Detection of Extremist Ideation on Social Media Using Machine Learning Techniques

Shynar Mussiraliyeva[1(✉)], Milana Bolatbek[1], Batyrkhan Omarov[1,2(✉)], and Kalamkas Bagitova[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan
mussiraliyevash@gmail.com, batyahan@gmail.com
[2] International Kazakh-Turkish University, Turkistan, Kazakhstan

Abstract. At present, the number of terrorist attacks carried out by lone terrorists under the influence of propaganda and extremist ideology, as well as by organized terrorist communities with a network and poorly connected structure, is increasing. The main means of information exchange, recruitment and promotion for such structures is the Internet, namely web resources, social networks and e-mail. In this regard, the task of detecting, identifying topics of communication, connections, as well as monitoring the behavior and forecasting of threats emanating from individual users, groups and network communities that generate and distribute terrorist and extremist information on the Internet arises.

The paper is devoted to the research and application of machine learning methods aimed at solving the problems of detecting potentially dangerous information on the Internet. The study examines the development of a corpus in Kazakh language for detecting extremist messages, and explores machine learning algorithms that used to detect content that contains calls for terrorist attacks and propaganda materials.

Keywords: Extremist ideation detection · Machine learning · Natural language processing · Classification

1 Introduction

The Internet is one of the main means of information exchange and propaganda for terrorist and extremist communities. The paper develops the proposed methods based on machine learning, using a sample search script to detect electronic messages, documents, and web resources containing extremist information, as well as users and communities in social networks that distribute such information. In this scenario, material with extremist content is available, and you need to find semantically similar materials in a social network. Using the method of semantic analysis based on orthonormal non-negative matrix factorization, the sample keywords that form search queries for the social network and the characteristic topics of the sample are highlighted. Based on orthonormal nonnegative matrix factorization, the semantic analysis method identifies the sample keywords that form search queries for the social network, and the characteristic topics

N. T. Nguyen et al. (Eds.): ICCCI 2020, LNAI 12496, pp. 743–752, 2020.
https://doi.org/10.1007/978-3-030-63007-2_58

of the sample. Search results for keywords in the social network contain a lot of "noise" – documents containing keywords, but semantically far from the original sample [1]. To filter noise, an estimate of the relevance of the found documents to the sample is calculated using a projection on the topics identified in the sample. Documents with extremist content are characterized by multilingualism, accidental and deliberate grammatical errors, deliberate distortion of semantically important words, and the presence of links and hashtags, which significantly complicates semantic analysis [2]. To solve these problems, we use n-gram representation of documents and "enrichment" of document texts (pumping out and automatically annotating information by links and hashtags and including them in the document body). The software prototype, which implements the described approaches, is applied to the analysis of real data from social networks.

2 Related Works

Over the past decade, terrorist and extremist organizations have significantly increased their presence on the Internet and social networks, actively using these tools to recruit new members and train them, prepare and organize terrorist attacks, promote violence, distribute extremist literature, etc. [3–5]. Using the Internet-a free and open resource-allows you to quickly and anonymously distribute any information, address directly to the audience of social networks and forums, without fear of censorship, present in traditional mass media. Activities aimed at identifying terrorists and related individuals, preventing the spread of extremist materials, and preventing upcoming terrorist attacks require analysis of all information received from representatives of extremist groups [6, 7]. In this context, the analysis of Internet resources comes to the fore. Due to the huge volume of information distributed over the Internet, its linguistic diversity and the requirement to monitor it in real time, it is necessary to use automatic text analysis procedures to identify potentially dangerous users, timely removal of extremist materials, and analysis of information about terrorists and upcoming terrorist attacks [8]. The main tasks in creating automatic tools for analyzing terrorist information are to select suitable data for testing algorithms and to develop algorithms that are suitable for solving the problem of detecting terrorist activity.

2.1 Development and Analysis of the Extremist Text Corpus

Chen et al. provides examples of collecting, analyzing, and visualizing publicly available terrorist materials using the so – called "shadow Internet" - a segment of the network that can only be accessed using special SOFTWARE and remain completely anonymous [9]. For the study, the authors took lists of terrorist groups (664 organizations) and their sites from US government sources, and downloaded their contents (3.6 million web pages) in English, Arabic, and Spanish. The finished corpus of extremist texts in English is described in [10]. All texts were written in Arabic and later translated into English. The case has a diverse markup (syntactic, semantic, anaphoric markup, as well as temporary markers and events), which was carried out automatically, and then checked manually.

