On Detecting Online Radicalization and Extremism Using Natural Language Processing

Shynar Mussiraliyeva
Al-Farabi Kazakh National University
Almaty, Kazakhstan

Milana Bolatbek
Al-Farabi Kazakh National University
Almaty, Kazakhstan

Batyrkhan Omarov
Al-Farabi Kazakh National University
Almaty, Kazakhstan

Zhanar Medetbek
Al-Farabi Kazakh National University
Almaty, Kazakhstan

Gulshat Baispay
Al-Farabi Kazakh National University
Almaty, Kazakhstan

Ruslan Ospanov
Al-Farabi Kazakh National University
Almaty, Kazakhstan

***Abstract*— Due to the activity of terrorist propaganda on the Internet and social networks, as well as given the high dynamics of the emergence of new sites and accounts of extremist orientation, it is important to quickly detect content that demonstrates a tendency to extremism in the prevention of extremist and terrorist activities. This article is intended to explore the possibilities of automatic recognition of extremist content using machine learning from this point of view. This article is devoted to the application of machine learning methods for solving the problem of security, in part – countering terrorism and extremism using information from the Internet.**

Keywords—extremism detection, machine learning, NLP, text processing, classification.

I. INTRODUCTION

In the modern world, there is an increasing number of terrorist acts carried out by extremist groups or individuals influenced by extremist ideas. These can be planned attacks organized by a large terrorist community or attacks by lone terrorists. However, the situation has changed significantly over the past 20 years. In the 90s, there was actually a General rather static picture [1] of what terrorist groups exist, how and by whom they are financed, what type of terrorist attacks and against which States or structures they can carry out, who is the leader or organizer of specific actions. Currently, more and more often there are lone terrorists, nothing but Internet communication, not associated with the masterminds of the terrorist attack. Often they themselves are the organizers of the attack, being under the influence of propaganda or extremist ideology, also spread over the Internet. In addition, even large terrorist and extremist organizations are moving from a vertically hierarchical to a horizontally networked or loosely connected structure in order to increase their survivability. Therefore, the role of the Internet as a means of exchanging information and spreading propaganda within terrorist and extremist communities is increasing many times [2].

Moreover, the idea of a so-called "clean state" is inherent not only to "skinheads", but also to religious extremists, who in turn call for the creation of such a "clean state" on a religious basis. It is clear that the behavior motivated by such ideas has a strict orientation, aimed in this case against persons of a different nationality or religion. This is also mixed with hatred of the existing government, which, according to extremists, condones the life of the "culprits" of all Russian troubles, which leads to an even wider spread of extremist ideas. These ideas are the Foundation for the formation of informal extremist youth groups.

The system of views imposed by extremists is attractive to young people because of the simplicity and unambiguity of their postulates, the promise of the opportunity to immediately, immediately, see the result of their aggressive actions. The need for personal participation in the complex and painstaking process of economic, political and social development is replaced by primitive calls for the complete destruction of existing foundations and their replacement with utopian projects.

Quite a lot of extremist crimes are committed by minors. Therefore, in order to prevent extremist crime and curb the criminal situation in this area, it seems appropriate to strengthen preventive work among young people, including minors, by implementing educational and preventive measures. Adolescents should be taught the basics of tolerance by organizing, for example, tolerance lessons, educational programs, and seminars on tolerance.

In this paper, we explore problem of extremist intended contents in the internet and ways to automatically identify such contents by applying machine learning and natural language processing techniques, also propose our corpus of extremist intended messages and experiments on the proposed corpus to classification of radical messages.

II. LITERATURE REVIEW

Classification of texts is one of the main tasks of computational linguistics, since it reduces a number of other tasks: determining the thematic affiliation of texts, the author of the text, the emotional color of statements, etc. In order to ensure information and public security, it is important to analyze content containing illegal information in telecommunications networks (including data related to terrorism, drug trafficking, preparation of protest movements or mass riots).

The importance of analysis of information when solving problems of counter-terrorism is currently understood on many different levels. From the published materials, three main areas of research can be identified in the field of applying machine learning methods for detecting, monitoring and predicting terrorist and extremist activity on the Internet.

1. Collecting data to be analyzed. The following types of information sources are used as analyzed data on the Internet: websites of terrorist or extremist orientation, news websites, pages of users of the social network Twitter, corporate e-mail messages.

2. Methods of text information analysis [3]. Traditional classification approaches are used, such as decision trees,

logical regression, naive Bayesian classifier, support vector method, and others. Methods are used to identify structured information from unstructured or weakly structured data, such as Named Entity Recognition (NER). To create a feature space for describing text messages, we use traditional features – keywords and frequently used phrases.

