Computers, Materials & Continua
DOI:10.32604/cmc.2022.019189
Article

Applying Machine Learning Techniques for Religious Extremism Detection on Online User Contents

Shynar Mussiraliyeva[1], Batyrkhan Omarov[1,*], Paul Yoo[1,2] and Milana Bolatbek[1]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]CSIS, Birkbeck College, University of London, London, UK
[*]Corresponding Author: Batyrkhan Omarov. Email: batyahan@gmail.com

Received: 05 April 2021; Accepted: 09 May 2021

Abstract: In this research paper, we propose a corpus for the task of detecting religious extremism in social networks and open sources and compare various machine learning algorithms for the binary classification problem using a previously created corpus, thereby checking whether it is possible to detect extremist messages in the Kazakh language. To do this, the authors trained models using six classic machine-learning algorithms such as Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbors, Naive Bayes, and Logistic Regression. To increase the accuracy of detecting extremist texts, we used various characteristics such as Statistical Features, TF-IDF, POS, LIWC, and applied oversampling and undersampling techniques to handle imbalanced data. As a result, we achieved 98% accuracy in detecting religious extremism in Kazakh texts for the collected dataset. Testing the developed machine learning models in various databases that are often found in everyday life “Jokes”, “News”, “Toxic content”, “Spam”, “Advertising” has also shown high rates of extremism detection.

Keywords: Extremism; religious extremism; machine learning; social media; social network; natural language processing; NLP

1 Introduction

Over the past fifty years, the ideologies of extremism, radicalism, and terrorism have clearly increased, as evidenced by the rapid increase in the number of terrorist incidents worldwide and the severity of deaths associated with each incident, as shown by the global terrorism database (GTD) [1]. Unfortunately, the number of terrorist attacks against countries of the Organization for economic cooperation and development (OECD) in 2015 was the highest since 2000. It was the second-worst year in terms of the number of deaths after 2001, as reported in the Global Terrorism index [2]. In 2015 alone, there were more than eleven thousand terrorist attacks globally, as a result of which more than twenty-eight thousand people were killed [3]. While in 2016, a study by Peace Tech Lab showed that 1,441 terrorist attacks occurred worldwide with more than fourteen thousand deaths [4]. While in the first half of 2017, the number of terrorist attacks reached 520, as a result of which, according to the map of terrorist incidents, 3565 people were killed, and that year, terrorism was responsible for 0.05% of global deaths [5].

Since 2014, the threat of lone-wolf attacks has increased significantly. This was facilitated by the call of the Islamic State of Iraq and Syria (ISIL/ISIL/DAESH) to its supporters on September 22, 2014, to carry out terrorist attacks on countries participating in or supporting the global coalition against DAESH, including many OECD countries. As a result, the United States was heavily targeted by ISIS-inspired attacks, with almost a third of all attacks targeting OECD countries from 2014 to mid-2016 occurring in the United States [6]. A study by the Institute for Economics and Peace found that religious extremism has increased dramatically since 2000 and embodying the leading ideology of terrorism in the Middle East and North Africa, sub-Saharan Africa, and South Asia [7].

Although the terms “radicalism” and “terrorism” are widely used, they remain poorly defined and are often confused because they are exonyms by nature [8]. In this context, the following definition of violent extremism is accepted as “encouraging, condoning, justifying or supporting the Commission of a violent act to achieve political, ideological, religious, social or economic goals”. In comparison, the term radicalism is defined as “the process of developing extremist ideologies and beliefs” [9,10]. On the other hand, Islamist radicalism is defined as “a militant methodology practiced by Sunni Salafi Islamists who seek the immediate overthrow of existing regimes and the non-Muslim geopolitical forces that support them in order paving the way for an Islamist society that will develop through military force” [11,12].

The use of technologies such as artificial intelligence, machine learning, and data mining in the fight against terrorism, radicalism, and violent extremism, especially in social networks, has attracted the attention of researchers over the past seventeen years [13–16]. Thus, intelligence and security Informatics has become a trending interdisciplinary field of research where advanced information technologies, systems, algorithms, and databases are studied, developed, and developed for international, national, and domestic security-related applications [17]. Several universities are working with local and national security agencies to establish research centers for the study of terrorism. Prominent examples of such institutions are the Chicago security and threat project (CPOST), based at the University of Chicago [18,19], and the national consortium for the study of terrorism and responses to terrorism, which is the center of excellence of the US Department of homeland security, based at the University of Maryland [20,21].

In this article, we explore the problem of detecting religious extremist thoughts and calls for extremism in online social sites, focusing on understanding and detecting extremist thoughts in online user content. We conduct a thorough analysis of content, language preferences, and topic descriptions to understand extremist appeals from a data mining perspective. Six different sets of informative features were identified, and several training algorithms were compared to identify extremist thoughts in the data. This is a new application of automatic detection of religious extremism in content with a combination of our proposed effective feature design and classification models.

This article makes a notable contribution and innovation to the literature in the following ways:

(1) Application of knowledge detection and data mining to detect the specific nature of religious extremism and calls to commit extremist acts in online user content. Previous work in this area has been done by psychological experts with statistical analysis; this approach reveals knowledge about extremist ideas in data analysis.

(2) Data corpus and platform: this article presents the Vkontakte social network and collects a new set of data for detecting extremist messages and calls to extremism. We used the Vkontakte social network [22] as it is the most popular social network among Kazakhstani youth [23]. Fig. 1 illustrates the results of surveys about utilizing social networks among young people in Kazakhstan. The data set is collected from a social network widely used in CIS countries and is classified into two categories (containing and not containing extremist messages or calls to extremism) by psychologists.

(3) Models and benchmarking: instead of using basic models with simple functions to detect extremist messages, this approach (1) identifies informative functions from various perspectives, including statistical, syntactic, linguistic, word embedding, and thematic functions; (2) chooses the best model to identify extremist texts by comparing various classifiers such as Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbors, Naive Bayes, and Logistic Regression and (3) provides benchmarks for detecting calls to extremism.

