Natural Text Anonymization Using Universal Transformer with a Self-attention*

Aleksandr Romanov [1]
alexx.romanov@gmail.com

Anna Kurtukova [1]
av.kurtukova@gmail.com

Anastasia Fedotova [1]
afedotowaa@icloud.com

Roman Meshcheryakov [2]
mrv@ipu.ru

[1]Tomsk State University of Control Systems and Radioelectronics
Tomsk, Russian Federation
[2]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russian Federation

Abstract

The paper focuses on the anonymization of natural language text in Russian. The problem of anonymization is topical in connection with the need to conduct studies aimed at assessing the effectiveness and stability of methods of attribution of the text to its intentional distortion by various techniques of anonymization. The paper presents a technique for anonymizing a Russian text based on a fast correlation filter, dictionary synonymization and a universal transformer model with a self-attention mechanism. The automated system developed on its basis is tested on an experimental corpus of Russian texts. The texts obtained with its help are analyzed by the authorship identification system. The effectiveness of attribution of anonymous texts by a specialized software system was reduced to a level of random guessing, which allows to name the proposed methodology effective.

Keywords: *anonymization, authorship, fast correlation filter, deep learning, transformer, self-attention*

1 Introduction

Every year there are more and more software products which allow to communicate on the Internet through text messages while maintaining anonymity. The development of such technologies leads to an increase in the number of offenses in cyberspace. However, technical means are not always enough to identify the subject that intentionally committed an illegal

action. In such cases, it becomes necessary to conduct an expert examination using various techniques and tools for identifying the authorship. This makes the task of attributing the author of a natural language text an important aspect of information security.

Existing software for authorship attribution allows to take into account various linguistic and statistical parameters and are quite effective in most cases. However, they are not resistant even to most trivial anonymization techniques. That is why an additional research aimed at assessing the effectiveness and sustainability of existing methods of attribution of a natural language text to its intentional distortions by various anonymization techniques is quite relevant. Therefore, the goal of this study is to create a technique for anonymizing a Russian text and a software system to realize it.

The problem of anonymization of natural language texts is frequently discussed in the works of foreign researchers, and the proposed approaches demonstrate positive results. However, a lack of techniques and software for Russian-language texts should be noted, which makes this study more relevant.

An automated anonymization system for text documents is presented in [Mamede et al., 2016]. The system was tested on different styles and types of texts using different anonymization methods, such as suppression, tagging, random substitution and generalization. The authors disclosed that all methods have their drawbacks, but the method of generalization as solution for anonymization of the text was recognized as the most acceptable, because it allows preserving the natural appearance of the text and its readability. The evaluation showed that the use of the tagging method facilitates the reading of anonymous text, preventing some semantic deviations caused by the substitution of words in the original text. The advantage of the system is the possibility to replace easily its modules in order to support new methods of anonymization, other languages, or to improve the module performance. Three experts using a data set consisting of 75 documents from two different corpora evaluated the system. The experts evaluated anonymized texts using the suppression method. The results show that readers were able to compare 67% of the optimized texts with the original.

The article [Sardina et al., 2018] has two main objectives: to compile a new corpus in Spanish with annotated anonymized spontaneous dialogue data, and to investigate techniques for automating the sensitive data identification task solution, in a setting where initially no annotated data from the target domain are available. Several methods that can successfully anonymize data have been investigated by the authors. Randomization into data changes without loss of value by adding noise, and the aggregation method of data reduction, has been effective only on structured setting in the form of graphs or tables. These methods are considerably oriented to the anonymization of structured datasets in the form of graphs or tables. A better suited classification of techniques specifically oriented to the anonymization of unstructured textual data is suppression: a neutral place-holder replaces the item to be anonymized, e.g. "XXXX", "ANON", tagging: a label indicating its category or identifier is used to replace the item to be anonymized, e.g. "LOC", "LOCATION453", and generalization: the item to be anonymized is substituted by one of the same category. It is noted that the best results are achieved by a combination of these methods. The experiment was carried out on the ESPort corpus, which includes a selection of 1170 spontaneous spoken human-human dialogues from phone calls. The corpus has been anonymized using the substitution technique, which implies that the result is a readable natural text, and it contains annotations of some linguistic and extra-linguistic phenomena annotations like laughter, repetitions, mispronunciations.

The technique [Nguyen-son et al., 2015] is based on the assessment of a threshold of frequency metric to improve the naturalness of fingerprinted messages. As a metric, a combi-

nation of precision and distribution estimates was used, calculated on the number of degrees of generalization of sensitive phrases and the loss of information obtained from them. Based on the proposed technique and generalization method, a web application designed to anonymize personal information in messages before posting on Facebook was created. In addition, authors used synonimization to create fingerprints - identifiers for each message so that if personal information is disclosed, the identity of the person who provided it can be established. The approach was tested on personal messages - the corpus included more than 55000 samples of identifying phrases that were distributed among groups: hometown, education, work, religion, politics, sports and personal interests. The accuracy of the technique was 92%.