3. Research on the topology of Internet communities. This direction includes identifying key nodes, calculating their metrics (connectivity, power, and others), and building user behavior models to assess the impact of individual nodes on the community as a whole. Initially, these methods were used to solve quite "civil" tasks, such as marketing, research of gaming or consumer communities, but many of these methods have successfully found their application in the field of counter-terrorism [4].

Most of the research aimed at studying the typological features of extremist sites (linguistic and structural) and their subsequent identification is based on materials from the Dark Web Project database. Such studies are part of the work carried out within the framework of a new actively developing direction — Terrorism Informatics, which studies the phenomenon of terrorism using quantitative methods of analysis on a large data set [5]. There are three main areas of work in the field of identifying dangerous content in the Network related to linguistic analysis [6]: 1) Sentiment Analysis of Internet text, including the use of formal grammatical parameters (frequency of n-gram letters 1, POS, words, frequency of punctuation marks, service words, indices of lexical diversity, etc.), i.e., the analysis of statements for the presence of a positive/negative assessment in it; 2) affect analysis, which, unlike tonality analysis, is aimed at identifying not just a sign of appreciation, but specific emotions experienced by the author of the text (joy, anger, sadness, etc.), based on various types of language parameters (lexical, syntactic, semantic) using various machine learning methods; 3) author's analysis, aimed at identifying the author of a specific Internet text from a given circle of people. An experiment conducted on the material of English and Arabic extremist forums using different types of language parameters (lexical, morphological, syntactic, structural, semantic — a total of 301 parameters) [7] showed that the analysis of such parameters as the frequency of punctuation marks and service words is of great importance for attribution of texts in both English and Arabic. Using the entire set of parameters, high accuracy of the models was achieved. However, this study, as the authors themselves point out [8], has limitations: the experiment was conducted on a limited circle of authors (20), whereas in real life it is usually necessary to determine the author of the text from a larger number of suspects, or even the circle of suspects is not limited [9]. In addition, the authors point out the need to test their methods on the material of other languages, texts in which are presented in the Dark Web Project.

III. Review of research on identifying extremist content

The main task of detecting extremist content is to analyze the content, for which a number of solutions are proposed. Usually, authors identify a set of words that are used to distinguish terrorist sites from anti-terrorist sites. Thus, the NB (Naive Bayesian classifier) and the point-to-point mutual information (PMI) method were used to study the classification of texts in expert-created samples of Internet texts in Russian, Bashkir, and Tatar. The classes were chosen as "drugs", "violence", "nationalism", "extremism", etc. It is shown that taking into account the selected morphological features has a different effect on the quality of classification [9].

Recent foreign research has focused on studying the online behavior of extremist users, mainly through content analysis to identify distinctive text features that can help in automatically detecting these users. Recently, the identification of different traits or attitudes of Twitter users ' extremism has attracted a lot of attention from researchers, especially in connection with its role in the rise of the popularity of ISIS [10]. [11] presented an approach to detecting terrorist events using a large number of data sources using semantic graphs. A dynamically updated multilingual expandable Open Source EOS corpus was used. The algorithm uses semantic graphs constructed from sentences in corpus articles, fixing various types of relationships between sentences in the document. The online algorithm supports and updates the dynamic graph over time. Detected events are presented as sets of suggestions that are more informative than accepted representations of events. [13] proposed an automated approach to identifying right-wing extremist users on Twitter. The approach is based on the assumption that linguistic variables (vocabulary and certain semantic patterns) serve as informative predictors of the user's fundamental interests. For users, the frequency of words used (unigrams and bigrams) is analyzed.

[14] describe a network for investigating the dynamics of extremist interaction and escalation in online social networks and their relevance to criminal radicalization processes.

[15] studied online extremist communities (ACC). The goal of searching for an OEC in a large data set can be formalized as an attempt to find a relatively small subgraph in a large, annotated heterogeneous g network. The complete network G is a directed weighted graph with vertex sets V1 ... Vn. in the first stage, community optimization algorithms and knowledge about the criminal network are used to get an idea of the social network and machine learning in Phase II. Phase II retains only the peaks that match the target to the hidden community. A study [16] suggests a new search method that uses mood analysis to identify the most radical users in online forums. The method was evaluated on four Islamic forums containing about 1 million messages. The content was analyzed using Part-of-Speech tags, mood analysis, and a special algorithm.

