	0.92
FLMAL	0.93	0.70	0.85	0.73	0.985	0.96	0.92	0.92

Figure 4: Comparative analyses of machine leaning model TF feature results



The experiments are carried out using several machine learning models such as decision tree classifier, support vector classifier, multinomial naive bayes, logistic regression, random forest classifier, support vector classifier, multinomial naive bayes, logistic regression, Random Forest classifier, and the proposed hybrid classifier with three different feature engineering techniques and comparatively analyzed with the proposed approach. The evaluation and results are extracted by using precision, f1-score, recall, and accuracy evaluation parameters. Figure 4 presents the results of machine learning models that are used in the experimenting process with the TF feature engineering technique. The comparative analysis of results has been conducted by using presented machine learning models and evaluation parameters. The random forest model presents the highest results with 98% accuracy, 96% precision, 91% recall, and 93% f1 score.

Multinomial naive bayes model shows relatively low results such as gain 92% accuracy, 84% precision, 68% recall, and 73% f1 score. Logistic regression model gain 97% accuracy, 96% precision, 86% recall, and 90% f1-score. Support vector classifier model gain 95% accuracy, 96% precision, 76% recall, and 82% f1-score. Decision tree model gain 96% accuracy, 88% precision, 91% recall, and 89% f1-score. Ada boost classifier gain 97% accuracy, 96% precision, 85% recall, and 90% f1-score.

Figure 5: Comparative analyses of machine leaning model TF-IDF feature result

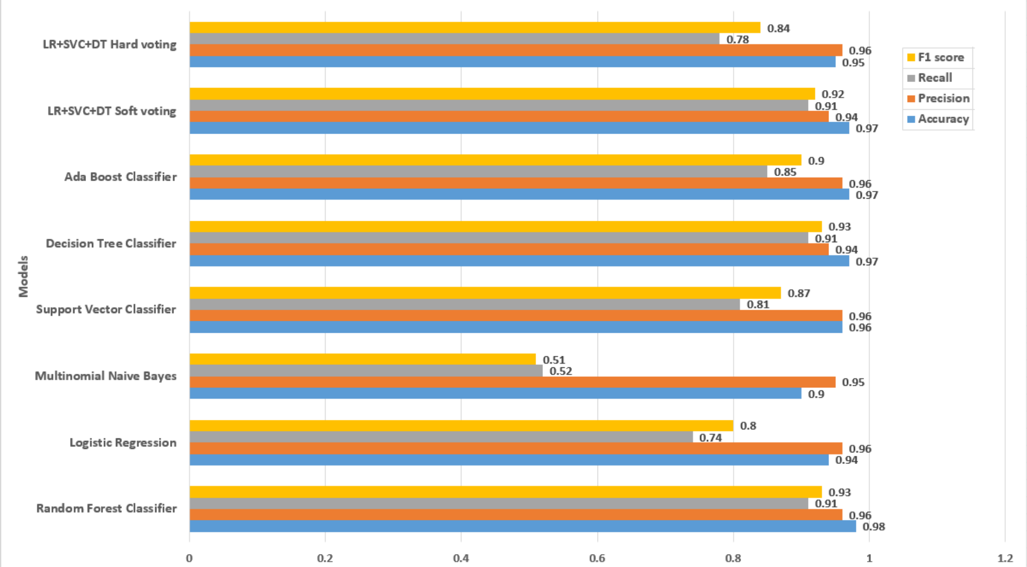


Figure 6: Experimental results with feature fusion and multiple machine learning models in comparison of proposed ECEOFF system