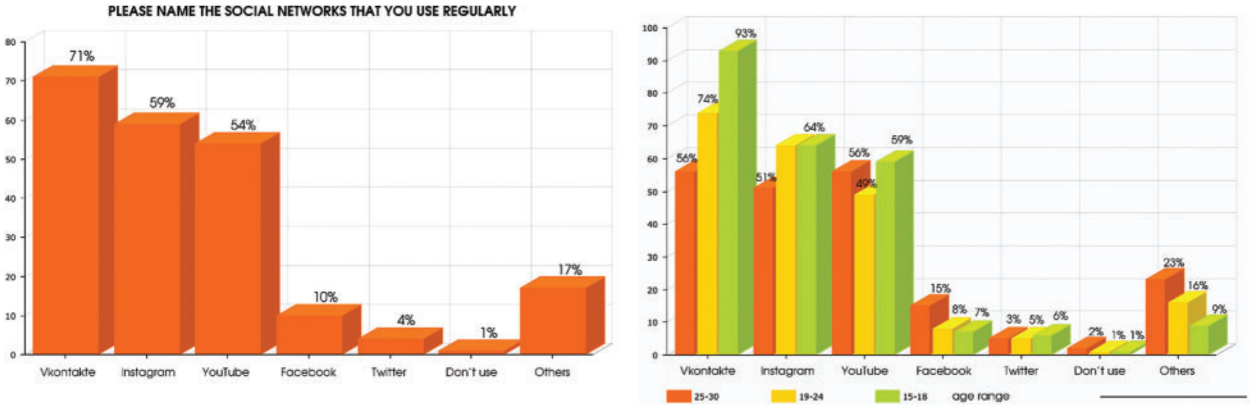

Figure 1: Trends of Kazakhstan youth: the most popular social networks among youth in Kazakhstan

The overall structure of the paper is as follows. In Section 2, we do a review on the related works. There, we tell about web-crawlers that proposed to collect, classify, and interpret the extremism information on the internet, machine learning techniques that used to identify extremism related texts, and about analyzing online user contents. Section 3 describes data collection, data annotation, data exploration, and preparation process. Section 4 describes feature extraction and text classification methods. Section 5 demonstrates the experiment results that were conducted to different algorithms and their comparison. In Section 6, we discuss opportunities in practical use and limitations of current research. In the end, we conclude and talk about the future of the research.

2 Related Works

Ashcroft et al. [24] tried to identify Jihadist messages on Twitter automatically. Within the article, researchers center on tweets which include English hashtags associated with ISIS. The authors used 3 dissimilar features such as stylometric features, temporal features and sentiment features. Be that as it may, one of the most confinements of their approach is that it is exceedingly subordinate on the information. Moreover, in [25] the researchers centered on identifying Twitter users included with "Media Mujahideen", a Jihadist bunch who disseminate purposeful content online. They utilized a machine learning approach employing a combination of data-dependent and data-independent features. The test was based on a restricted set of Twitter accounts, making it troublesome to generalize the outcomes to a more complex and reasonable scenario.

In [26], the authors proposed to apply LSTM-CNN model, which works as follows: (i) CNN model is applied for feature extraction, and (ii) LSTM model receives input from the CNN model and retains a sequential correlation by taking into account the previous data for capturing the global dependencies of a sentence in the document concerning tweet classification into extremist and non-extremist. Authors experimented with multiple Machine Learning classifiers such as Random Forest, Support Vector Machine, KN-Neighbors, Naive Bayes, and deep learning classifiers.

In [27], a sentiment analysis tool and a decision tree are used to differentiate pro-extremist web pages from anti-extremist pages, news pages, and pages that did not relate to extremism.

The novelty of the research [28] is to improve the algorithm of naive Bayes on detecting a sentiment that leads to terrorism on Twitter. To increase the accuracy, user behavioral analysis has been proposed to embed into the algorithm after the sentiment classification process has been done.

In [29], the authors searched for lexical, psycholinguistic and semantic features that allow automatic detection of extremist texts. The researchers performed morphological analysis, syntactical analysis of the corpus, as well as semantic role labelling (SRL) and keyword extraction (noun phrases).

The work [30] points at identifying right-wing radical content Twitter profiles written in German. The authors created a bag-of-words frequency profile of all tokens used by authors in the entirety of all messages in their profile.

In [31], an Exploratory Data Analysis (EDA) using Principal Component Analysis (PCA), was performed for tweets data (having TF-IDF features) to reduce a high-dimensional data space into a low-dimensional space. Furthermore, the classification algorithms like naive Bayes, K-Nearest Neighbors, random forest, Support Vector Machine and ensemble classification methods (with bagging and boosting), etc., were applied PCA-based reduced features and with a complete set of features.

In [32], the authors made a detailed analysis of the use of affect technologies to analyze online radicalization. Influence analysis was applied to a wide range of domains, such as radical forums, radical magazines, and social networks (Twitter, Facebook and YouTube). As classifiers, in this work, both Logistic Regression and Linear SVM are considered. In this work, the SIMON method is adapted to extract radicalization detection features by using radically oriented lexicons.

Research [33] focuses on the sentimental analysis of social media multilingual (Urdu, English and Roman Urdu) textual data to discover the intensity of extremism's sentiments. The study classifies the incorporated textual views into four categories, including high extreme, low extreme, moderate, and neutral, based on their level of extremism.

In [34], a context-sensitive computational method to investigating radical content on Twitter breaks down the influence prepared into building blocks. The authors show this handle employing a combination of three relevant measurements—religion, ideology and hate—each explaining a degree of radicalization and highlighting autonomous features to render them computationally open. The paper makes three commitments to solid examination: (i) Advancement of a computational method established within the relevant measurements of religion, ideology, and hate, which reflects procedures utilized by online Islamist radical bunches; (ii) An in-depth investigation of important tweet corpora concerning these measurements to prohibit likely mislabeled users; and iii) a system for comprehension online extremism as a handle to help counterprogramming. In this paper, researchers utilize Word2Vec with skip-grams to produce contextual dimension models.

In [35], and experience and the results of collecting, analyzing, and classifying Twitter data from affiliated members of ISIS, as well as sympathizers are presented. Authors used artificial intelligence and machine learning classification algorithms to categorize the tweets, as terror-related, generic religious, and unrelated. In addition, researchers built their own crawler to download tweets from suspected ISIS accounts. Authors report the K-Nearest Neighbour classification accuracy, Bernoulli Naïve Bayes, and Support Vector Machine (One-Against-All and All-Against-All) algorithms.

It should be noted that all the above-mentioned literature contains studies to determine extremist texts in English and other languages. At the moment, the authors of the study have not been able to find any work on the definition of extremist messages in the Kazakh language.

3 Data

Before classifying texts to extremist-related or neutral, we need to define danger criteria. One solution is to prepare a set of keywords. For the definition, a set of key phrases was prepared, applied to explore data in the Vkontakte social network [22]. Referring to the indicated keywords or phrases in the text, the software package infers that the text is applicable for further study. Fig. 2 shows the entire data collection, analysis of posts, and classification of texts.