The paper [Kacmarcik et al., 2006] explores techniques for reducing the effectiveness of standard authorship attribution methods. The authors consider two levels of anonymization: shallow and deep. In the test set, authors show that shallow anonymization can be achieved by making 14 changes per 1000 words to reduce the likelihood of identifying author by 17%. For deep anonymization the unmasking work of Koppel and Schler is adapted. The possibility of creating a tool to support document anonymization has been explored on the assumption that the author has undertaken basic preventative measures (like spellchecking and grammar checking). For experiments, the authors have chosen a standard data set, the Federalist Papers. The support vector machine (SVM) is used for each feature set. However, modifying the document to increase or decrease the frequency of a term will necessarily impact the frequencies of other terms and thus affect the document stylometric signature. One limitation to this approach is that it applies primarily to authors that have a reasonably sized texts corpus. Finally, simple SVMs less resilient to obfuscation attempts than Koppel and Schler's unmasking approach. Classifiers with a minimum number of features are susceptible even to trivial methods of entanglement. The accuracy of the technique is 86.86%.

The system presented in [McDonald et al., 2012] defines the steps necessary to anonymize documents and implements them. This system has been implemented via tool JStylo-Anonymouth [Authorship Attribution], which has been released under an open source license (GPL 3). The software allows attribution of authorship, calculation of features most conducive to the identification process, and offers ways to change feature vectors to ensure anonymity. The authors use the K-means clustering method. The results show that 80% of the study participants were able to anonymize their documents in terms of a fixed corpus and limited feature set used. However, it was found that it was difficult to make changes to the pre-written documents, which was a serious shortcoming of this approach.

The research in [Simi et al., 2017] is devoted to prevention of attacks on a person's privacy based on confidential information from social networks. To prevent such attacks K-anonymization is used. In general, k-anonymization is used to achieve this purpose. The technique is ineffective for authors that have a reasonably sized corpus. The authors of [Simi et al., 2017] propose three effective algorithms, most often mentioned in scientific papers, which allow to use various anonymization strategies for a more complete assessment. The authors conducted tests for the algorithms incognito, Samarati and Sweeney. The data sets obtained from UCIrvine were used for study. In the course of the tests, the parameters of the k-th value and the size of the dataset were changed. A dependency was established: the greater the value of k, the greater the time spent on anonymization. Experiments demonstrate that, as the number of k value expands the time taken for anonymization increases. According to the results of the tests, the authors concluded that among the three algorithms, the Samarati algorithm has the advantage, since it provides effective anonymization even on a large amount of data.

In the article [Maeda et al., 2016] the method of anonymization of unstructured texts is proposed using the dictionary of anonymization and a quasi-identifier (information identifying the set of connected objects, for example, "nation - place of birth"). The system breaks parts of quasi-identifiers into alternate characters, for example, "" in order to prevent the re-identification of private information. The anonymization dictionary is created from a list of quasi-identifiers. Further, an accelerated process of anonymization based on heuristics and set theory is proposed. The advantage of this method is the maximum preservation of the author's text.

The authors of the article [Brennan et al., 2009] investigated adversarial attacks and their devastating effect on the robustness of existing statistical methods of analysis in authorship recognition. The results of the study are based on the participation of 15 individual authors. Each author had to submit approximately 5000 words of sample writing. Each writing sample had to be from some sort of a formal source. This was intended to eliminate slang and abbreviations. Then the authors of the texts made an obfuscation attack to hide their own style, and also sought to imitate the style of another author. Then, three attribution methods were applied to the resulting corpus: a statistical technique using the signature (accuracy 95%), the approach using neural networks (accuracy 78.5%) and the classification based on synonyms (accuracy 91.6%). Based on the results, it was concluded that all three methods were not effective enough in such attacks. The obfuscation attack reduces the effectiveness of the techniques to the level of random guessing and the imitation attack succeeds with 68-91% probability depending on the stylometric technique used. The authors highlight the following reasons for the negative results: test subjects were unfamiliar with stylometric techniques, without specialized knowledge in linguistics, and spent little time on the attacks.

The article [Quiring et al., 2019] is devoted to the related topic of anonymization of program source codes. The paper presents a new, based on machine learning, method of "attack on authorship" of source code. The essence of the approach is to perform a number of semantic code transformations that mislead deanonymization algorithms, but look plausible to the developer. The attack is guided by Monte-Carlo tree search that enables to operate in the discrete domain of source code. The "black box" strategy allows creating non-targeted attacks that prevent correct identification, as well as targeted attacks that mimic the style of the developer. To verify the result, a series of experiments were conducted using the source code of 204 programmers. The experiment has shown that the author's technique significantly affects the methods of attribution - their accuracy is reduced from 88% to 1%. Another experiment was aimed at investigating the effect of targeted attacks. The result showed that in a group of programmers, each person can impersonate another developer in 77 of cases%.

2 Anonymization Technique Based on the Fast Correlation Filter, Dictionary Synonymization and Universal Transformer

As a rule, approaches to the anonymization of natural language texts are based on classical mathematical algorithms and statistics. Seldom modern methods operate with machine learning (ML) algorithms, despite their high efficiency in related problems of text mining [Kurtukova et al., 2019a]-[Romanov et al., 2018]. This is due to the specifics of the anonymization process - it is necessary to modify the text exclusively so that its meaning is not ultimately

distorted, which requires the researcher to study all stages of the technique.

The technique presented in Fig. 2 is based on the assumption that the deep NN architectures intended for the text generation can improve the process of "obfuscation" of the text by adding new words and figures of speech which do not affect the general meaning in any way, and consists of the following steps: 1. Text features extraction and calculation of average values of text features based on the corpus of Russian-language texts.

2. Filtering the calculated features with a fast correlation filter and selection the most informative features for further smoothing.