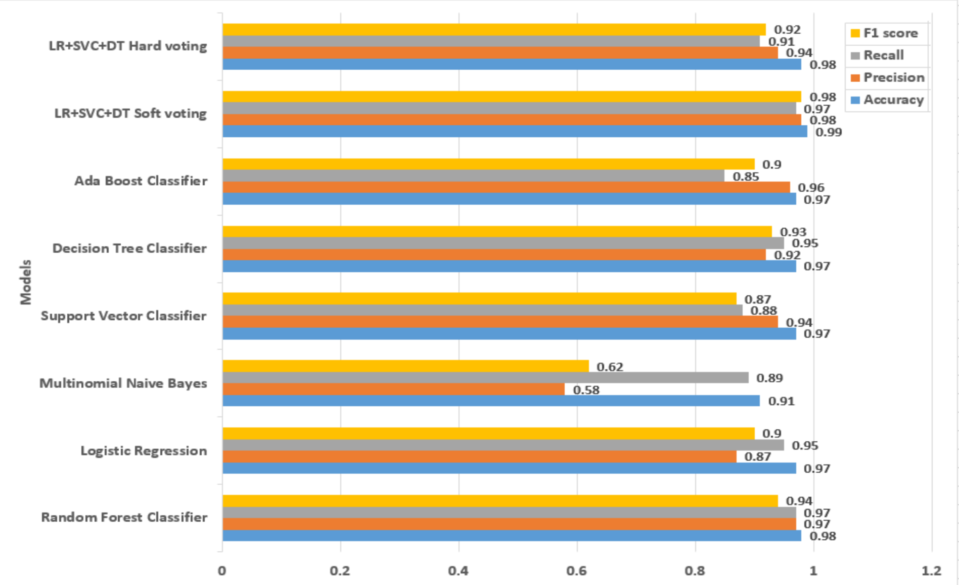
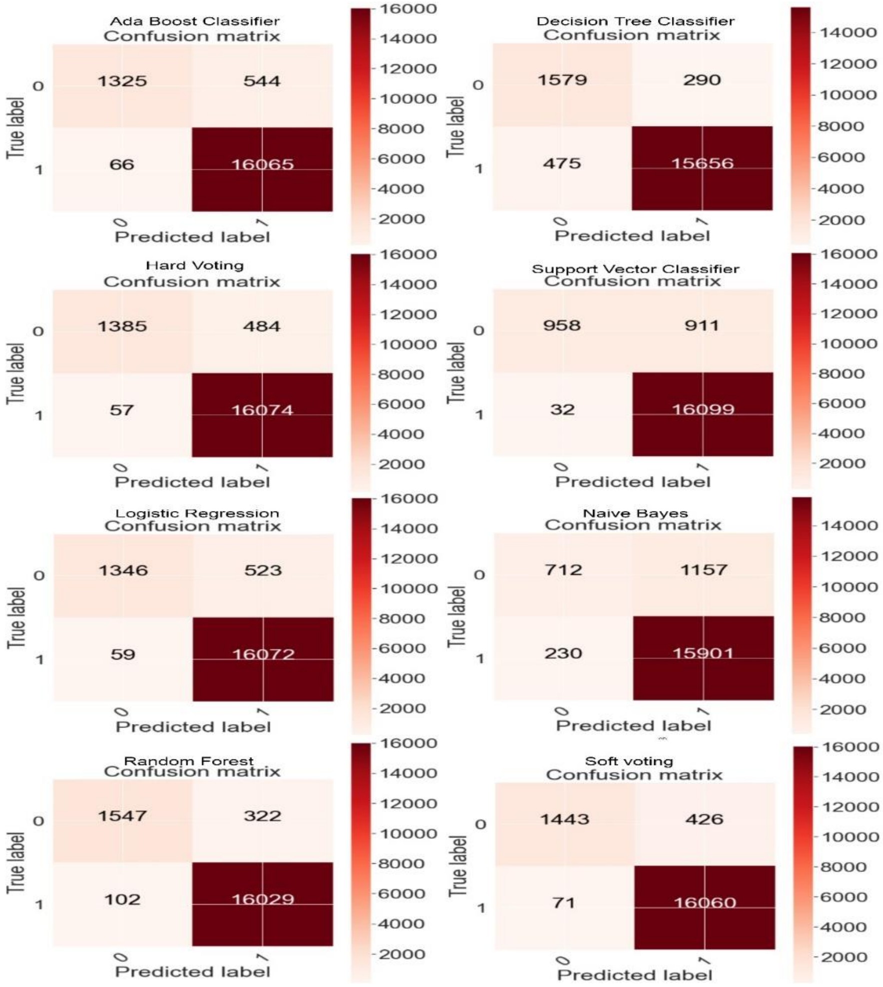
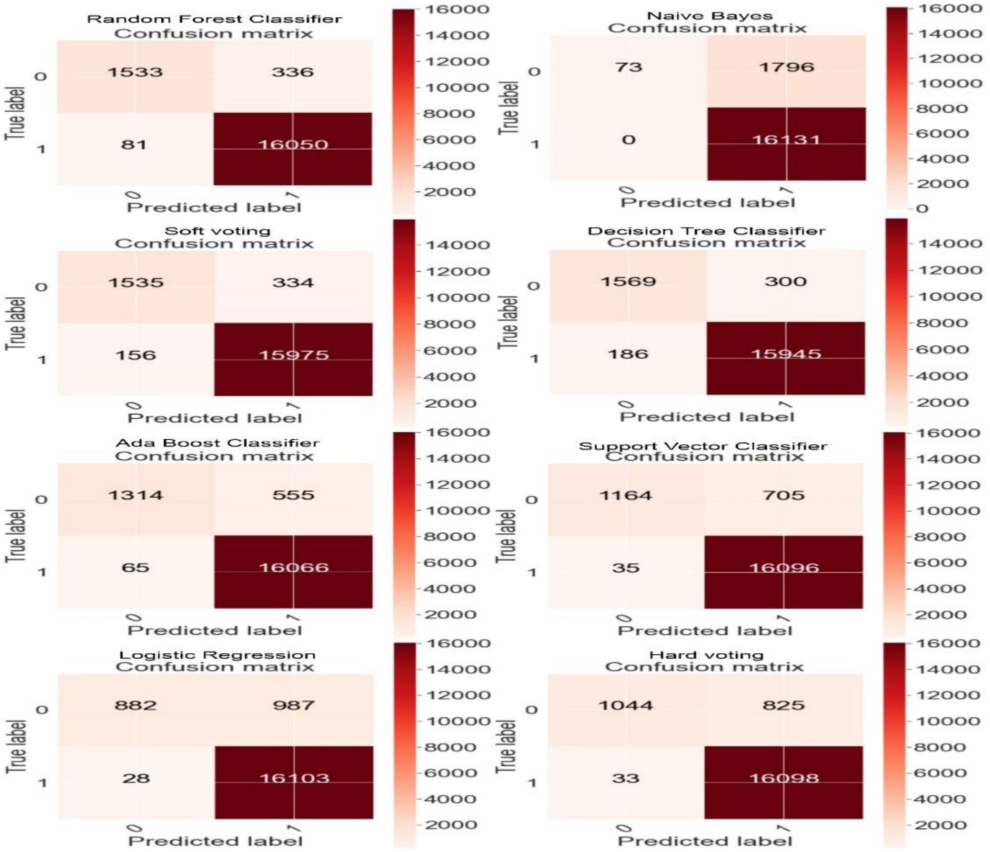