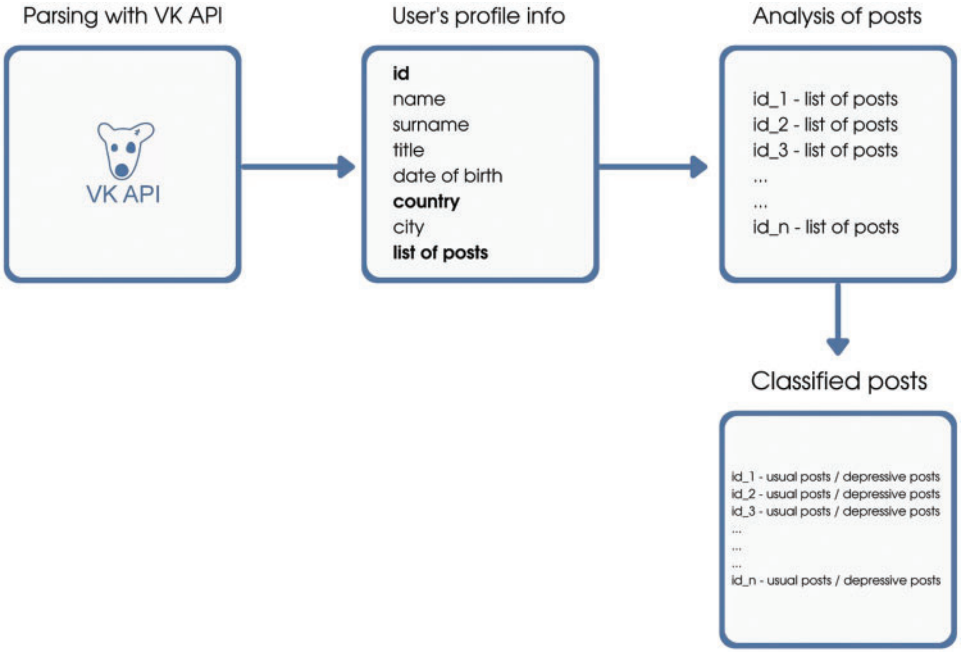

Figure 2: Scheme of data acquisition, analysis and classification of posts

The accomplishment of data acquisition may differ depending on the data source but keeps the main concept of its structure. The main goal of the part of the software responsible for data retrieval from open sources is to accomplish actions promptly and effectively. To gain high efficiency, it is necessary to use the built-in methods for receiving data from sources (API). In case of absence of such methods, then it is necessary to acquire the required data from HTTP requests.

There are three modules of the software package:

1) Information collection module is responsible for obtaining data from open sources and transmitting it for further treatment; A Python framework was built to parse data from the VK social network. We used official VK API [36] and partially parsed open accounts in Kazakhstan.
2) Keyword search module is responsible for finding keywords in a large amount of data; since we already had a list of keywords and key phrases often found in extremism related messages; we applied a linear search for words in each text, partitioning it into tokens. Keywords or key phrases for searching for possible dangerous messages were developed and approved by experts.
3) Document ranking module is responsible for identifying whether the data is related to extremism.

3.1 Information Collection Module

To collect data, we use the Vkontakte social network. Fig. 3 illustrates a schema of the data collection process. We use Python 3.6 to create a parser for data collection. Interaction with the social network API was performed using the requests library. The Pycharm Community Edition 2018 software was chosen as the development environment. To get the data, we use The VK API, a ready-made interface that allows getting the necessary information from the Vkontakte social network database using HTTPS requests. Components of the request are given in Tab. 1. Tab. 1 lists the components of a simple users query.get which as a request url looks like this ‘https://api.vk.com/method/users.get?user_id= 210700286&v= 5.92’.

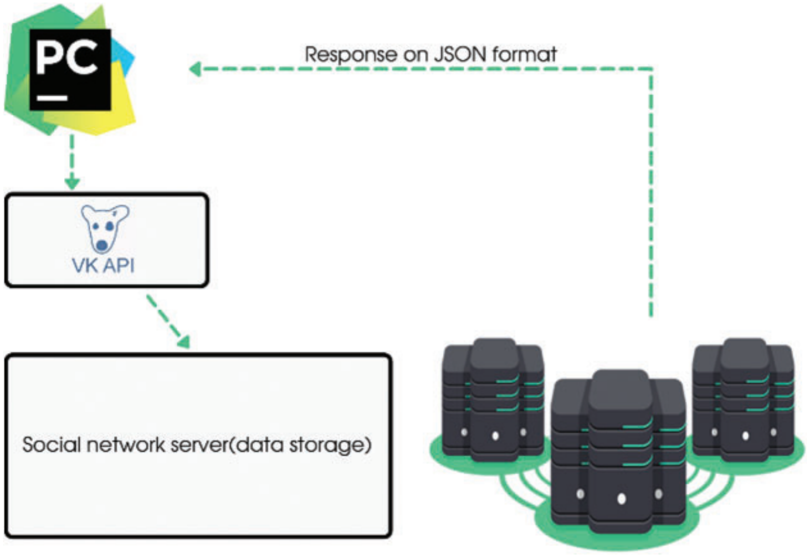

Figure 3: Data collection schema

Table 1: Query component

Query parameter	Explanation
Component	Value
https://	Connection protocol
api.vk.com/method	API service address
User.get	Name of the API Vkontakte method
?user_id = 210700286&v = 5.92	Query component

All methods in the system are divided into sections. In the transmitted request, you must pass the input data as getting parameters in the HTTP request after the method name. If the request is successfully processed, the server returns a JSON object with the requested data. The response structure for each method is strictly defined. The rules are specified on the pages describing the method in the official documentation.

To analyze the data, Python 3.7 programming language was applied with pandas, NumPy, matplotlib, plotly, bokeh, cufflinks, spacy, googletrans packages as main libraries for calculation and visualization. Full description of programming code was given in Google Colaboratory [37] notebook by the address https://colab.research.google.com/drive/1osZ0oEAgmna2OTK5gpTG4_24f3P–dsX?usp=sharing [38].

3.2 Keyword Search Module

What does "keywords confirming the possibility of defining a post as extremist" mean? There is a certain set of words that are often used by people who have decided to commit extremism or to call for extremism. In General, these words are directly related to the idea of life and death. Still, sometimes, in posts written by people who call for extremism, they try to avoid using words that directly mean their attempt at extremism. But they try to use synonyms for these same words, allowing us to find their posts using more and more new sets of keywords.

Keywords associated with extremism were identified from the previous topic. For example, kafir, kill, blow up, end, etc. These keywords will help you search for extremist posts on social networks.

As you find extremist posts, the keyword database will be updated, thereby providing a more accurate definition of extremist posts.

3.3 Document Ranking Module

Document ranking module-responsible for determining whether the information is dangerous. Word2vec vectorizer and deep learning algorithms such as Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BiLSTM) were used to rank documents by hazard level. More information about feature processing methods is given in the next sections.

Data Annotation Module. We collected the extremism ideation texts from Vkontakte social network and manually checked all the posts to ensure they were correctly labelled. Our annotation rules and examples of posts appear in Tab. 2.