3. Text correction by smoothing the identifying features.

4. Generation of anonymized text by the "universal transformer" model [Dehghaniet al., 2019] based on input dictionary-smoothed text.

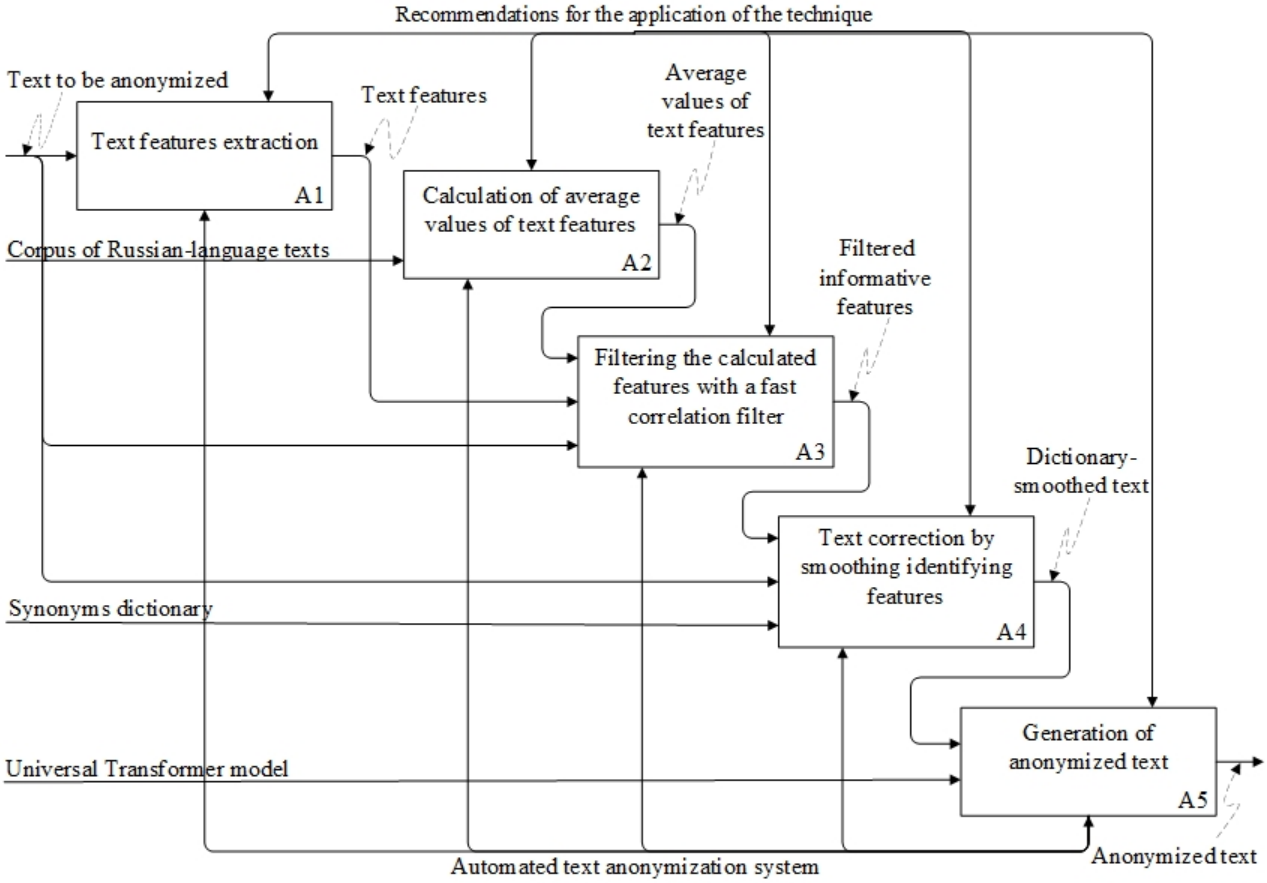

Figure 1: Anonymization technique based on the fast correlation filter, dictionary synonymization and universal transformer

For identification of the author, about a thousand different groups of statistical characteristics are used [Romanov et al., 2011]:

- lexical (punctuation, special symbols, lexicon, jargon, dialectics, archaism);
- morphological (lemmas, morphemes, grammar classes);
- syntactic (complexity, position of words, prevalence, sentiment analysis);
- structural (headings, fragmentation, citation, links, design, placement parameters);
- content-specific (keywords, emoticons, acronyms and abbreviations, foreign words);

- idiosyncratic stylistic features (spelling and grammatical errors, anomalies);
- document metadata (steganography, data structures).

However, five of the most informative features of the author's style have been allocated which could affect the authorship identification process:

- unigrams (frequencies of letters of the Russian alphabet);
- trigrams (frequencies of triples of letters of the Russian alphabet);
- Sharov-words (frequencies of all words from the dictionary of S. Sharov [Sharov]);
- punctuation (frequency of punctuation marks);
- parts of speech (distribution of words among parts of speech).

Based on the features, the average frequency of occurrence in the training corpus and in the anonymous text are calculated. The resulting identifying values are passed to the fast correlation filter. It accepts the input a full set of available for analysis features and uses a measure of symmetrical uncertainty to determine the dependencies between the features:

$$SU(X,Y) = 2\frac{H(X) - H(X/Y)}{H(X) + H(Y)} = SU(Y,X),$$

where $H(X)$, $H(Y)$ – are the entropies of random variables having accordingly i и j states.

$$H(X) = -\sum_i P(x_i) log_2(P(x_i)),$$

where $H(X|Y)$ – conditional entropy:

$$H(X/Y) = -\sum_i P(y_j) \sum_i P(x_i/y_j) log_2(P(x_i/y_j)),$$

where $P(x_i)$, $P(y_j)$ – prior probabilities for all values X and Y, $P(x_i/y_j)$ – posterior probability X with known Y.