[17] propose a new method for detecting extremist online content based on a multi-modal approach, including text (syntactic and semantic) functions, behavioral and psychological functions. Text functions include functions calculated using text mining and natural language processing techniques, such as word bag, n-grams, word frequency, and so on. The experiment was conducted on the Twitter platform, the data was collected by filtering tweets based on keywords representing top terrorist organizations, according to the US National counterterrorism center.

A number of main problems related to the identification and analysis of extremist messages are presented in [18]. In [18] when identifying extremist texts, the result is messages of illegal content found in the General flow of text information generated by social media. And the result of the analysis of illegal messages is the identified behavioral, structural and linguistic features of certain extremist groups.

In recent years, many scientists have turned to Twitter research. Analysis of messages from this social network allows you to predict changes in the exchange rate on the market, study the reaction of society to social and political events, people's attitude to new technologies, and much more. The authors [19] automatically divide tweets into "extremist" and "other" based on lexical features (the presence of religious terms, offensive and negatively colored words, etc.), using SVM and the nearest neighbor method (KNN) for binary classification. For the study, the authors collected 45 million rubles. messages from the social network Twitter, 10,487,000 of which were marked up and made up a training sample. The finished corpus of extremist texts in English is described in [20]. The volume of the corpus is 100 texts (42,480 word usage). All texts were written in Arabic and later translated into English. The corpus has a variety of markup (syntactic, semantic, anaphoric), which was carried out mathematically, and then checked manually. In addition, time markers and events are marked in the texts. In [21], a number of classifiers were used to identify Twitter messages that support the activities of extremist Islamic groups. Each network message was represented as a feature vector, in which the position of each feature depends on the frequency of its occurrence in the message. Stylometric features were used for classification: service words, frequency words, punctuation features, hashtags, bigrams on symbols and words. [22] proposes the dataset for detecting depressive posts on social media. The authors [23] set the task of detecting recruitment messages from extremist organizations and proposed a manually marked-up corpus of 192 randomly selected social media messages in English to test methods for solving this problem. We obtained a high result on the quality indicator of the binary classification AUC [24] and concluded that the proposed problem is solvable.

IV. Extremism Detection Experiment and Results

A. Data Collection

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.

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

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.

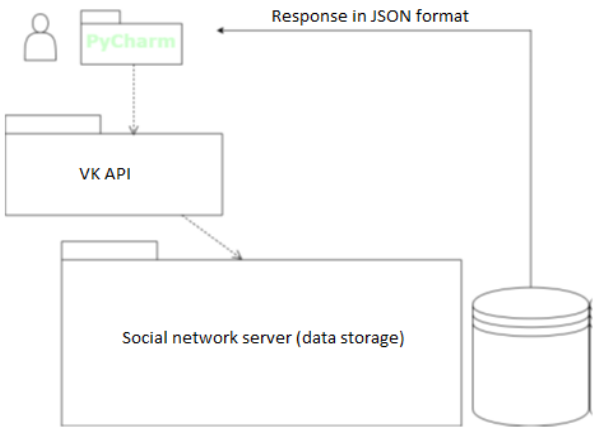

Fig. 1. Example of a figure caption.

B. Applying Machine Learning

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 Figure 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.

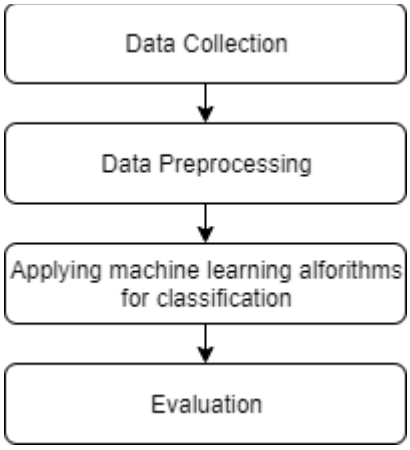

Fig. 2. Example of a figure caption.

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

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

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 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.

V. Discussion and Conclusion

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.

Classification of texts is one of the main tasks of computational linguistics, since it reduces a number of other tasks: determining the thematic affiliation of texts, the author of the text, the emotional color of statements, etc.

Formally, the task of classifying texts can be described as follows. There are many documents and many possible categories (classes). You need to build a classifier that relates the selected document to one of several predefined categories based on the content of the document. The most widespread modern approach to classification is based on machine learning methods. According to these methods, a set of rules or decision criteria for a text classifier is calculated automatically based on training data. Training data includes sample documents from each class.

The solution of the classification problem consists of four consecutive stages: preprocessing and indexing documents, reducing the dimension of the feature space, building and training a classifier using machine learning methods, and evaluating the quality of classification.