In [11], a corpus of texts was created containing illegal texts of seven categories (terrorism, ideological texts, religious hatred, separatism, nationalism, aggression and

calls for unrest, fascism) and neutral texts with similar vocabulary. Various extensions of the standard corpus platform for studying specialized text corpora have been proposed [12, 13]. In [14], research is conducted on the use of methods for analyzing the corpus of illegal texts.

2.2 Extremist Ideation Detection

According to a study [15], the use of social networks to track the spread of radical ideas and extremist threats has attracted the attention of researchers for more than 10 years. In the last 3 years, there has been a surge in research interest in identifying and predicting the text content of messages in open social networks. The authors [16] note that Twitter is the most common data source, and various methods of information retrieval and machine learning are used for content analysis. Clustering, logistic regression, and dynamic Query Expansion are more suitable for predicting terrorist attacks, riots, or protests. A common component of various approaches and methods is named Entity Recognition (NER), which allows you to extract structured information from unstructured or semi-structured documents. To detect radicalism and extremism in real time, the K-Nearest neighbor method, the Naive Bayes classifier, the support Vector Machine (SVM) method, decision trees, Topical Crawler/Link Analysis, and others are most often used [17, 18].

In works based on the analysis of publicly available information on the Internet (Twitter, text documents of free access), one of the main tasks is to identify terrorist bandits and other terrorists. The difficulty lies in the fact that, first, communication on forums is carried out in different languages, and also, perhaps, in their combination (the same applies to documents posted on the Internet). And secondly, the fact that a simple search for keywords or specific phrases does not allow you to distinguish terrorist attacks from, for example, news agencies. In addition, terrorist sites are often disguised as news sites and religious forums. The number of sites is huge, which makes their analysis in manual mode ineffective, so for the correct identification of real sites and forums associated with certain terrorist groups, automatic means of effective selection and filtering are necessary. It is more difficult to determine whether the information being distributed belongs to one of the terrorist groups, since different terrorist groups may be ideologically close and use similar vocabulary [19].

In [20], authors investigated the possibility of creating methods for automatically detecting aggressiveness in social media texts. Identification of psycholinguistic characteristics of the text and determination of the percentage of words and phrases from the specified dictionaries is performed. In [21], the analysis of texts with extremist content is carried out, on the basis of which psychological criteria are derived, according to which the expert should evaluate the text.

In [22], decision trees to classify texts presented as graphs to classify texts. The subgraphs obtained as a result of the analysis of documents allow you to select several words, the presence of which in the text clearly determines its belonging to the terrorist site. At the same time, the absence of all these words means that the document is not exactly a terrorist document.

A similar problem, for which several different approaches are used, is considered in [23]. This is an attempt to automatically identify radical content released by jihadist groups on Twitter. To do this, we compare the results of classifying tweets into radical

and non – radical using SVM methods with linear kernel functions, AdaBoost, and the naive Bayesian classifier.

In [24], the problem of identifying tweets that promote hatred and extremism is solved as a binary classification problem using the k-neighbor and LIBSVM methods. It is shown that the classification using LIBSVM is more accurate.

Another area of research on extremist texts on the Internet is to determine the type of Internet user activity. In this work [25], the task of identifying extremist users is solved based on Twitter records, and it is also evaluated whether an ordinary user will choose extraterritorial materials and whether users will respond to contacts initiated by extremists. In this case, the analysis can be performed on aggregated data after the fact or in real-time forecast mode.

In [26], we present the Advanced Terrorist Detection System (ATDS), which is designed to track real-time access to anomalous content, which may include websites created by terrorists, by analyzing the content of information received by users via the Internet. ATDS functions in learning and recognition mode. In training mode, ATDS determines the typical interests of a pre-defined group of users by processing web pages that these users have accessed for some time. In recognition mode, ATDS monitors real-time Internet traffic generated by the controlled group, analyzes the content of web pages, and signals if the information received is not part of the group's typical range of interests and is similar to the interests of terrorists. The system analyzes arbitrary text data, which is used to determine the typical interests of users (groups of users) using the k-means clustering method.

As you can see, the development of approaches to the presentation of text information, its processing, building effective and accurate algorithms for analyzing texts, identifying their topics is an important and relevant scientific direction, which is paid much attention in the world. It should be noted that there are practically no researches devoted to the analysis of terrorist information for Kazakh language. Apparently, this is due to the lack of systematized data for testing algorithms, and the lack of expressed need for automatic processing and searching for information on the Internet (since such processing is performed manually by experts).