The closer the SU value is to the unit, the higher is the dependence of the features on each other. Thus, a search is made for the subset which best describes the author's style, and the remaining uninformative features are excluded from the further process of anonymization.

The obtained informative features are calculated in the anonymized text and smoothed in accordance with the following principles:

- Words frequencies are compared with the Sharov's dictionary and they are replaced with synonyms that have the lowest frequency according to the dictionary.
- For character unigrams and trigrams, words are detected which have the high frequency of occurrence of specific n-grams, they are replaced with synonyms which contain other sets of n-grams in priority.
- Punctuation marks are divided into functional groups: isolating (for a text), separating and emphasizing (for a sentence). For text anonymization on the basis of punctuation, punctuation marks are replaced within an isolated functional group according to average statistics.
- When considering the frequency of occurrence of different parts of speech, they are replaced by equivalent structures according to the replaced part of speech. Trigram replacements indirectly affect the frequency of the occurrence of unigrams in the text, which in turn brings this indicator closer to the average value.

The final step is to submit the corrected text to the input of the deep learning model. For this purpose, a transformer model was chosen [Wang et al., 2019] and [Vaswani et al., 2019].

This solution is due to the special popularity of this architecture for solving related text mining problems [Sun et al., 2019]-[Zihang et al., 2019] among modern deep learning models and demonstrates results superior to simpler architectures.

The transformer processes the input text sequence at the level of words and characters, and also uses the self-attention mechanism to study the context. The main advantage of a transformer over simple recurrent neural networks (RNN) and hybrid neural networks (HNN) is the model training speed. This is achieved by parallel processing of words and setting correspondences between them (one word correlates with the other words in a sentence, forming a context).

The classic transformer modification called "Universal Transformer" (see Fig. 2) is used in the article. The characteristic of this model is the use of a more computationally efficient recurrence apparatus: several modules of uniform, parallel-in-time functions of recurrent transformations. Also, the universal transformer operates on the basis of an adaptive algorithm which regulates the amount of computing resources spent on processing one element of the sequence. If the element is a word that has several different meanings, this algorithm increases the number of iterations. Iterations are designed to improve the model's understanding of the context. In addition, in this case, the algorithm reduces them when processing simple and unambiguous elements.

At each stage, the transformer does not process the text sequentially, but simultaneously, after that it checks the received interpretation of each character or the word using the self-attention mechanism.

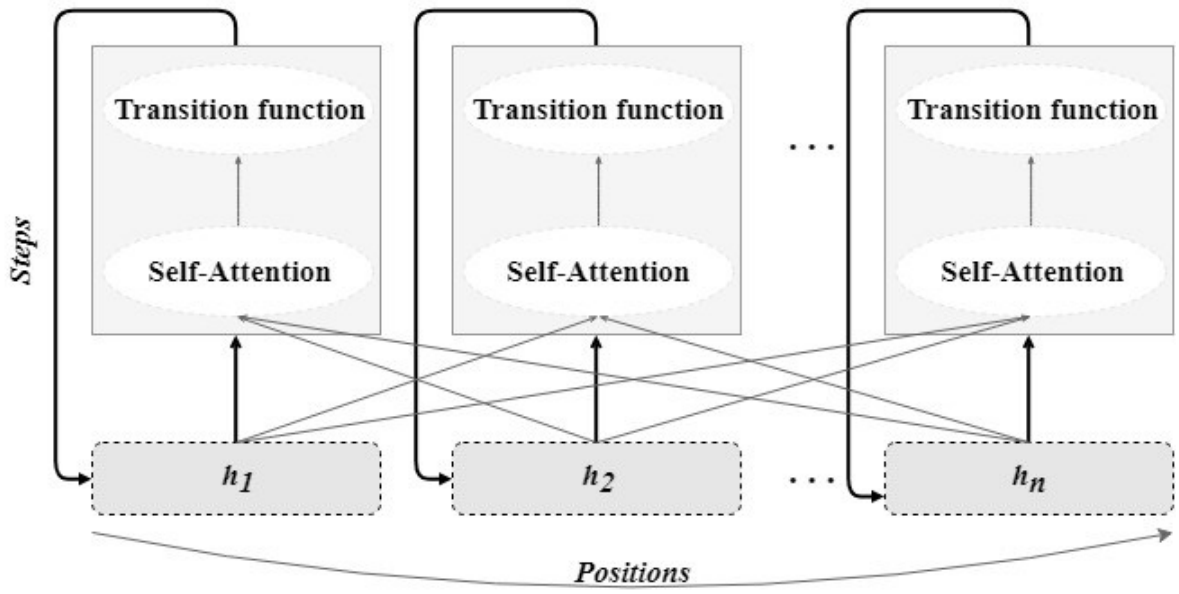

Figure 2: Universal transformer model

The model repeatedly checks a number of representations of tensors of the first rank (indicated in the figure as h) sequentially for each position, combining information from different positions in combination using the self-attention mechanism and recurrence transformation functions.

