ISBN: 978-989-8533-90-6 © 2019

EXTREMIST TEXT DETECTION IN SOCIAL WEB

Dmitry Devyatkin[1], Ivan Smirnov[1], Fyodor Solovyev[2], Margarita Suvorova[1]
and Andrey Chepovskiy[3]
[1]*Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia*
[2]*Institute of Computing for Physics and Technology, Moscow, Russia*
[3]*National Research University Higher School of Economics, Moscow, Russia*

ABSTRACT

Online or cyber extremism is one of the critical problem for the security of Russia and other countries as social web is widely used for radical activity and propaganda. This paper considers the problem of extremist text detection in Russian social media. We propose models and methods for identification of extremist text in Russian, which apply deep linguistic parsing and statistical processing of texts. We also present the dataset of terrorist, religious hate, racism and other radical texts in Russian and results of experiments on this dataset. It was shown, that low-dimensional psycholinguistic and semantic features of texts allow detecting extremist texts with quite good performance while lexical features allow recognizing topics of the detected extremist texts.

KEYWORDS

Cyber Extremism, Radical Texts Detection, Topic Identification, Psycholinguistic Features, Semantics, Extremist Lexis

1. INTRODUCTION

Due to their anonymity and quick diffusion of information social media are often misused for malicious intent. Extremist online recruitment and propaganda is one of them. Automatic detection of illegal activities on the Internet is a challenging problem for many reasons: great number of new messages per day, specificity of language for online communication and dynamic changes of content. In this paper we consider the problem of extremist text detection in the Russian-language segment of the Internet. By extremist texts we mean written materials that foment hatred toward a particular social or ethnic group, race, religion; vindicate terrorism, propagate superiority of a particular social or ethnic group, race, religion, etc. It stems from the language of Russian law "On Countering Extremist Activities".

Detection of extremist texts can be considered as a topic classification task where analyzed objects are texts form social networks and blogs. However, such an approach requires creating human annotated corpora [Cohen, K., et al, 2014] and pre-defined set of characteristics, e.g. keywords, results of complete linguistic analysis, and extracted named entities. Unfortunately, there are no such open corpora in Russian for a number of reasons.

This paper structured as follows: Section 1 contains a review of related works; Section 2 describes our methods for extremist text detection and classification; Section 3 describes a dataset of extremist text in Russian we created for the current research; and in Section 4 we provide results of experiments.

2. RELATED WORK

The problem of online extremism detection attracts attention of researchers since the early 21st century. By now there are vast amount of literature on this topic, including extensive reviews of existing solutions and methods with comprehensive analysis of their limitations, as well as trends and state-of-the-art in this field in general [Agarwal, S., & Sureka, A., 2015], [Correa, D. & Sureka, A., 2013]. Authors came to a conclusion that Link Based Bootstrapping and Text Classification are the most popular methods for extremist content detection. Text classification examines the linguistic features and makes decisions based on a machine

learning classifier. Such an approach is akin to the one we use in our research. Moreover, Internet content is dynamic and trends to change patterns in time. This is the main limitation of Link Based Bootstrapping which Text Classification does not have. H. Chen presents several studies for open-source terrorism information collection, analysis, and visualization [Chen, H., 2007]. About 3,6 million web-pages were collected from sites of terrorist groups. This dataset is available for terrorist researchers on the Dark Web Portal. J.R.Scanlon and M.S.Gerber have set themselves a task to identify the recruitment activities of violent groups [Scanlon, J.R. & Gerber, M.S., 2014]. For this purpose they annotated manually 192 post of Chen's Dark Web Project dataset. Researchers achieved an 89% area under the curve (AUC) by SVM classifier and came to a conclusion that automatic detection of online terrorist recruitment is a feasible task.

Scientific studies rely heavily on open-source models and instruments, especially when it comes to PhD students or postdocs. In the field of social media analytics there are quite a number of them [Agarwal, S., et al, 2015]. Another challenge is that extremist users can create new accounts after being suspended. To deal with this problem one has to track extremists' accounts, analyze their behavior and connections to be able to identify multiple accounts of the same user [Klausen, J., et al, 2016]. Works on statistical analysis are relevant to our research as well. Statistical analysis implies such text characteristics as average length of words and sentences, frequency of special characters, syntax errors, frequency words of size n, structure of phrases and sentences, etc. Such methods of statistical analysis are usually applied for plagiarism detection and author identification [Zurini, M., 2015]. It is worth noting that in most of the observed approaches, simple lexical or statistical attributes are used as features for classification. This can make the approaches topic-dependent reduce their capabilities in real applications.