Thus, the development of automatic tools for thematic analysis will significantly improve the efficiency of solving problems of searching the Internet for documents and individual messages of terrorist and extremist orientation, which, in turn, will lead to the possibility of preventing upcoming terrorist attacks, reducing the influence of extremist groups and increasing the level of national security.

3 Development of the Extremist Intended Texts Corpus

Text data is necessary for analyzing what is said, thought, or felt in texts. Unfortunately, when it comes to analyzing extremist behavior, it is difficult to find a suitable selection of texts. Many document collections from social networks and media, are shared collections and should be filtered according to the research area. Because of the complexity and lack of an appropriate subject area of the corpus in the Kazakh Language we decided to create our own corpus of extremist intended texts. The corpus consists of several parts as extremist intended posts that contains 3000 words and 15 000 words with non-extremist posts, which include religious texts and texts from news portals.

In order to collect data we use Vkontakte social network that is popular in Commonwealth of Independent States. Figure 1 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 Vkontakte API that is a ready-made interface that allows to get the necessary information from the Vkontakte social network database using https requests to the server. Components of the request were given in Table 1.

Table 1. Query components.

Component	Value
https://	Connection protocol
Api.vk.com/method	Address of the API service
User.get	Name of the Vkontakte API method
?user_id = 210700286&v = 5.92	Query parameters

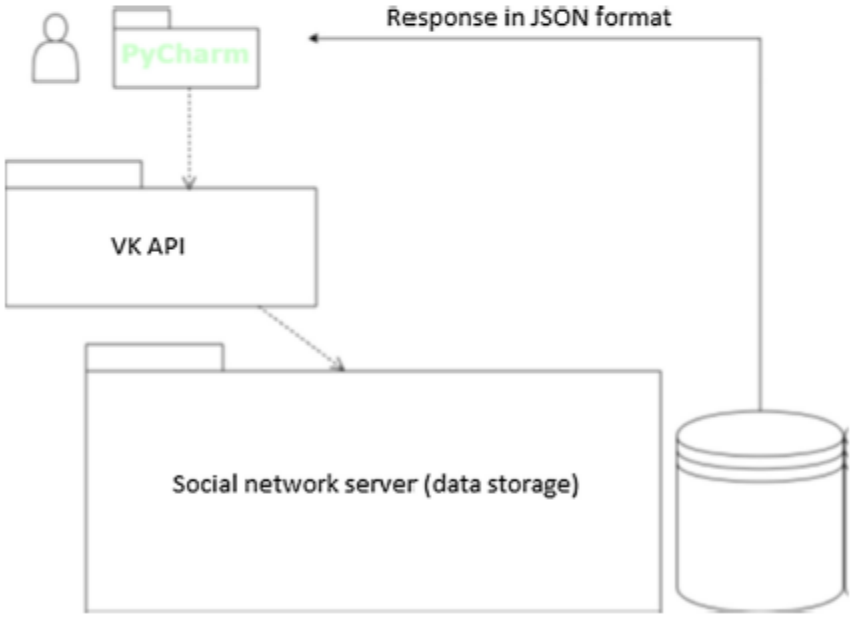

Fig. 1. Data collection schema

Methods are conditional commands that correspond to a specific database operation. For example, users.get-method for getting information about the user, account.getinfo-method for returning information about the current user, etc.

All methods in the system are divided into sections. In the transmitted request, after the method name, you must pass the input data as GET parameters in the http request. If the request is processed successfully, 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.

4 Experiment Results

In order to test the corpus, we approached the extremist text detection problem as a classification task. We performed the step-by-step process outlined in Fig. 2. We analyse and do primary pre-processing the collected data. In this step, we labeled all the texts to two classes, that class 1 means extremist behavior, and class 0 means non-extremist behavior. To preprocess the data we applied StringToWordVector that fulfills tokenization, stemming, and stop/frequent word removal.

To classify documents into two classes, we experimented with machine learning models as Gradient boosting with word2vec, Random forest with word2vec, Gradient boosting with tf-idf, and Random forest with tf-idf. The selected algorithms have demonstrated their efficiencies in various studies of text classification.

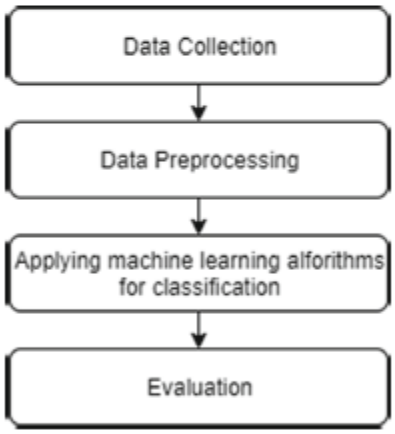

Fig. 2. Overview of the research

For research purposes, we conducted four experiments using a USB enclosure to classify emotional sentences.