In this work, a universal transformer, trained on the corpus of Russian-language texts

and their brief descriptions, reflecting the main meaning and playing the role of labels in this task, is able to generate a new expanded text based on the input sample. thus confusing the source text and distorting identifying features that indirectly indicate the author's style.

3 Experiment Setting and Results

An automated system for anonymization of natural-language text in Russian based on the presented technique was developed. Python was chosen as a programming language [Chollet, 2017], as it is especially popular for text analysis. To perform morphological analysis, the pymorphy2 library was used by default [Korobov], which has a dependent module pymorphy2-dicts which contains a collection of Russian-language dictionaries OpenCorpora. An electronic version of N. Abramov's synonym dictionary [Synonyms dictionary] was used in the module of smoothing of features. The automated system provides an ability to connect and use other morphological analyzers. The universal transformer model was implemented on the basis of the architecture offered by the modern tensor2tensor deep learning library [Tensor2Tensor Documentation] and has been modified due to additional embeddings in accordance with the specifics of the text anonymization task for the Russian texts. For the experiments, the corpus of the Russian-language texts collected from the M. Moshkov electronic library [Moshkov Library], containing the texts of 23 writers with a total volume of 115 samples was used.

The software system starts its work by calculating the average values of identifying features that are found in the corpus of Russian texts. The calculation results (see Table 1) were ranked by the frequency of occurrence of the grams for each of the identifying features.

Table 1: The set of identifying features

	Parts of the speech		Punctuation		Trigrams		Sharov		Uniframs	
№	Feature	Frequency	Feature	Frequency	Feature	Frequency	Feature	Frequency	Feature	Frequency
1	Noun	0.2717	,	1.4323	"что"	0.0159	и	0.0363	о	0.1107
2	Verb in personal form	0.1422	-	0.1306	"про"	0.0119	в	0.0278	а	0.0835
3	Adjective	0.1069	"	0.1138	"при"	0.0108	не	0.0207	е	0.0826
4	Preposition	0.1044	?	0.1005	"как"	0.0102	он	0.0189	и	0.0674
5	Union	0.1014	!	0.0893	"был"	0.0096	на	0.0165	н	0.0651
6	Pronoun	0.0791	:	0.0745	"это"	0.0095	я	0.0156	т	0.0601
7	Adverb	0.0652	...	0.0704	"все"	0.0086	что	0.0125	с	0.0518
8	Particle	0.0538	;	0.0137	"под"	0.0082	тот	0.0114	л	0.0504
9	Infinitive	0.0239	)	0.0119	"так"	0.0068	быть	0.0112	в	0.0443
10	Adverbial participle	0.0109	(	0.117	"раз"	0.0061	с	0.0111	р	0.0438

Then, the frequencies of features for the sample text to be anonymized are calculated. Tables 2 - 5 show the frequency of occurrence of features of the whole corpus, for a specific text and their deviation from the average. After smoothing the features according to the dictionary of synonyms to medium frequencies, reanalysis of frequency analysis was carried out, the results of which are also presented in the tables.

Based on the results, it can be concluded: the module for smoothing of the informative text features of an automated system performs its functions correctly and has a significant impact on the process of text anonymization. The next step is to obtain the corrected text

by the universal transformer and to start the process of generating the final text. For a new corpus, process begins with training and its speed directly depends on such factors as the number of samples in the training corpus, the input sequence length and the complexity of the occuring words.

To correct the values of weights and hyperparameters of the model, cross entropy metrics and Kullback-Lybler distances, traditional for complex text analysis tasks, were calculated. The final value of the loss function was 3.28 for the test part of the original corpus, which is a positive result.

Table 2: Indicator of occurrence of punctuation

№	Feature	Frequency in the corpus	Frequency in the text	Deviation	Frequency after smoothing
1	,	1.4323796	1.0180969	-0.4142827	1.3306129
2	-	0.1306444	0.1005123	0.0301321	0.1402113
3	"	0.1138403	0.1010302	0.0128101	0.1120889
4	?	0.1005719	0.1506129	0.0500410	0.1015788
5	!	0.0893705	0.0408639	-0.0485065	0.0801210
6	:	0.0745130	0.0321074	-0.042405	0.0603214
7	...	0.0704761	0.0277291	-0.0427469	0.0688701
8	;	0.0137191	0.0011675	-0.0125515	0.0116751
9	)	0.0119047	0.0026269	-0.0092776	0.0162754
10	(	0.0116949	0.0026269	-0.0090678	0.0162754

Table 3: Indicator of occurrence of trigrams

№	Feature	Frequency in the corpus	Frequency in the text	Deviation	Frequency after smoothing
1	что	0.0159385	0.0128798	0.0076731	0.0128798
2	про	0.0119698	0.0165072	0.0010719	0.0100090
3	при	0.0108387	0.0108821	-0.0030945	0.0106718
4	как	0.0102530	0.0114078	-0.0018135	0.0114078
5	был	0.0095941	0.0050467	-0.0073249	0.0085225
6	это	0.0095326	0.0057302	0.0065621	0.0093382
7	все	0.0086597	0.0097781	-0.0013886	0.0071147
8	под	0.0082749	0.0119861	0.0013155	0.0092302
9	так	0.0067989	0.0066764	-0.0020908	0.0066764
10	раз	0.0061580	0.0083587	0.0004178	0.0059444