Figure 7: The results of TF feature comparing actual and predicted outcomes



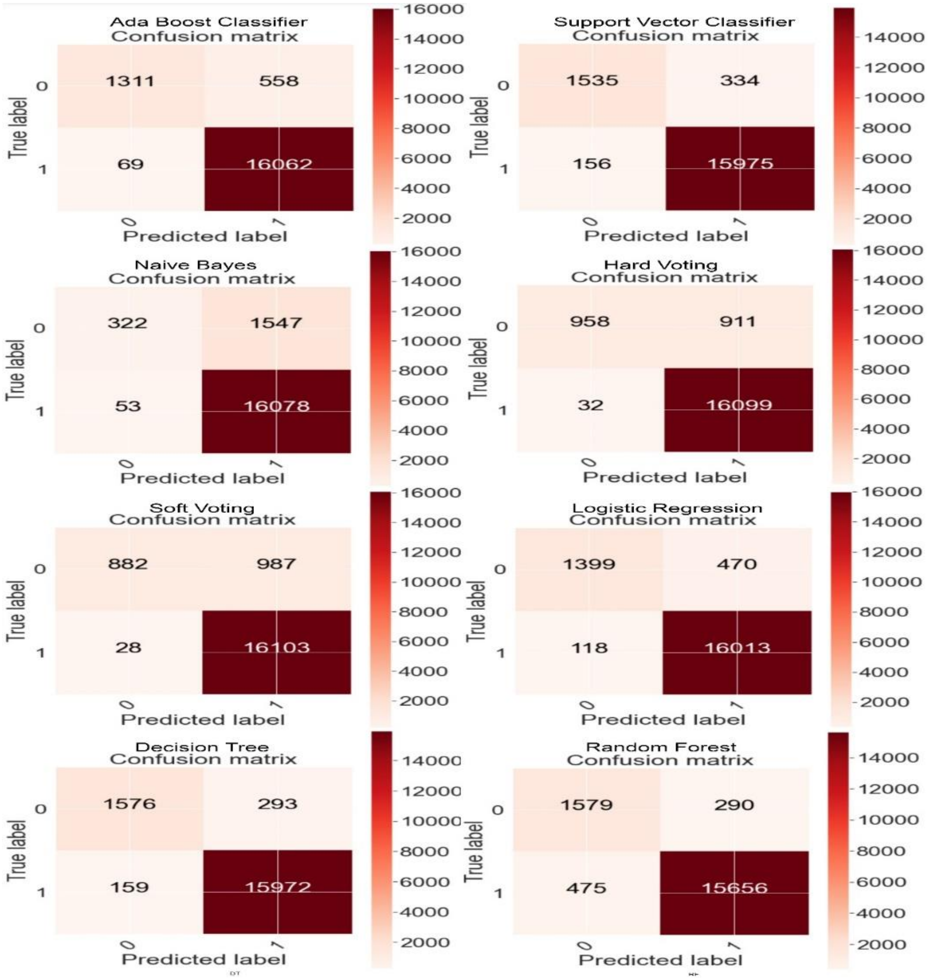
The proposed model based on the combination of three models such as LR+SVC+DT also shows 97% accuracy, 97% precision, 87% recall, and 91% f1 score results with hard voting. Gain 97% accuracy, 96% precision, 88% recall, and 92% f1 score with soft voting. Comparatively other models in Figure 4 almost show 95% above accuracy results but the highest results gain with random forest using TF features except for multinomial naive bayes.