3. METHODS FOR EXTREMIST TEXT DETECTION AND DESIGN OF EXPERIMENTAL STUDY

3.1 Detection of Extremist Texts

Instead of widely used topic-dependent lexical features we studied more complex linguistic-based psycholinguistic and semantic features. The features [Vybornova, O., et al, 2011] were extracted from Russian text corpus. The values of psycholinguistic features are estimated on the basis of morphology of lexical units of the analyzed texts. Some examples of psycholinguistic features are:

1. Verbs-adjectives ratio in the text fragment.
2. Verbs-nouns ratio in the text fragment.
3. Infinitive-verb ratio.
4. Number of the pronouns in the first person plural form.
5. Number of the pronouns in the first person singular form.
6. Number of verbs in past tense, first person, singular.
7. Number of pronouns in the third person plural form.
8. Number of pronouns in the third person singular form.
9. Number of impersonal verbs.

Values of semantic features are calculated as frequencies of semantic roles in the corpus. We used SRL for the Russian language [Lyashevskaya O., 2010]. Some examples of semantic roles are present in the Table 1.

Table 1. Examples of semantic roles

Role	Definition
Causative	Reason for action or occurrence of a feature (e.g. reward for bravery).
Destructive	The object of destructive influence (e.g. blow up the house).
Situative	External situation, natural or social factors that determine the state of the subject (e.g. to carry out reforms under conditions of economic instability).
Comitative	Concomitant action, feature, an accompanying object, or accomplice (e.g. to discuss something with colleagues)

In addition to topic-independence, the proposed approach allows to significantly reduce the feature set dimensionality. That could leads to reducing of overfitting in small datasets. For representation of psycholinguistic and semantic feature of a text t we used the following approach. Let $\Phi = \{\varphi_1,\varphi_2,...,\varphi_n\}$ be a set of all semantic and psycholinguistic features with cardinality n, $\varphi_i(t)$ be a value of a feature φ_i in a text t, where $i = 1,..,n$. Then values of psycholinguistic and semantic features for each text t can be represented as a vector:

$$t_{psy} = (\varphi_1(t),\varphi_2(t),...,\varphi_n(t)) \quad (1)$$

We created an experimental dataset, where each text is represented as a vector of values of these features (see expression 1) and tagged as extremist or neutral. Lexical features were excluded. During the experiment we evaluated the contribution of psycholinguistic and semantic features for text classification (into extremist and neutral). Classification quality was estimated by calculating F_1-score during 5-fold cross-validation. For classification we applied logistic regression, linear SVM, random forest, gradient boosting. For this experiment we used open source library of machine learning methods - Scikit-learn - where all above methods are realized [Pedregosa, F., et al., 2011].

3.2 Extremist Topic Identification

The meaning of concept of extremism is heterogeneous covering a variety of types of offences. We have developed a classification of the text collection being analyzed in our work. The classification comprises the following 7 topics: terrorism, ideological extremism, religious hatred, separatism, nationalism, aggression and calls for riots, fascism. The topics will be considered in detail in the experimental dataset description.

We propose the following framework for the problem of extremist text topic identification. Given a text corpus $C = \{C_1,..,C_K\}$ with texts falling in K topics all the texts $C_i = \{t_1^i,...,t_{C_i}^i\}$ within each topic C_i, $i = 1..K$ are concatenated into a single text T_i, $i = 1..K$. Given a text T_i a frequency dictionary $D_i = \{(x_1^i, f_1^i),...,(x^i{}_{|D_i|}, f^i{}_{|D_i|})\}$ is built, where x^i_j is a j-th feature - a fragment of text obtained from the source text and f_j^i a frequency (number) of such fragments normalized by the total number of obtained text fragments. A set of features $\{x_1^i,...,x^i_{|D_i|}\}$ of dictionary D_i is denoted as X_i. Feature may be represented as letter n-gramm, initial word forms or certain parts of speech, nominal groups, functional words. Each feature is related to single feature class F determined by the way the feature was obtained, e.g.: a class of initial word forms, a class of numerals, a class of 4-gramms, etc. A set $F_l = \{F_{l1},...,F_{l|Fl|}\}$ of considered feature classes induces a set F_l of obtained features.

$$F^{\wedge}_l = F_{l1}... \cup F_{|Fl|} \quad (2)$$

In our work we considered the following classes of features:

- Single alphabet symbols
- Sequences of symbols of lengths from 2 to 8
- Initial forms of nouns, adjectives, verbs, participles, adverbs, quantitative numerals, collective numerals, pronouns, last names, first names, middle names, vocabulary words
- Abbreviations
- Conjunctions
- Particles
- Prepositions
- Interjections
- Word stems
- Nominal and verbal groups