Table 4: Indicator of occurrence of Sharov's features

№	Feature	Frequency in the corpus	Frequency in the text	Deviation	Frequency after smoothing
1	и	0.03636	0.0326469	-0.0348241	0.03595942
2	в	0.02779	0.0280010	-0.0163191	0.0273947
3	не	0.02069	0.0317091	0.0110191	0.0256901
4	он	0.01894	0.0403210	0.0213810	0.0223512
5	на	0.01659	0.0086473	-0.0079427	0.0146579
6	я	0.01563	0.0069060	-0.0138601	0.0102431
7	что	0.01255	0.0143562	-0.0061967	0.0132680
8	тот	0.0114	0.0031809	0.0019673	0.0113363
9	быть	0.01122	0.0077013	-0.0048563	0.0117470
10	с	0.01115	0.0140214	-0.0378629	0.0120301

Upon training, the obtained model is used to anonymize the user sample. The transformer, which processes incoming sentences one at a time, forms a new text, preserving its originally intended meaning, by generating new phrases and figures of speech. The text anonymized by the model is written line by line to the output file. The user receives reference information about the changes made and recommendations in case he wants to anonymize manually, without referring to the automatically generated text.

To assess the effectiveness and sustainability of the authorship attribution technique of a natural language text to its intentional distortions, it was decided to conduct a series of additional experiments with the automated system for authorship identification "Avtoroved" [Kurtukova et al., 2019a]. This system uses classical NN, SVM, and the QSUM method and demonstrates a high authorization accuracy of 95-98% for the texts written in the Russian language.

Table 5: Indicator of occurrence of unigrams

№	Feature	Frequency in the corpus	Frequency in the text	Deviation	Frequency after smoothing
1	о	0.1107492	0.1095906	-0.0011585	0.1033789
2	а	0.0835207	0.0806022	-0.0029183	0.0799414
3	е	0.0826149	0.0848894	0.0022745	0.0797453
4	и	0.0674710	0.0602521	-0.0072189	0.0659077
5	н	0.0651345	0.0556855	-0.0094490	0.0631874
6	т	0.0601586	0.0577851	-0.0023733	0.0599754
7	с	0.0518844	0.0477657	-0.0041185	0.0502248
8	л	0.0504705	0.0473027	-0.0031677	0.0485175
9	в	0.0443202	0.0424008	-0.0019193	0.0404249
10	р	0.0438877	0.0454505	0.0015628	0.0431268

The authors' texts of various styles and lengths were involved in the experiment: Agafonov V., Grossman V.S., Bykov V., Bulgakov M., Knorre F., Druzhnikov Yu., Koval Yu., Krivin F., Kaledin S., Degen I. The results of the analysis with "Avtoroved" of different-sized corpora of the original, unchanged samples and anonymized ones by the developed software system are presented in Table 6.

Thus, the developed technique is even resistant to multi-stage analysis and the identification of informative features and allows to evaluate objectively the reliability of attribution techniques and software systems based on them to intentional distortions of the source text. The proposed technique reduces the accuracy of the authorship identification of a Russian-language text by half, which is a serious result for the study.

Table 6: Results for Russian-language texts of different lengths

Size of corpus	Accuracy on the original corpus	Accuracy on the anonymized corpus
10 texts	92.7%	39%
20 texts	93.1%	39.7%
30 texts	95%	42.1%
40 texts	96.8%	43%
50 texts	97%	46%

An example of text anonymization was presented on an extract from the novel of E. I. Zamyatin "The Scourge of God" (Fig. 3 and Fig. 4).

Below is an extract of the novel anonymized by the proposed technique, where the uppercase signs are smoothed with a synonyms dictionary, italics are the context generated by the transformer, and highlighting is the transformer's self-attention mechanism.

This example was demonstrated to 10 linguistic experts. Expert ratings were distributed on a ten-point scale. The maximum score was assigned if the meaning of the anonymized text does not differ from the original one and the text is completely clear to the expert. The minimum, if the meaning of the anonymized text differs from the original, the text is not clear to the expert, since most of the changes in the text do not correspond to semantics.

Беспокойство было всюду в Европе, оно было в самом воздухе, им дышали.

Все ждали войны, восстаний, катастроф. Никто не хотел вкладывать денег в новые предприятия. Фабрики закрывались. Толпы безработных шли по улицам и требовали хлеба. Хлеб становился все дороже, а деньги с каждым днем падали в цене. Вечное, бессмертное золото вдруг стало больным, люди потеряли в него веру. Это было последнее, ничего прочного в жизни больше не осталось.

Прочной перестала быть самая земля под ногами. Она была как женщина, которая уже чувствует, что ее распухший живот скоро изрыгнет в мир новые существа - и она в страхе мечется, ее бросает в холод и жар.

Была зима, когда птицы замерзали на лету и со стуком падали на крыши, на мостовую. Потом настало такое лето, что деревья цвели три раза, а люди умирали от лихорадочного жара земли. В июльский день, когда земля лежала с черными, пересохшими, растрескавшимися губами, по ее телу прошла судорога. Земля выла круглым, огромным голосом. Птицы с криком носились над деревьями и боялись на них сесть. Молча падали на дали стены, церкви, дома. Люди бежали из городов как животные и стадами жили под открытым небом. Время исчезло. Никто не мог сказать, сколько часов или дней это длилось.

Вся покрытая холодным потом, земля наконец затихла. Все бросились в церкви. Сквозь трещины в сводах зияло раскаленное небо. Пламя свечей пригибалось от человеческих испарений, от тяжести выбрасываемых вслух человеческих грехов. Бледные священники кричали с амвонов, что через три дня мир разлетится в куски, как положенный на уголья каштан.