Table 2: Annotation rules

Categories	Rules	Examples
Extremist text	(i) Expressing extremist thoughts	For the blessing of God, I will do jihad.
	(ii) Including potential extremist actions	I came to the Syrian land in Jihad.
Nonextremist text	(i) Formally discussing extremism	Last time, the global extremism rate was increasing.
	(ii) Referring to other's extremism	A man rams three motorbikes in what prosecutors say was an "Islamist-motivated" attack.
	(iii) Not relevant to extremism	I love this TV show and watch every week.

4 Methods and Technical Solutions

Before attributing the text to extremism related, it is necessary to define the criteria of "danger". One solution is to define a set of keywords. This method of determining the types of information was used in the developed software package. For the definition, a set of keywords was compiled, which was used to analyze information in the social network Vkontakte. Based on the presence or absence of the specified keywords in the text, the software package concludes that the text is suitable for further research. In our study, we used statistical features, parts of speech (POS), Linguistic Inquiry and Word Count (LIWC), TF-IDF word frequency features.

To understand the informativeness of these feature sets, we visualise the features on the collected corpus in 2-dimensional space by using principal component analysis (PCA) [39] in Fig. 4. From Fig. 4, we can observe a clearer separation between the two colours. This indicates that it should be easier for our classifier to separate both groups.

4.1 Classification Models

Extremism related message detection in social networks content is a standard supervised learning classification problem. Taking into account a corpus $\{x_i, y_i\}_i^n$ consisting texts $\{x_i\}_i^n$ with labels $\{y_i\}_i^n$, we developed a supervised classification models to learn the function from the training data pairs of input objects and supervisory signals [40]:

$$Y_i = F(x_i) \tag{1}$$

where $y_i = 1$ represents that x_i is "extremist intended text", $y_i = 0$ denotes "not extremist intended text." The training of the classification problem is to minimize the classification error in the training data. The prediction error is to be introduced as a loss function L(y, F(x)) where y is the real label and F(x) is the predicted label. In general terms, the goal of training is to obtain an optimal prediction model F(x) by solving below optimization task:

$$\hat{F} = \arg\min_F E_{x,y}[L(y, F(x))] \tag{2}$$

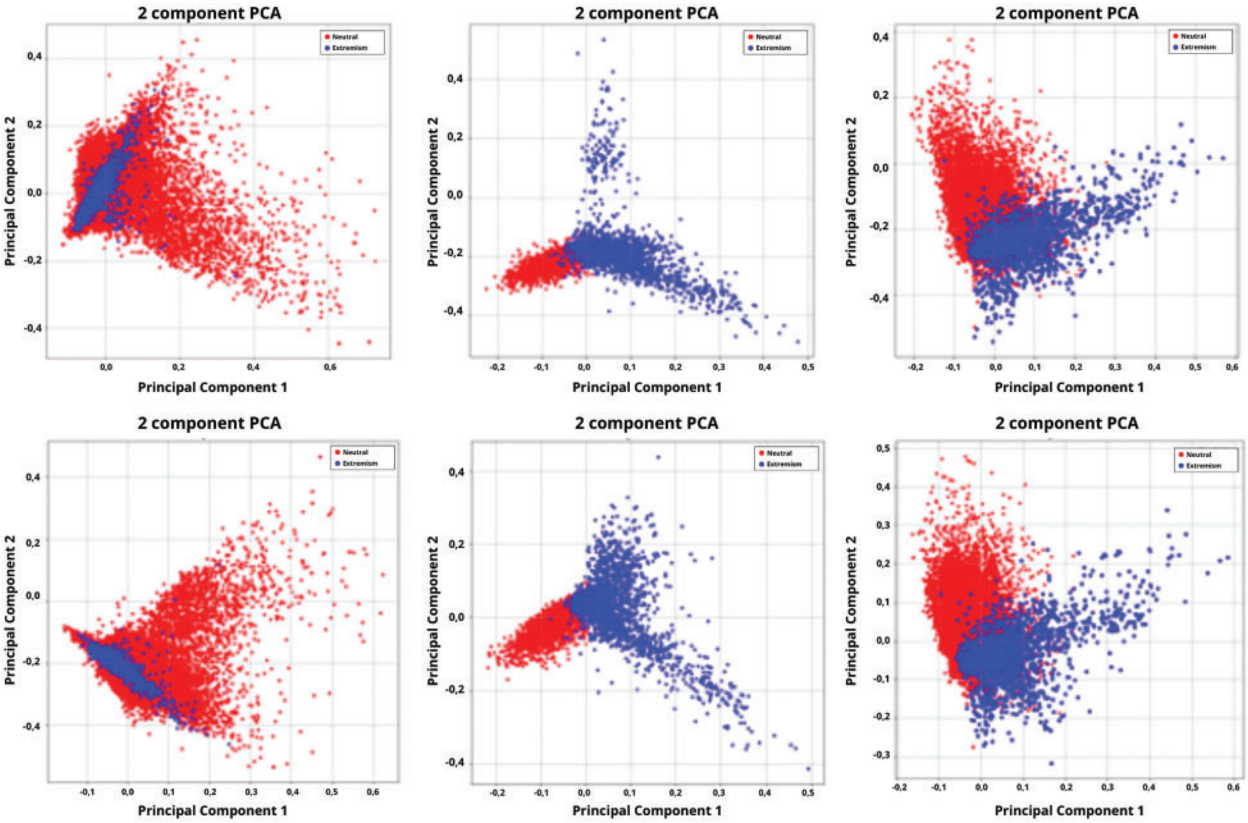

Figure 4: Visualization of extracted features using PCA

Fig. 5 demonstrates a schema of extremism related texts classification. The features include statistical features, LIWC features, POS, TF-IDF vectors, and as well as oversampling and undersampling techniques to handle imbalanced data. All extracted features were input to the classifiers.

4.2 Evaluation Method

Our task is to detect extremism related content of each of the users in the chosen data. We start performing text classification methods using the entire space of dimensional objects extracted from the data set. As basic characteristics, we utilize N-gram probabilities, LIWC categories, the LDA model, and their multiple combinations of functions based on collected training data.

Confusion matrix: this is a method for summarizing the results of classification. Accuracy alone is misleading if the number of observations is not balanced in each class. This gives an idea of our model for getting the correct one and differentiating it from the error. This clearly shows that the correct classification of a low extreme class is less, which is why its accuracy and recall work poorly.

Precision and recall: accuracy is also called positive predictive value. This is the proportion between the corresponding instances among the extracted instances. The recall is the sensitivity, and it is the proportion between the retrieved relevant instances compared to the total number of relevant instances. In classification, the accuracy is a true positive (TP) divided by the total number of labelled (TP + FP) belonging to this class. Recall that in classification, the total number of true positives (TP) is divided into instances that actually belong to the class (TP + FN).