To filter the obtained features we used topical importance measure *Tic(w,τ),* calculated as follows [Mbaikodzhi, E., et al, 2012]:

$$idf(w,\tau) = log_2 \frac{|\tau|}{m(w,\tau)}, \quad (3)$$

$$\Delta I(w,c,\sigma) = idf(w, c \setminus \sigma) - idf(w,\sigma), \quad (4)$$

$$Tic(w,c,\sigma) = \Delta I(w,c,\sigma) H(\Delta I(w,c,\sigma)) \quad (5)$$

where *H(.)* - Heaviside function, *w* - some feature, *m(w,τ)* - count of texts in some set *τ* where *w* occurs, *c* - the experimental dataset *σ* - topical subset of the dataset. We filtered all features with zero topical importance. This way we significantly reduced the feature set dimension. In the experiment, we applied two models for features representation: a "bag of words" model and fastText word embeddings [Bojanowski, P., et al, 2016]. We used fastText because it allows creation of word embeddings for words missing in a training dataset. That is essential for social network analysis. As in the first experiment classification quality was estimated by calculating F_1-score during 5-fold cross-validation. For this experiment we also used Scikit-learn library.

4. DATASET OF EXTREMIST TEXTS

Unfortunately, there is no extremist dataset in Russian. Therefore, first of all we had to create a text corpus for the research. It includes not only extremist texts, but also topically similar texts that do not fall under the category of extremism (we call them neutral): political blogs, sites of opposition members, news, texts and posts from social networks about religion. When choosing texts for the corpus, we were guided by the law "On Countering Extremist Activities" and experts' opinion on what can be considered extremism and what cannot. The corpus includes 493 manually collected texts (650 000 words), 368 of them are extremist texts. The concept of extremism is broad and includes different types of offenses. Therefore, we classified all text on seven categories:

1. Terrorism (27 texts, 3,296 words) – materials from web-sites of outlawed in Russia organizations (the Islamic State, Hizb ut-Tahrir, etc.), where their ideology is propagated.
2. Ideological texts (26 texts, 21,131 words) affirms the superiority of some religion over others, or spreads false interpretations of sacred books.
3. Religious hatred (55 texts, 16,697 words) – texts that calls for cruelty against representatives of other religions, forming a negative image of other religions, attributing dangerous intentions to persons of another religion.
4. Separatism (7 texts, 852 words) - materials that disseminate the idea of separating some regions from the Russian Federation, and contain insults and threats against ethnic groups residing there.

5. Nationalism (208 texts, 19,399 words) - texts that affirm the initial hostility of a certain ethnic group, call for the physical destruction of its representatives, require restrictions of their rights and freedoms in Russia.
6. Aggression and calls for insurgency (43 texts, 6,757 words) - calls for unauthorized rallies and riots, deposing the government, threaten government officials and their families.
7. Fascism (13 texts, 2,059 words) - texts that justify neo-fascism and genocide, discuss forbidden fascist books, etc.

At the moment the corpus is not large enough and unbalanced: nationalism is larger than all other categories. In the future the above classification can be changed - similar categories can be combined and heterogeneous categories can be divided. Besides, we plan to expand corpus by new texts and apply the proposed method for Tatar text classification.

5. RESULTS OF EXPERIMENTS AND DISCUSSION

5.1 Results of Extremist Text Detection

We estimate F_1 binary score for the smallest class, because of the unbalanced dataset. On the basis of this experiment (Table 2) we make a conclusion that linear classification methods (e.g. linear SVM and logistic regression) do not work well with psycholinguistic and semantic features (F_1-score stands at 55%). As for more complex methods - ensembles of decision trees – they give satisfactory quality without lexical features (F_1-score reaches 76%). Psycholinguistic and semantic features do not depend on the topic of the message as much as lexical features do. Therefore, their application for extremist text detection is more promising. However, this group of features needs additional research to improve the quality of classification.

Table 2. Comparative analysis of text classification methods (F_1-score, %)

Classifier	Psycholinguistic and semantic features	Psycholinguistic and semantic features + bag of words	Psycholinguistic and semantic features + fastText
Log-regression	55±6	60±4	91±5
Linear SVM	55±3	58±6	91±3
Random forest	76±3	85±4	92±4
Gradient boosting	76±3	85±5	93±3

5.2 Results of Extremist Topic Identification

Specialized topic dictionaries of keywords and all their morphological forms have been created based on the experimental dataset. Words from Russian dictionaries fall into three levels of their topic relevance:

1. The first level is represented with commonly used words, phrases usually in negative tone. Those words frequently occur in the texts on the topic at the same time not being directly related to the topic.
2. The words of the second level occur in the texts on the given and similar topics, describing the given topic without a direct relation to it.
3. Third level words most frequently occur in the texts on the topic having the strongest relevance to the topic.