Figure 8: The results of TF-IDF feature comparing actual and predicted outcomes



The TF-IDF features are the weighted features based on the ratio of term frequency and inverse document frequency that create the feature matrix based on the word importance in a single document over all the documents. Figure 5 presents the comparative results of the machine learning models by using TFIDF features. Still, the random forest shows highest results but improved in comparison of TF features. The random forest gain 98% accuracy, 96% precision, 91% recall and 93% f1 score. The proposed approach based on the combination of features fusion and hybrid model in the presence of grid search hyper parameter tuning that increases the efficiency of the prediction and accuracy results. Figure 6 presents the results of the proposed approach. The random forest shows the same results of 98% accuracy as before but improve the prediction results by gaining 97% precision, 97% recall and 94% f1 score.

Figure 9. The results of fused features comparing actual and predicted outcomes



The proposed approach and model gain 99% accuracy, 98% precision, 97% recall and 98% f1 score results with soft voting. The automatic hyper-parameter tuning implemented by using grid search method to improve the efficiency and carried out experiments using voting approach for the comparison of the machine learning models. The grid search automatically set the best hyper parameters for each model to get highest results then voting is applied on the model to combines them. After this, the fused features are passed to the hybrid model to perform training. The fused features increase the importance of the features and proposed model increases the accuracy results. Figure 6 shows the comparative analysis of the proposed hybrid model using soft voting with the rest of the machine learning models.

The proposed approach and model gain 99% accuracy, 98% precision, 97% recall and 98% f1 score results with soft voting. The automatic hyper-parameter tuning implemented by using grid search method to improve the efficiency and carried out experiments using voting approach for the comparison of the machine learning models. The grid search automatically set the best hyper parameters for each model to get highest results then voting is applied on the model to combines them. After this, the fused features are passed to the hybrid model to perform training. The fused features increase the importance of the features and proposed model increases the accuracy results. Figure 6 shows the comparative analysis of the proposed hybrid model using soft voting with the rest of the machine learning models.

The comparative analysis was also studied based on the confusion matrixes of the TF, TFIDF, and Fused features using machine learning models and the proposed model. Figures 7, 8, and 9 are the confusion matrixes that are created based on thetrue labels and predicted labels. The 0 presents the extreme contents and 1 presents normal contents. The confusion matrix presented the correctly classified results and false classified results.

5) CONCLUSION

The proposed study is considered to be more helpful for digital profiling and digital policing. Data about crimes and criminalsare already stored in databases for future prediction. This data is more important for law enforcement agencies and think tanksof a country. The world is going towards digitization and making more tools and developing new techniques for data analysisand implementation. Data from social media is now used for the prediction of the future of the country. The proposed approach is based on the state of art structure of preprocessing, feature fusion, and grid search-based implementation of a hybrid model that shows the highest accuracy results of 99% with the Twitter dataset. A comparative analysis has been done in this researchto evaluate the proposed approach in a scientific manner. The aims of determination of extremism in social media text data using Twitter dataset are successfully achieved and prevention of crime and terrorism will be controlled in a real-life scenario.

References

[Ahmad et al., 2019]Ahmad, S., Asghar, M. Z., Alotaibi, F. M., and Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Computing and Information Sciences, 9(1):1–23.