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{6}$$

$$specificity = \frac{TN}{TTN + FN} \tag{7}$$

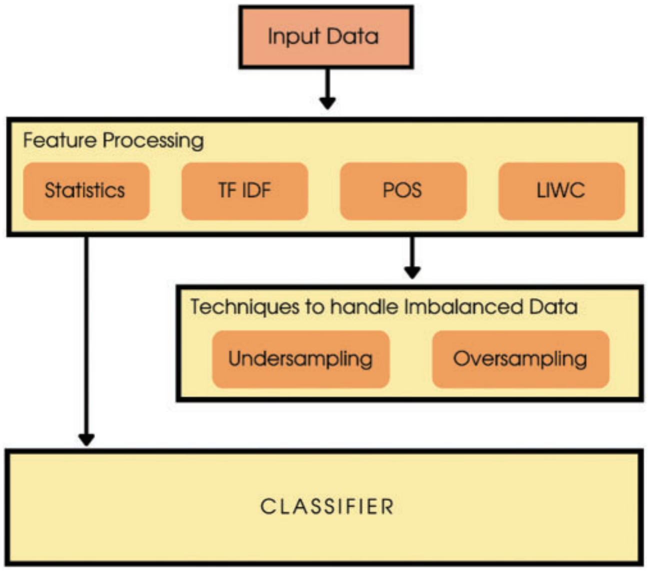

Figure 5: Classification of extremism related texts

Receiver Operating Characteristic (ROC): ROC is usually used for binary classification to study the output quality of the classifier. To find the ROC for classification with multiple labels, you must binarize the output data. One curve is drawn for each label, but each indicator is treated as a binary forecast.

5 Experiment Results and Evaluation

In this section, we compare the results of applying different machine learning algorithms for religious extremism classification using different combinations of features. In current research, we consider the following most common methods of classifier construction and training: Decision

Tree, Random Forest, Support Vector Machine, k-nearest neighbors, Logistic Regression, Naïve Bayes.

5.1 Feature Processing

In this section, we compare the results of applying different machine learning algorithms for religious extremism classification using different combinations of features. In current research, we consider the following most common methods of classifier construction and training: Decision Tree, Random Forest, Support Vector Machine, k-nearest neighbors, Logistic Regression, Naïve Bayes.

As shown in Tab. 3, the performance of all methods improves by combining more features as a whole. This observation confirms the informativity and efficiency of the acquired features. Nevertheless, the contribution of each feature varies considerably, which indicates oscillations in the outcomes of separate methods. The Support Vector Machine and Logistic Regression methods show the best productivity of the applied methods when using all groups of features as input data. Random Forest and Naïve Bayes also show good results in F1.

Table 3: Comparison of different methods using different features

Methods	features	Acc.	Prec.	Rec	F1	AUC
	Statistical features	0.5689	0.5369	0.5478	0.4658	0.5367
	Statistical features + TF-IDF	0.8204	0.2423	0.7593	0.3673	0.8622
SVM	Statistical features + TF-IDF + POS	0.8412	0.2512	0.6625	0.3643	0.8263
	Statistical features + TF-IDF + POS + LIWC	0.1065	0.0641	0.8834	0.1196	0.5357
Decision tree	Statistical features	0.5387	0.5397	0.5454	0.4894	0.5465
	Statistical features + TF-IDF	0.9444	0.9529	0.201	0.332	0.6472
	Statistical features + TF-IDF + POS	0.9444	0.8969	0.2159	0.348	0.6395
	Statistical features + TF-IDF + POS + LIWC	0.9444	0.8812	0.2208	0.3532	0.6274
	Statistical features	0.5267	0.5347	0.5446	0.4674	0.5875
	Statistical features + TF-IDF	0.9368	1.0	0.0794	0.1471	0.9179
RF	Statistical features + TF-IDF + POS	0.9369	1.0	0.0819	0.1514	0.9151
	Statistical features + TF-IDF + POS + LIWC	0.9364	1.0	0.0744	0.1386	0.914
	Statistical features	0.5311	0.5308	0.5481	0.4872	0.5634
	Statistical features + TF-IDF	0.9335	0.8421	0.0397	0.0758	0.5847
KNN	Statistical features + TF-IDF + POS	0.9354	0.8158	0.0769	0.1406	0.6105
	Statistical features + TF-IDF + POS + LIWC	0.9351	0.7037	0.0943	0.1663	0.701
Naïve Bayes	Statistical features	0.5298	0.5303	0.5496	0.4992	0.5475
	Statistical features + TF-IDF	0.9681	0.8942	0.6079	0.7238	0.9739
	Statistical features + TF-IDF + POS	0.9625	0.806	0.598	0.6866	0.9687
	Statistical features + TF-IDF + POS + LIWC	0.9543	0.7304	0.531	0.6149	0.9599
	Statistical features	0.5185	0.5364	0.5464	0.4736	0.5634
	Statistical features + TF-IDF	0.9601	0.9568	0.4392	0.602	0.9759
LR	Statistical features + TF-IDF + POS	0.9598	0.9418	0.4417	0.6014	0.9759
	Statistical features + TF-IDF + POS + LIWC	0.9409	0.6647	0.2804	0.3944	0.9336

The AUC performance measurement in each classification is the area under the receiver operating characteristic curve with all extracted features. As we noticed from the results, the AUC performance rises with the increasing of features.

The Logistic Regression method achieves the highest AUC of 0.9759. In addition to this, the majority of other methods have AUC value above 0.9. The receiver operating characteristic (ROC) curves of these methods are shown in Fig. 6.

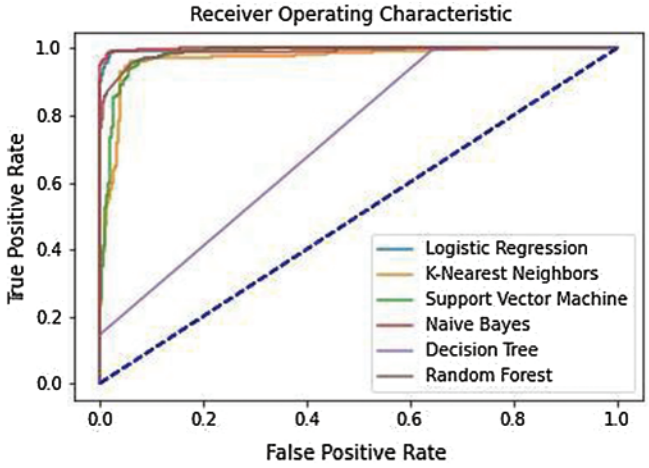

Figure 6: The receiver operating characteristic curve of six methods with all processed features

5.2 Extremism Ideas in Neutral Topics

To evaluate the extremism related text classification with other specific online communities, we expanded our corpus and tested our models in "news", "toxic content", "spam", "advertising", "jokes". The results are illustrated in Fig. 7. They show more than 90% accuracy in detecting extremism related texts from the other domains. Thus, using the features extracted using our approach was an effective way to classify reports of extremist ideas from another area.