Such a division is caused by the polysemy of the grammatical morphemes in the Russian language leading to a diversity of meanings for different phrases. The following dictionaries have been developed by the present day:

1. Extremism: first level – 1249 words, second level – 766 words, third level –310 words.
2. Terrorism: first level – 1060 words, second level – 254 words, third level –322 words.
3. Nationalism, social hatred: first level – 207 words, second level – 205 words, third level – 311 words.
4. Fascism: first level – 92 words, second level – 239 words, third level – 187 words.
5. Violence, cruelty: first level – 214 words, second level – 460 words, third level – 261 words.

Results of the second experiment are presented in Table 3.The separatism topic was excluded from the table because its texts were not detected by any classifier. The experiment shows that fastText embeddings model allow significantly improve the quality of topic classification in our unbalanced and small dataset, compared with the "bag of words" model. We believe that this is due to the smaller size of the feature set space produced by the embedding model. It can also be related to the property of embedding models to generate similar vectors for semantically similar words. Unlike the previous task in this experiment, linear classification methods reach better results than ensembles. This indicates that in this task we should rather continue to replenish and balance our experimental dataset than to complicate the models of machine learning that we used.

Table 3. Results of extremist topic identification using lexical features

	"Bag of words" model, F_1-score, %				fastText model, F_1-score, %			
	SVM	Log-regression	Random-forest	Gradient-boosting	SVM	Log-regression	Random-forest	Gradient-boosting
Aggression	_	_	_	_	56±2	57±4	35±2	47±7
Fascism	_	_	_	_	36±8	34±7	_	_
Ideological texts	33±5	35±2	32±4	25±2	34±5	33±3	30±3	27±5
Nationalism	86±9	86±1	86±1	85±1	85±6	86±1	83±1	83±1
Religious hatred	37±3	32±3	32±3	29±6	44±1	41±1	30±2	29±4
Terrorism	70±3	69±3	40±4	50±7	75±2	74±3	70±2	60±4

6. CONCLUSION AND FUTURE WORK

This research resulted in creation of experimental datasets in Russian that can be used for training and testing machine learning methods for extremist text detection. Specific lexis that characterizes different categories of illegal texts was extracted. The proposed psycholinguistic and semantic features were used in our previous research for sentiment analysis [Vybornova, O., et al, 2011]. It was shown that extremist text detection can be performed with satisfactory quality using machine-learning methods and set of psycholinguistic and semantic features. It was also shown that fastText embeddings allow significantly improve a quality of extremist topic identification for texts from social networks.

ACKNOWLEDGEMENT

The reported study was funded by RFBR according to the research projects № 16-29-09546, № 18-00-00606(18-00-00233), № 19-07-00806.

REFERENCES

Agarwal, S., Sureka, A., 2015. Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. *arXiv preprint arXiv:1511.06858*

Agarwal, S., et al, 2015, Open source social media analytics for intelligence andsecurity informatics applications. *International Conference on Big Data Analytics.* Hyderabad, Telangana State, India, pp. 21–37.

Bojanowski, P., et al, 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*

Chen, H., 2007, Exploring extremism and terrorism on the web: the dark web project. *Pacific Asia Workshop on Intelligence and Security Informatics*. Chengdu, China, pp. 1–20.

Cohen, K., et al, 2014. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, Vol. 26, No.1, pp. 246–256.

Correa, D., Sureka, A., 2013. Solutions to detect and analyze online radicalization: a survey. *arXiv preprint arXiv:1301.4916*

Klausen, J., et al, 2016. Finding online extremists in social networks. *arXiv preprint arXiv:1610.06242*

Mbaikodzhi, E., et al, 2012. The method of automatic classification of short text messages. *Information Technologies and Computing Systems*, No.3, pp. 93–102.

Lyashevskaya O., 2010, Bank of Russian constructions and valencies. *LREC 2010*, Malta, Valletta, May, pp. 1802-1805.

Pedregosa, F., et al., 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, Vol. 12(Oct), pp. 2825–2830.

Scanlon, J.R., Gerber, M.S., 2014. Automatic detection of cyber-recruitment by violent extremists. *Security Informatics*, Vol. 3, No.1, pp. 1–10.

Vybornova, O., et al, 2011. Social tension detection and intention recognition using natural language semantic analysis. *Proceedings of European Intelligence and Security Informatics Conference.* Athens, Greece, pp. 277–281.

Zurini, M., 2015. Stylometry metrics selection for creating a model for evaluating the writing style of authors according to their cultural orientation. *Informatica Economica*, Vol. 19, No.3, pp. 107-119.