Этот срок прошел. Земля по временам еще, чуть-чуть вздрагивала, но она уцелела. Люди вернулись в дома и начали жить. По ночам они знали, что все прежнее кончилось, что теперь жизнь надо мерить месяцами, днями. И они жили бегом, коротко, задыхаясь, спеша. Как богач перед смертью торопится все раздать, так женщины, не жалея, раздавали себя направо и налево. Но они теперь не хотели больше рожать детей, груди им стали не нужны, они пили лекарства, чтобы стать безгрудыми.

И как женщины - незасеянными, бесплодными оставались поля. Деревни пустели, а города переливались через край, в городах не хватало домов. Было нечем дышать в театрах и в цирках, не замолкала музыка, огни не потухали всю ночь, красные искры сверкали в шелку, в золоте, в драгоценностях - и в глазах.

Figure 3: Source text "The Scourge of God"

Беспокойство было *в то время* ВЕЗДЕ *повсеместно* в Европе, оно было *даже* в самом КІСЛОРОДЕ, им дышали….

Все ждали СРАЖЕНІЯ восстаний, катастроф, битв, *трагедий*! Никто не ЖЕЛАЛ вкладывать денег в новые, *не существовавшие ранее* предприятия. ЗАВОДЫ *спешно* закрывались *из-за нехватки ресурсов*. МАССЫ безработных *бродяг* шли, *двигались* по улицам *и переулкам* и требовали, *просили* хлеба. Хлеб становился все дороже, а деньги с каждым днем *ежедневно* падали в цене. ПОСТОЯННОЕ, бессмертное *драгоценное* золото вдруг *неожиданно* стало больным, люди УТРАТИЛІ в него веру, *надежду*. Это было КРАЙНЕЕ, ничего прочного *и стабильного* в жизни больше не осталось...

Прочной перестала, *прекратила* быть самая ПОЧВА под ногами *ступивших на неё*. Она была как женщина, *девушка,* которая уже ОЩУЩАЕТ, что ее распухший живот скоро изрыгнет в СВЕТ новые существа, *создания* - и она в страхе мечется, ее бросает в холод и жар.

Была *в то время* зима, когда птицы ДРОГЛІ на лету и со стуком падали *вниз* на ДОМА на мостовую! Потом настало, *случилось* такое *жаркое* лето что деревья цвели *целых* три раза, а люди умирали от лихорадочного жара земли. В июльский *летний* день, когда ПОЧВА НАХОДІЛАСЬ с черными, пересохшими, растрескавшимися губами, по всему ее телу прошла судорога, *дрожь*. Земля КРІЧАЛА, вопила круглым, БОЛЬШІМ голосом! Птицы с ВІЗГОМ носились над *цветущими* КУСТАМІ и боялись на них сесть. БЕСШУМНО, *осторожно* падали на дали стены ХРАМА дома. НАРОДЫ *дико* бежали из *своих* городов как животные и стадами *скитались,* ОБІТАЛІ под открытым небом. Время ІІСТЕКЛО, *не осталось ни минуты*. Никто *не знал* и не мог СООБЩІТЬ, сколько ВРЕМЕНІ или дней это длилось.

Вся *целиком и сплошь* покрытая *ледяным*, холодным потом, земля *наконец* ЗАМОЛЧАЛА! Все бросились в *ближайшие* церкви. Сквозь РАСКОЛЫ в сводах зияло раскаленное небо. ОГОНЬ свечей пригибалось от ЛЮДСКІХ испарений, от тяжести *и груза* выбрасываемых вслух, *наружу* человеческих грехов. Бледные священники *громко* кричали с амвонов, что СПУСТЯ три дня СВЕТ разлетится в ЧАСТІ, как положенный на уголья каштан.

Этот МОМЕНТ *времени* прошел… ПОЧВА по временам еще *изредка,* чуть-чуть вздрагивала, но она уцелела и *осталась в полной сохранности*. Люди вернулись в *свои* дома и СТАЛІ жить *спокойной и размеренной жизнью*. По ВЕЧЕРАМ они знали, что все прежнее кончилось, *прекратилось* что теперь, с этого момента, жизнь надо СЧІТАТЬ месяцами, днями. ІІ они СУЩЕСТВОВАЛІ БЫСТРО коротко задыхаясь, ТОРОПЯСЬ *и спеша*. Как богач перед своей смертью СПЕШІТ все раздать, *отдать*, так ДЕВУШКІ, не жалея, раздавали себя направо и налево. Но они СЕЙЧАС *не хотели*, не ЖЕЛАЛІ больше рожать детей, груди им стали не нужны, они пили ТАБЛЕТКІ, чтобы стать безгрудыми.

ІІ как женщины - ПУСТЫМІ, бесплодными, *осиротевшими* оставались поля. СЕЛА пустели, а города переливались, заполнялись через край, *больше, чем, нужно*, в городах не хватало КВАРТІР. Было нечем дышать *везде, где бы они не были*, в театрах и в цирках, не УТІХАЛА *громкая* музыка, огни не КОНЧАЛІСЬ всю ночь, красные *яркие* искры МЕРЦАЛІ в шелку, в золоте в драгоценностях - и в глазах.