In real world data, a class imbalance is a frequent problem, where one class contains a small number of data points, and another contains a large number of data points. In our dataset, we have met a class imbalanced problem, where 1% of all data is religious extremism related data; the other part is neutral data. In order to solve a class imbalanced problem we did experiments using oversampling and undersampling techniques.

Tab. 4 demonstrates the imbalanced classification results. The KNN method gave the best result in imbalanced classification with the maximum classification accuracy, recall, f1-score, and AUC ROC curve applying oversampling, maximum precision in undersampling. In these experiments, KNN gains better performance in accuracy, recall, f1-score, and AUC ROC than most models using oversampling and the best precision using undersampling. Fig. 8 demonstrates receiving operating characteristics for imbalanced data classification.

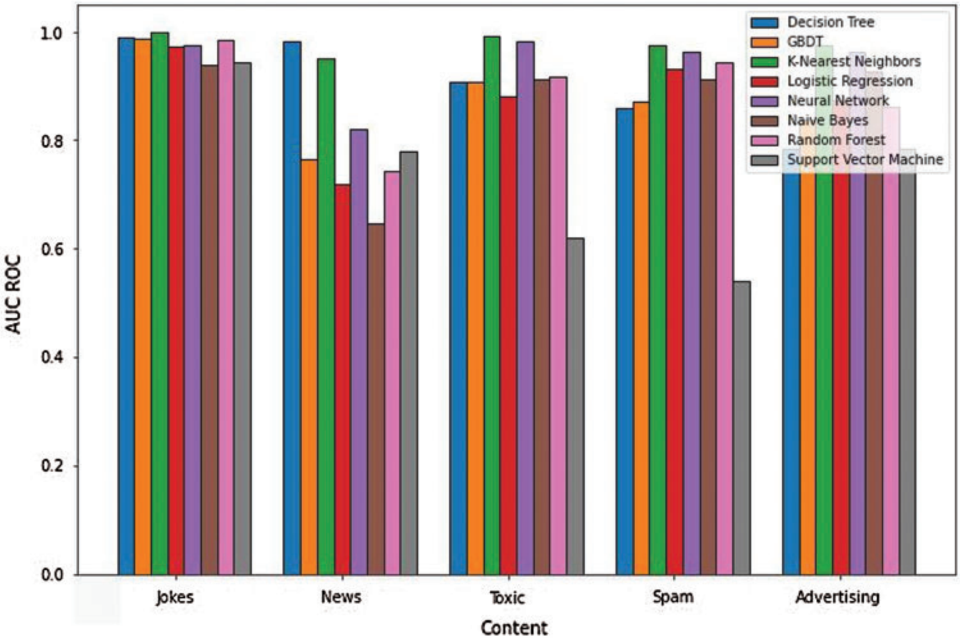

Figure 7: Classification for extremist related content *vs.* other domain texts

Table 4: Comparison of different methods using different features applying sampling techniques

Methods	Data	Features	Acc.	Prec.	Rec	F1	AUC
Support vector machine	Oversampling	Statistical features	0.5689	0.5305	0.5478	0.4658	0.5367
		Statistical features + TF-IDF	0.7727	0.6029	0.8537	0.7067	0.8878
		Statistical features + TF-IDF + POS	0.7679	0.7644	0.3997	0.5249	0.8114
		Statistical features + TF-IDF + POS + LIWC	0.3158	0.3088	0.9145	0.4617	0.5614
	Undersampling	Statistical features	0.5696	0.5374	0.5487	0.4608	0.5345
		Statistical features + TF-IDF	0.9287	0.8916	0.9757	0.9318	0.9855

(Continued)

Table 4: Continued

Methods	Data	Features	Acc.	Prec.	Rec	F1	AUC
		Statistical features + TF-IDF + POS	0.9139	0.8699	0.973	0.9186	0.9841
		Statistical features + TF-IDF + POS + LIWC	0.4253	0.4589	0.8437	0.5945	0.4836
Decision tree	Oversampling	Statistical features	0.5647	0.5317	0.5480	0.4646	0.5370
		Statistical features + TF-IDF	0.7921	0.9279	0.3816	0.5408	0.6869
		Statistical features + TF-IDF + POS	0.7861	0.9538	0.35	0.5121	0.6726
		Statistical features + TF-IDF + POS + LIWC	0.7832	0.9488	0.3427	0.5035	0.6688
	Undersampling	Statistical features	0.5648	0.5367	0.5421	0.4607	0.5350
		Statistical features + TF-IDF	0.6743	0.6056	0.9973	0.7536	0.7148
		Statistical features + TF-IDF + POS	0.6568	**1.0**	0.3127	0.4764	0.7306
		Statistical features + TF-IDF + POS + LIWC	0.6487	**1.0**	0.2965	0.4574	0.7169
Random Forest	Oversampling	Statistical features	0.5607	0.5340	0.5405	0.4605	0.5301
		Statistical features + TF-IDF	0.7978	0.9858	0.375	0.5434	0.9506
		Statistical features + TF-IDF + POS	0.8003	0.9851	0.3831	0.5517	0.9519
		Statistical features + TF-IDF + POS + LIWC	0.7994	0.986	0.3801	0.5486	0.9479

(Continued)

Table 4: Continued

Methods	Data	Features	Acc.	Prec.	Rec	F1	AUC
	Undersampling	Statistical features	0.5648	0.5325	0.5447	0.4670	0.5332
		Statistical features + TF-IDF	0.8964	0.9933	0.7978	0.8849	0.9852
		Statistical features + TF-IDF + POS	0.8816	0.9829	0.7763	0.8675	0.9819
		Statistical features + TF-IDF + POS + LIWC	0.8843	0.9831	0.7817	0.8709	0.9829
KNN	Oversampling	Statistical features	0.5674	0.5345	0.5478	0.4663	0.5324
		Statistical features + TF-IDF	**0.996**	0.9904	**0.9973**	**0.9939**	**0.9997**
		Statistical features + TF-IDF + POS	0.992	0.9781	0.9973	0.9876	0.999
		Statistical features + TF-IDF + POS + LIWC	0.9401	0.8442	0.9973	0.9144	0.9981
	Undersampling	Statistical features	0.5678	0.5302	0.5405	0.4678	0.5345
		Statistical features + TF-IDF	0.5451	**1.0**	0.0889	0.1634	0.9849
		Statistical features + TF-IDF + POS	0.5841	**1.0**	0.1671	0.2864	0.9852
		Statistical features + TF-IDF + POS + LIWC	0.821	0.8696	0.7547	0.8081	0.9151
Naïve Bayes	Oversampling	Statistical features	0.5654	0.5314	0.5450	0.4658	0.5336
		Statistical features + TF-IDF	0.9541	0.9392	0.9161	0.9275	0.9869