[Akhter et al., 2018]Akhter, N., Zhao, L., Arias, D., Rangwala, H., and Ra- makrishnan, N. (2018). Forecasting gang homicides with multi-level multi- task learning. In International Conference on Social Computing, Behavioral- Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pages 28–37. Springer.

[Aldera et al., 2021]Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., and Alothaim, A. (2021). Exploratory data analysis and classification of a new arabic online extremism dataset. IEEE Access, 9:161613–161626.

[Alghamdi and Selamat, 2022]Alghamdi, H. and Selamat, A. (2022). Tech- niques to detect terrorists/extremists on the dark web: a review. Data Technologies and Applications.

[Andleeb et al., 2019]Andleeb, S., Ahmed, R., Ahmed, Z., and Kanwal, M. (2019). Identification and classification of cybercrimes using text mining technique. In 2019 International Conference on Frontiers of Information Technology (FIT), pages 227–2275. IEEE.

[Araque and Iglesias, 2020]Araque, O. and Iglesias, C. A. (2020). An ap- proach for radicalization detection based on emotion signals and semantic similarity. IEEE Access, 8:17877–17891.

[Araque and Iglesias, 2022]Araque, O. and Iglesias, C. A. (2022). An ensem- ble method for radicalization and hate speech detection online empowered by sentic computing. Cognitive Computation, 14(1):48–61.

[A.R.Zaidi, 2021]A.R.Zaidi (2021). Detecting suspicious communi- cation. https://www.kaggle.com/syedabbasraza/detecting-suspicious- communication/data.

[Asif et al., 2020]Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., and Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. Telematics and Informatics, 48:101345.

[Bisong, 2019]Bisong, E. (2019). Matplotlib and seaborn. In Building machine learning and deep learning models on google cloud platform, pages 151–165. Springer.

[Chaparro et al., 2021]Chaparro, L. F., Pulido, C., Rudas, J., Victorino, J., Reyes, A. M., Estrada, C., Narvaez, L. A., and Gómez, F. (2021). Quantifying perception of security through social media and its relationship with crime. IEEE Access, 9:139201–139213.

[de Carvalho et al.,]de Carvalho, V. D. H., Costa, A. P. C. S., et al. Exploring text mining and analytics for applications in public security: An in-depth dive into a systematic literature review.

[Feizollah et al., 2019]Feizollah, A., Ainin, S., Anuar, N. B., Abdullah, N. A. B., and Hazim, M. (2019). Halal products on twitter: Data extraction and sentiment analysis using stack of deep learning algorithms. IEEE Access, 7:83354–83362.

[Fraiwan, 2022] Fraiwan, M. (2022). Identification of markers and artificial intelligence-based classification of radical twitter data. Applied Computing and Informatics.

[Gaikwad et al., 2021a]Gaikwad, M., Ahirrao, S., Phansalkar, S., and Kotecha, K. (2021a). Multi-ideology isis/jihadist white supremacist (miws) dataset for multi-class extremism text classification. Data, 6(11):117.

[Gaikwad et al., 2021b]Gaikwad, M., Ahirrao, S., Phansalkar, S., and Kotecha, K. (2021b). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. IEEE Access, 9:48364–48404.

[Hunter, 2007]Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. Computing in science & engineering, 9(03):90–95.

[Iqbal et al., 2021]Iqbal, F., Batool, R., Fung, B. C., Aleem, S., Abbasi, A., and Javed, A. R. (2021). Toward tweet-mining framework for extracting terrorist attack-related information and reporting. IEEE Access, 9:115535–115547.

[Iqbal et al., 2019]Iqbal, F., Hashmi, J. M., Fung, B. C., Batool, R., Khattak, A. M., Aleem, S., and Hung, P. C. (2019). A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. IEEE Access, 7:14637–14652.

[Johnson et al., 2020]Johnson, N., Turnbull, B., Maher, T., and Reisslein, M. (2020). Semantically modeling cyber influence campaigns (cics): ontology model and case studies. IEEE Access, 9:9365–9382.

[Karami et al., 2020]Karami, A., Lundy, M., Webb, F., and Dwivedi, Y. K. (2020). Twitter and research: a systematic literature review through text mining. IEEE Access, 8:67698–67717.

[Kathuria et al., 2019]Kathuria, R. S., Gautam, S., Singh, A., Khatri, S., and Yadav, N. (2019). Real time sentiment analysis on twitter data using deep learning (keras). In 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pages 69–73. IEEE.