In addition to developing methods for searching the Internet for extremist content, the task of diagnosing the propensity of the author of an Internet text to extremist behavior is important, since it is well known that recruiters operating through social networks do not immediately switch to calls to join the ranks of extremists. First, they gain confidence in the victim, communicate with her on various topics to find out as much information about her as possible. In this regard, it is relevant to develop tools that would allow Web users to determine the propensity of the author of an Internet text to extremism through linguistic analysis using quantitative methods.

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.

VI. 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] Ferrara, E., Wang, W. Q., Varol, O., Flammini, A., & Galstyan, A. (2016, November). Predicting online extremism, content adopters, and interaction reciprocity. In International conference on social informatics (pp. 22-39). Springer, Cham.

[2] Nouh, M., Nurse, R. J., & Goldsmith, M. (2019, July). Understanding the radical mind: Identifying signals to detect extremist content on twitter. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 98-103). IEEE.

[3] Agarwal, S., & Sureka, A. (2015, February). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology (pp. 431-442). Springer, Cham.

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

[5] Wei, Y., & Singh, L. (2018). Detecting users who share extremist content on twitter. In Surveillance in Action (pp. 351-368). Springer, Cham.

[6] Johansson, F., Kaati, L., & Sahlgren, M. (2017). Detecting linguistic markers of violent extremism in online environments. In Artificial Intelligence: Concepts, Methodologies, Tools, and Applications (pp. 2847-2863). IGI Global.

[7] Wei, Y., & Singh, L. (2017, May). Using network flows to identify users sharing extremist content on social media. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 330-342). Springer, Cham.

[8] Fu, T., Huang, C. N., & Chen, H. (2009, June). Identification of extremist videos in online video sharing sites. In 2009 IEEE International Conference on Intelligence and Security Informatics (pp. 179-181). IEEE.

[9] Agarwal, S., & Sureka, A. (2015, March). Topic-specific YouTube crawling to detect online radicalization. In International Workshop on Databases in Networked Information Systems (pp. 133-151). Springer, Cham.

[10] Scrivens, R., Gaudette, T., Davies, G., & Frank, R. (2019). Searching for extremist content online using the dark crawler and sentiment analysis. Methods of Criminology and Criminal Justice Research (Sociology of Crime, Law and Deviance, Vol. 24), Emerald Publishing Limited, 179-194.

[11] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835.

[12] Scrivens, R., Davies, G., & Frank, R. (2018). Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. Behavioral sciences of terrorism and political aggression, 10(1), 39-59.

[13] Mashechkin, I. V., Petrovskiy, M. I., Tsarev, D. V., & Chikunov, M. N. (2019). Machine Learning Methods for Detecting and Monitoring Extremist Information on the Internet. Programming and Computer Software, 45(3), 99-115.

[14] Kursuncu, U., Gaur, M., Castillo, C., Alambo, A., Thirunarayan, K., Shalin, V., ... & Sheth, A. (2019). Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-22.

[15] Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

[16] Waseem, Z. (2016, November). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the first workshop on NLP and computational social science (pp. 138-142).

[17] Kaati, L., Omer, E., Prucha, N., & Shrestha, A. (2015, November). Detecting multipliers of jihadism on twitter. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 954-960). IEEE.

[18] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International workshop on natural language processing for social media (pp. 1-10).

[19] Myagkov, M., Kashpur, V. V., Baryshev, A. A., Goiko, V. L., & Shchekotin, E. V. (2019). Distinguishing Features of the Activity of Extreme Right Groups under Conditions of State Counteraction to Online Extremism in Russia. Region: Regional Studies of Russia, Eastern Europe, and Central Asia, 8(1), 41-74.

[20] Oussalah, M., Faroughian, F., & Kostakos, P. (2018, November). On Detecting Online Radicalization Using Natural Language Processing. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 21-27). Springer, Cham.

[21] Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels. me forum: Linguistic analysis and automatic detection. Journal of Language Aggression and Conflict, 7(2), 240-268.

[22] Narynov, S., Mukhtarkhanuly, D., & Omarov, B. (2020). Dataset of depressive posts in Russian language collected from social media. Data in brief, 29, 105195.

[23] Grover, T., & Mark, G. (2019, July). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 13, pp. 193-204).

[24] Litvinova, T., & Litvinova, O. (2019, October). Analysis and Detection of a Radical Extremist Discourse Using Stylometric Tools. In The 2018 International Conference on Digital Science (pp. 30-43). Springer, Cham.