(Continued)

Table 4: Continued

Methods	Data	Features	Acc.	Prec.	Rec	F1	AUC
		Statistical features + TF-IDF + POS	0.9503	0.9357	0.9076	0.9214	0.9829
		Statistical features + TF-IDF + POS + LIWC	0.9149	0.9046	0.8213	0.8609	0.9677
	Undersampling	Statistical features	0.5675	0.5354	0.5470	0.4607	0.5305
		Statistical features + TF-IDF	0.9758	0.9944	0.9569	0.9753	0.9981
		Statistical features + TF-IDF + POS	0.9771	0.9972	0.9569	0.9766	0.998
		Statistical features + TF-IDF + POS + LIWC	0.8533	0.9119	0.7817	0.8418	0.9524
Logistic Regression	Oversampling	Statistical features	0.5619	0.5315	0.5463	0.4698	0.5465
		Statistical features + TF-IDF	0.9746	0.9731	0.9469	0.9598	0.9957
		Statistical features + TF-IDF + POS	0.9744	0.9716	0.948	0.9597	0.9956
		Statistical features + TF-IDF + POS + LIWC	0.9505	0.959	0.8833	0.9196	0.9874
	Undersampling	Statistical features	0.5604	0.5307	0.5487	0.4696	0.5267
		Statistical features + TF-IDF	0.9771	0.9836	0.9704	0.9769	0.9976
		Statistical features + TF-IDF + POS	0.9785	0.9837	0.973	0.9783	0.9977
		Statistical features + TF-IDF + POS + LIWC	0.9785	0.9944	0.9623	0.9781	0.9979

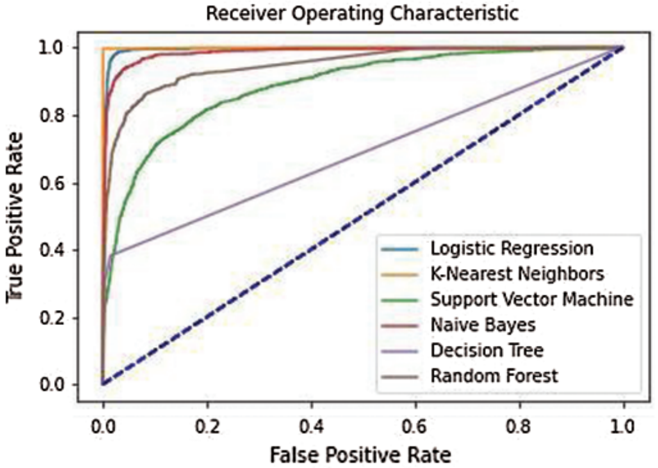

Figure 8: The ROC curve of six methods with all processed features using imbalanced classification

6 Discussion

6.1 Practical Use

Our research results demonstrate that the text-mining approach can be used to detect contents with religious extremism on the internet. As one of the most effective models, the logistic regression model and the Naïve Bayes algorithm conduct well on the given issue. The models that are applied in this research can be applied to instantly identify people with calls to extremism when they publish materials on their forum or blog entry. Because of the suitability and flexibility of the mentioned model, code for embedding in mobile applications, comments, blogs, forums add little workload. If religious extremism calls or thoughts are recognized in the pop-up window, the message can be immediately blocked.

6.2 Limitations

Firstly, the classification system in current research is limited to text messages in the Kazakh Language. Such models can be trained and tested in other languages if there is an appropriate dataset or corpus.

Secondly, our system can give a decision if an input text is religious extremism related or not. It cannot distinguish the level of hardness of extremism (as low, moderate, high extremism types). For that, it needs to create another corpus or the current corpus needs to be expanded with labelling of different levels of extremism.

Thirdly, by saying extremism detection, we can tell only about religious extremism. The other extremism types as violent, radicalization, racism, supremacism and ultranationalism, political extremism, anarchist, maoist, or single issue extremism are not considered in this research. For automatic detection of each type of extremism, it would be necessary to create a sufficient corpus that divided multiple classes and multiclassification algorithms would need to be applied.

Fourthly, our system can only claim to detect extremism texts, not a possible extremism attempt.

Fifthly, in this research, we use classical machine learning algorithms and features. In further research, we will propose our own methods to improve extremism detection rate by considering the

Kazakh language features. In the next part of our research, we are going to improve classification results by considering the uniqueness of the Kazakh Language.

By considering the relationship between religious extremism ideation and extremism facts, the acute focus should be devoted to people who use social networks to talk with thoughts of radical or extremist beliefs. The results of this research specify that these short statements have the capability to attract user's attention and cause serious anxiety. Future study may attempt to illuminate the true threat by exploring the social networks materials of those known to have committed extremist acts. Besides, carrying a prospective examination within which users give permission to having both extremist thought risk and social network posts monitored with relevant operations for adverse events would help to better understand the nature of social network behavior among those who experienced radical thoughts.

The given limitations during this study are going to be considered in the next step of our research.

7 Data Availability

The data used to confirm the results of this study is available in the Mendeley data resource at https://data.mendeley.com/datasets/h272z7xv9w/1.

8 Conclusion

The amount of text information is growing rapidly with the popularization of social networks, thus leaving many problems such as calls for extremism, suicide, and the dissemination of various information that will lead to psychological problems. Now, the prevention of these problems is the most important problem of the Internet society and it is extremely important to develop methods for automatic detection of such texts.

In this research, we studied the problem of automatic detection of religious extremism in online user content. By gathering and exploring depersonalized data from open groups and social network accounts, we implement a wealth of knowledge that can complement the understanding of religious extremism and calls to extremism. By applying machine learning, feature processing techniques to the constructed corpora, we have clearly shown that our framework can achieve high accuracy in detecting extremist ideas and calls to religious extremism from ordinary messages, thereby preventing the spread of extremism. In this paper, we deliver our knowledge in 1) understanding of extremist thoughts and calls to extremism by analyzing extremism related posts, comments, texts; 2) propose corpora to classify extremism ideation in the Kazakh language; 3) proffer machine learning methods, techniques and features in detecting extremist ideas.

Funding Statement: This work was supported by the grant "Development of models, algorithms for semantic analysis to identify extremist content in web resources and creation the tool for cyber forensics" funded by the Ministry of Digital Development, Innovations and Aerospace industry of the Republic of Kazakhstan. Grant No. IRN AP06851248. Supervisor of the project is Shynar Mussiraliyeva, email: mussiraliyevash@gmail.com.

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

References

[1] Global Terrorism Database (GTD), 2021. [Online]. Available: https://www.start.umd.edu/research-projects/global-terrorism-database-gtd.

[2] Country Reports on Terrorism, 2019. [Online]. Available: https://www.state.gov/reports/country-reports-on-terrorism-2019/.