Table 2 illustrates the performance of each methods that applied to identify extremist texts using the extremist texts corpus. For each method we compare accuracy, precision, recall, F1 score, and AUC to evaluate quality of corpus and performance of the algorithms. All the methods shown precision around 90%. Table 2 confirms that, the models classify extremist and non-extremist texts very good showing more that 90% accuracy. It means that, quality of the extremist texts corpus is quite good. In spite of this

Table 2. Comparison of different machine learning models on the corpus

Model	Accuracy	Precision	Recall	F1 score
Gradient boosting with word2vec	89	87	86	86
Gradient boosting with tf-idf	85	84	84	85
Random forest with word2vec	87	86	84	85
Random forest with tf-idf	83	84	83	81

result, we should complement the corpus in order to get more precision in identifying extremist texts. The experimental results illustrate that from our collected corpus, we can successfully classify extremist behavior in the texts.

Table 3. TF-IDF values of most frequently used words in the corpora

Keyword	TF-IDF value
allah (аллаһ)	25.62
jihad (жиһад)	22.62
alla (алла)	19.92
djihad (джихад)	17.1
allah (аллах)	16.72
sog'ys (соғыс)	14.3
jihad (жихад)	11.43
sogys (согыс)	8.4
ka'pir (кәпір)	7.98
tozaq (тозақ)	6.28
tozak (тозак)	5.88
kafir (кафир)	3.4

Table 3 shows the TF-IDF values of most frequently used words in the corpora. These words can be used to improve the reliability of detecting the extremist orientation in the text. In the future it is planned to assign emotional tones to the revealed words, which will later be used to create algorithms and software for analyzing the tonality of the text (sentiment analysis) [27].

For classification authors applied machine learning methods such as linear SVC, multinomial naive Bayes, logistic regression, classification trees and random forest. For this experiment authors used open source library of machine learning methods - Scikit-learn. Classification results are given in Table 4.

Table 4. Text classification results

Model	Accuracy
Linear SVC	0.61
Multinomial naive Bayes	0.81
Logistic regression	0.70
Classification trees	0.51
Random forest	0.83

5 Conclusion and Future Work

In this paper, we applied text classification techniques using natural language processing technologies for the detection of extremist behavior. To complete our task, we applied various classification algorithms.

Our experimental results show that the problem can be successfully solved. Experiments show that we can achieve high accuracy in extremist text classification using the collected corpus.

In this article, we used individual words as attributes without any additional syntactic or semantic knowledge. In the future, we plan to include information about emotions that can positively affect the accuracy of the task.

Ideally, text analysis methods are applied to cases containing thousands or even millions of documents. In this case, less than 200 records were used that can be identified with certainty as extremist behavior. Further analysis of language models will require a larger corpus. To achieve a larger corpus, we will use internal semi-automatic methods that will ensure sufficient representation of each topic in the corpus.

Using a large corpus, researchers can identify features such as the presence of emotions, cause-and-effect relationships, or language models associated with extremist behavior that can be used to teach machine learning algorithms. The main purpose of the case is to use it as an ML resource.

However, despite these limitations, the created corpus proved to be effective in training ML algorithms.

In the next step of this research we are going to supply the corpus with new texts, make balanced corpus, tonality of posts in social media, and increase the accuracy of extremist text classification.

Acknowledgements. This research has been funded by the Ministry of Digital Development, Innovations and Aerospace industry of the Republic of Kazakhstan (Grant No. AP06851248, "Development of models, algorithms for semantic analysis to identify extremist content in web resources and creation the tool for cyber forensics").