[Khan et al., 2019]Khan, D. M., Rao, T. A., and Shahzad, F. (2019). The classification of customers' sentiment using data mining approaches. Global Social Sciences Review, 4:146–156.

[Lal et al., 2020]Lal, S., Tiwari, L., Ranjan, R., Verma, A., Sardana, N., and Mourya, R. (2020). Analysis and classification of crime tweets. Procedia computer science, 167:1911–1919.

[Lee et al., 2022]Lee, E., Rustam, F., Washington, P. B., El Barakaz, F., Al- jedaani, W., and Ashraf, I. (2022). Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. IEEE Access, 10:9717–9728.

[Loper and Bird, 2002]Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. arXiv preprint cs/0205028.

[Mahesh, 2020]Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9:381–386.

[Mashechkin et al., 2019]Mashechkin, I. V., Petrovskiy, M., Tsarev, D. V., and Chikunov, M. N. (2019). Machine learning methods for detecting and monitoring extremist information on the internet. Programming and Computer Software, 45(3):99–115.

[Meliana et al., 2019]Meliana, N., Fadlil, A., et al. (2019). Identification of cyber bullying by using clustering methods on social media twitter. In Journal of Physics: Conference Series, volume 1373, page 012040. IOP Publishing.

[Mittal and Patidar, 2019]Mittal, A. and Patidar, S. (2019). Sentiment analysis on twitter data: A survey. In Proceedings of the 2019 7th International Conference on Computer and Communications Management, pages 91–95.

[Mouhssine and Khalid, 2018]Mouhssine, E. and Khalid, C. (2018). Social big data mining framework for extremist content detection in social networks. In 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), pages 1–5. IEEE.

[Oliphant, 2006]Oliphant, T. E. (2006). A guide to NumPy, volume 1. Trelgol Publishing USA.

[Pedregosa et al., 2011]Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. The Journal of machine learning research, 12:2825–2830.

[Razzaq et al., 2019]Razzaq, A., Asim, M., Ali, Z., Qadri, S., Mumtaz, I., Khan, D. M., and Niaz, Q. (2019). Text sentiment analysis using frequency- based vigorous features. China Communications, 16(12):145–153.

[Roy et al., 2020]Roy, P. K., Tripathy, A. K., Das, T. K., and GAO, X.-Z. (2020). A framework for hate speech detection using deep convolutional neural network. IEEE Access, 8:204951–204962.

[Salloum et al., 2017]Salloum, S. A., Al-Emran, M., Monem, A. A., and Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J, 2(1):127–133.

[Sangher and Singh, 2019]Sangher, K. S. and Singh, A. (2019). A systematic review–intrusion detection algorithms optimisation for network forensic analysis and investigation. In 2019 International Conference on Automation, Computational and Technology Management (ICACTM), pages 132–136. IEEE.

[Sharif et al., 2019]Sharif, W., Mumtaz, S., Shafiq, Z., Riaz, O., Ali, T., Hus- nain, M., and Choi, G. S. (2019). An empirical approach for extreme behavior identification through tweets using machine learning. Applied Sciences, 9(18):3723.

[Shoeibi et al., 2021]Shoeibi, N., Shoeibi, N., Hernández, G., Chamoso, P., and Corchado, J. M. (2021). Ai-crime hunter: An ai mixture of experts for crime discovery on twitter. Electronics, 10(24):3081.

[Torregrosa et al., 2021]Torregrosa, J., Bello-Orgaz, G., Martinez-Camara, E., Del Ser, J., and Camacho, D. (2021). A survey on extremism analysis using natural language processing. arXiv preprint arXiv:2104.04069.

[Tundis et al., 2018]Tundis, A., Bhatia, G., Jain, A., and Mühlhäuser, M. (2018). Supporting the identification and the assessment of suspicious users on twitter social media. In 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), pages 1–10. IEEE.

[Tundis et al., 2019]Tundis, A., Jain, A., Bhatia, G., and Muhlhauser, M. (2019). Similarity analysis of criminals on social networks: An example on twitter. In 2019 28th International Conference on Computer Communication and Networks (ICCCN), pages 1–9. IEEE.

[Yu et al., 2019]Yu, D., Xu, D., Wang, D., and Ni, Z. (2019). Hierarchical topic modeling of twitter data for online analytical processing. IEEE Access, 7:12373–12385.

[Yu et al., 2020]Yu, H., Hu, Y., and Shi, P. (2020). A prediction method of peak time popularity based on twitter hashtags. IEEE Access, 8:61453–61461