[3] M. L. Ferreira, P. F. Graciano, S. R. Leal and M. F. D. Costa, "Night of terror in the city of light: Terrorist acts in Paris and Brazilian tourists' assessment of destination image," *Revista Brasileira de Pesquisa em Turismo,* vol. 13, no. 1, pp. 19–39, 2019.

[4] M. Al-Zewairi and G. Naymat, "Spotting the islamist radical within: Religious extremists profiling in the united state," *Procedia Computer Science,* vol. 113, pp. 162–169, 2017.

[5] H. Ritchie, J. Hasell, C. Appel and M. Roser, Terrorism. Our world in data, 2021. [Online]. Available: https://ourworldindata.org/terrorism.

[6] A. Shehabat, T. Mitew and Y. Alzoubi, "Encrypted jihad: Investigating the role of telegram app in lone wolf attacks in the west," *Journal of Strategic Security,* vol. 10, no. 3, pp. 27–53, 2017.

[7] Five key questions answered on the link between peace & religion, 2021. [Online]. Available: https://www.economicsandpeace.org/wp-content/uploads/2015/06/Peace-and-Religion-Report.pdf.

[8] D. Koehler, "How and why we should take deradicalization seriously," *Nature Human Behaviour,* vol. 1, no. 6, pp. 1–3, 2017.

[9] B. Schuurman and M. Taylor, "Reconsidering radicalization: Fanaticism and the link between ideas and violence," *Perspectives on Terrorism*, vol. 12, no. 1, pp. 3–22, 2018.

[10] R. Scrivens, S. Windisch and P. Simi, "Former extremists in radicalization and counter-radicalization research," in *Radicalization and Counter-Radicalization,* vol. 25, pp. 209–224, 2020.

[11] M. A. Adraoui, "Borders and sovereignty in islamist and jihadist thought: Past and present," *International Affairs,* vol. 93, no. 4, pp. 917–935, 2017.

[12] T. Abbas, "The symbiotic relationship between Islamophobia and radicalisation," *Critical Studies on Terrorism*, vol. 5, no. 3, pp. 345–358, 2012.

[13] Z. U. Rehman, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz *et al.,* "Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning," *Computers, Materials & Continua,* vol. 66, no. 2, pp. 1075–1090, 2021.

[14] S. Ahmad, M. Z. Asghar, F. M. Alotaibi and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," *Human-centric Computing and Information Sciences,* vol. 9, no. 1, pp. 1–23, 2019.

[15] S. Mussiraliyeva, M. Bolatbek, B. Omarov and K. Bagitova, "Detection of extremist ideation on social media using machine learning techniques," in *Int. Conf. on Computational Collective Intelligence*, Da Nang, Vietnam, pp. 743–752, 2020.

[16] E. Ferrara, "Contagion dynamics of extremist propaganda in social networks," *Information Sciences,* vol. 418, pp. 1–12, 2017.

[17] I. V. Mashechkin, M. I. Petrovskiy, D. V. Tsarev and M. N. Chikunov, "Machine learning methods for detecting and monitoring extremist information on the internet," *Programming and Computer Software,* vol. 45, no. 3, pp. 99–115, 2019.

[18] Chicago security and threat project (CPOST), 2021. [Online]. Available: https://cpost.uchicago.edu/.

[19] The Chicago Project on Security and Threats, 2021. [Online]. Available: https://www.uchicago.edu/research/center/the_chicago_project_on_security_and_threats/.

[20] The national consortium for the study of terrorism and responses to terrorism, 2021. [Online]. Available: https://start.umd.edu/.

[21] The national consortium for the study of terrorism and responses to terrorism based at the university of Maryland, a department of homeland security emeritus center of excellence, 2005. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/oup_coefactsheet_start_07192019.pdf.

[22] Vkontakte social network, 2021. [Online]. Available: https://vk.com/.

[23] Trends of Kazakhstan's youth, 2021. [Online]. Available: https://www.brif.kz/blog/?p = 3304.

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

[25] R. Torok, "Developing an explanatory model for the process of online radicalization and terrorism," *Sec. Informatics,* vol. 2, no. 6, pp. 1–10, 2013.

[26] S. Ahmad, M. Asghar, F. Alotaibi and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," *Human-centric Computing and Information Sciences,* vol. 9, no. 1, pp. 24, 2019.

[27] R. Scrivens and R. Frank, "Sentiment-based classification of radical text on the Web," in *2016 European Intelligence and Security Informatics Conf.*, Uppsala, Sweden, pp. 104–107, 2016.

[28] S. Azizan and I. Aziz, "Terrorism detection based on sentiment analysis using machine learning," *Journal of Engineering and Applied Sciences,* vol. 12, no. 3, pp. 691–698, 2017.

[29] D. Devyatkin, I. Smirnov, M. Ananyeva and M. Kobozeva, "Exploring linguistic features for extremist texts detection," in *2017 IEEE Int. Conf. on Intelligence and Security Informatics*, Attica, Greece, pp. 188–190, 2016.

[30] M. Hartung, R. Klinger, F. Schmidtke and L. Vogel, "Identifying right-wing extremism in German twitter platforms: A classification approach," in *Int. Conf. on Applications of Natural Language to Information Systems*, Liège, Belgium, pp. 320–325, 2017.

[31] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali *et al.,* "An empirical approach for extreme behavior identification through tweets using machine learning," *Applied Sciences,* vol. 9, no. 18, pp. 1–20, 2019.

[32] O. Araque and C. Iglesias, "An approach for radicalization detection based on emotion signals and semantic similarity," *IEEE Access,* vol. 8, pp. 17877–17891, 2020.

[33] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid and J. Shah, "Sentiment analysis of extremism in social media from textual information, *Telematics and Informatics,* vol. 48, pp. 101345, 2020.

[34] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan *et al.,* "Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate," in *Proc. of the ACM on Human-Computer Interaction*, Austin, Texas, USA, pp. 1–22, 2019.

[35] F. Mohammad, "Identification of markers and artificial intelligence-based classification of radical twitter data," *Applied Computing and Informatics,* vol. 16, no. 1, pp. 1–7, 2020.

[36] Description of VK API methods, 2021. [Online]. Available: https://vk.com/dev/methods.

[37] Google Colab, 2021. [Online]. Available: https://colab.research.google.com/.

[38] The program code, data, and obtained results in google colaboratory, 2021. [Online]. Available: https://colab.research.google.com/drive/1osZ0oEAgmna2OTK5gpTG4_24f3P–dsX?usp=sharing.

[39] J. Lever, M. Krzywinski and N. Altman, "Points of significance: Principal component analysis," *Nature Methods,* vol. 14, no. 7, pp. 641–642, 2017.

[40] S. Ji, C. P. Yu, S. F. Fung, S. Pan and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity,* vol. 2018, pp. 1–10, 2018.