References

1. Pande, N., Karyakarte, M.: A Review for Semantic Analysis and Text Document Annotation Using Natural Language Processing Techniques. Available at SSRN 3418747 (2019)

2. Alshemali, B., Kalita, J.: Improving the reliability of deep neural networks in NLP: a review. Knowl. Based Syst. **191**, 105210 (2019)
3. Yankah, S., Adams, K.S., Grimes, L., Price, A.: Age and online social media behavior in prediction of social activism orientation. J. Soc. Media Soc. **6**(2), 56–89 (2017)
4. Costello, M., Hawdon, J.: Who are the online extremists among us? sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. Violence Gend. **5**(1), 55–60 (2018)
5. Ferrara, E.: Contagion dynamics of extremist propaganda in social networks. Inf. Sci. **418**, 1–12 (2017)
6. Awan, I.: Cyber-extremism: Isis and the power of social media. Society **54**(2), 138–149 (2017)
7. Chetty, N., Alathur, S.: Hate speech review in the context of online social networks. Aggress. Violent. Beh. **40**, 108–118 (2018)
8. Kruglanski, A., Jasko, K., Webber, D., Chernikova, M., Molinario, E.: The making of violent extremists. Rev. Gen. Psychol. **22**(1), 107–120 (2018)
9. Chen, H.: Exploring extremism and terrorism on the web: the dark web project. In: Yang, Christopher C., et al. (eds.) PAISI 2007. LNCS, vol. 4430, pp. 1–20. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71549-8_1
10. Finlayson, M.A., Halverson, J.R., Corman, S.R.: The N2 corpus: a semantically annotated collection of Islamist extremist stories. LREC, pp. 896–902 (2014)
11. Chepovskiy, A., Devyatkin, D., Smirnov, I., Ananyeva, M., Kobozeva, M., Solovyev, F.: Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). In: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017, pp. 188–190. Institute of Electrical and Electronics Engineers Inc. (2017)
12. Ménard, P.A., Barriere, C.: PACTE: a colloaborative platform for textual annotation. In: Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13) (2017)
13. Anthony, L.: Visualisation in corpus-based discourse studies, pp. 197–224. A Critical Review, Corpus Approaches to Discourse (2018)
14. Wolfe, C.R., Dandignac, M., Reyna, V.F.: A theoretically motivated method for automatically evaluating texts for gist inferences. Behav. Res. Methods **51**(6), 2419–2437 (2019). https://doi.org/10.3758/s13428-019-01284-4
15. Danekenova, A., Zhussupova, G., Nurmagambetov, R., Shunayeva, S., Popov, V.: The most used forms and methods of citizens involvement in terrorist and extremist activity. J. Pol. & L. **12**, 1 (2019)
16. Nicholls, T., Bright, J.: Understanding news story chains using information retrieval and network clustering techniques. Commun. Methods Measures **13**(1), 43–59 (2019)
17. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W.: The automated detection of racist discourse in dutch social media. Comput. Linguist. Netherlands J. **6**, 3–20 (2016)
18. Narynov, S., Mukhtarkhanuly, D., Omarov, B.: Dataset of depressive posts in Russian Language collected from social media. Data Brief **29**, 105195 (2020)
19. Ahmad, S., Asghar, M.Z., Alotaibi, F.M., Awan, I.: Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Hum. Centric Comput. Inf. Sci. **9**(1), 24 (2019)
20. Scrivens, R., Gaudette, T., Davies, G., Frank, R.: Searching for extremist content online using the dark crawler and sentiment analysis. In: Methods of Criminology and Criminal Justice Research. Sociology of Crime, Law and Deviance, vol. 24, pp. 179–194. Emerald Publishing Limited (2019)
21. Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., Shah, J.: Sentiment analysis of extremism in social media from textual information. Telematics and Informatics, p. 101345 (2020)

22. Last, M., Markov, A., Kandel, A.: Multi-lingual detection of terrorist content on the web. In: Chen, H., et al. (eds.) WISI 2006. LNCS, vol. 3917, pp. 16–30. Springer, Heidelberg (2006). https://doi.org/10.1007/11734628_3
23. Enghin Omer Using machine learning to identify jihadist messages on Twitter. http://uu.diva-portal.org/smash/get/diva2:846343/FULLTEXT01.pdf
24. Sureka, A., Agarwal, S.: Learning to classify hate and extremism promoting tweets intelligence and security. In: 2014 IEEE Joint Year Informatics Conference (JISIC), 2014, pp. 320–320 (2014). https://doi.org/10.1109/jisic.2014.65
25. Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., Galstyan, A.: Predicting online extremism, content adopters, and interaction reciprocity arXiv:1605.00659 [cs.SI] (2016)
26. Elovici, Y., et al.: Detection of access to terrorrelated Web sites using an Advanced Terror Detection System (ATDS). J. Am. Soc. Inf. Sci. **61**, 405–418 (2010). https://doi.org/10.1002/asi.21249
27. Bolatbek, M., Mussiraliyeva, S., Tukeyev, U.: Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language. J. Math. Mech. Comput. Sci. Farabi Kazakh National Univ. **1**(97), 134–142 (2018)