Figure 4: Anonymized text "The Scourge of God"

For these estimates, the concordance and Pearson coefficients were calculated. The value of the expert's concordance coefficient was 0.76, which indicates a high degree of consistency of expert opinions. Its significance was confirmed by Pearson's score of 116.9. This means that obtained results make sense and can be taken into account in the study. The results of expert evaluation suggest that for all the considered text features, it was possible to achieve a satisfactory smoothing of indicators, and the text generated by the transformer does not affect the semantic, therefore, the functioning of the automated system is correct.

4 Conclusion

As part of the study, a technique for text anonymization based on smoothing out selected informative features, a fast correlation filter, and a universal transformer with a self-attention was proposed. The software system developed on its basis was tested on the corpus of Russian-language texts and showed a positive result.

The obtained results allow to reach the following conclusions:

- The features extracted by the fast correlation filter are quite informative.

- Smoothing with a synonyms dictionary is correct and smooths the frequency of occurrence to average values for the Russian language.

- The text generated by the universal transformer model is readable and meaningful, despite the changes made.

- Anonymized text can be recognized by the authorship identification system with an accuracy not exceeding 50%, and, therefore, can be used for anonymization.

The uniqueness of the developed system is due to the lack of similar solutions for the Russian language on the international market, a limited number of studies related to the problem under consideration, and the possibility of adapting the system to any other language.

In the future, we plan to continue the study, in particular, changing the technique by introducing new ML algorithms for filtering informative author's text features. It is assumed that an ensemble of two or more NN architectures will show the results better than those presented in this scientific work.

References

[Mamede et al., 2016] Mamede N., Baptista J, Dias F.(2016) Automated anonymization of text documents.2016 IEEE Congress on Evolutionary Computation (CEC). 2016. Pp. 1287-1294.

[Sardina et al., 2018] Sardina L. G., del Pozo A. and Aldezabal I. Automating the Anonymisation of Textual Corpora. 2018. 78 p.

[Nguyen-son et al., 2015] Nguyen-son H. Q., Tran M. T., Yoshiura H., Sohenara A. N., Echizen I. (2015) Anonymizing Personal Text Messages Posted in Online Social Networks and Detecting Disclosures of Personal Information. IEICE Transactions on Information and Systems. E98. Pp. 78–88.

[Kacmarcik et al., 2006] Kacmarcik G., Gamon M.(2006) Obfuscating Document Stylometry to Preserve Author Anonymity. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Pp. 444-451.

[McDonald et al., 2012] McDonald A. W., Afroz S., Caliskan A., Stolerman A., Greenstadt R. (2012) Use Fewer Instances of the Letter "i". Toward Writing Style Anonymization. PETS'12 Proceedings of the 12th international conference on Privacy Enhancing Technologies. Pp. 299-318.

[Authorship Attribution] Authorship Attribution and Authorship Anonymization Framework. URL: https://github.com/psal/jstylo.

[Simi et al., 2017] Simi M. S., Nayaki K. S., Elayidom M. S. An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity. IOP Conference Series: Materials Science and Engineering. 2017. 225 p.

[Maeda et al., 2016] Maeda W., Suzuki Y., Nakamura S. Fast text anonymization using k-anonyminity. Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services. 2016.

[Brennan et al., 2009] Brennan M., Greenstadt R.(2009) Practical attacks against authorship recognition techniques. Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference. 7 p.

[Quiring et al., 2019] Quiring E., Maier A. Rieck K. Misleading Authorship Attribution of Source Code using Adversarial Learning. 2019. ArXiv:1905.12386.

[Kurtukova et al., 2019a] Kurtukva A. V., Romanov A. S. (2019). The technique of deanonymization of the author of the source code based on the SVM and automatic filtering of features. In Proceedings of the XVI Conference Prospects for the development of basic sciences. Vol. 7. Pp. 92-94. (In Rus.) = Kurtukva A. V., Romanov A. S. Metodika deanonimizacii avtora ishodnogo koda na osnove mashiny opornyh vektorov i avtomaticheskoi filtracii priznakov: Perspektivy razvitiya fundamental'nyh nauk: sbornik trudov XVI mezhdunarodnoi konferencii studentov, aspirantov i molodyh uchenyh, 2019. - S. 92-94.

[Kurtukova et. al, 2019b] Kurtukova A. V., Romanov A. S. (2019). Identification author of source code by machine learning methods. SPIIRAS Proceedings. Vol. 18(3). Pp. 741-765. (In Rus.) = Kurtukva A. V., Romanov A. S. Identifikaciya avtora ishodnogo koda metodami mashinnogo obucheniya. Trudy SPIIRAN, 2019. - №18(3). - S. 741-765.

[Romanov et al., 2018] Romanov A. S., Vasileva M. I., Kurtukova A. V., Meshheryakov R. V. (2018). Sentiment Analysis of TextUsing Machine Learning Techniques. Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Saint Petersburg, Russia, November. 2017. Pp. 86-95. (In Rus.) = Romanov A. S., Vasileva M. I., Kurtukova A. V., Meshheryakov R. V. Analiz tonal'nosti teksta s ispolzovaniem metodov mashinnogo obucheniya. Sbornik trudov konferencii "The II international conference R. Piotrowski's Readings LE AL'2017": M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, 2018. - S. 86-95.

[Dehghaniet al., 2019] Dehghani M., Gouws S., Vinyals O., Uszkoreit J., Kaiser L. (2019) UNIVERSAL TRANSFORMERS. ArXiv:1807.03819.

[Romanov et al., 2011] Romanov A. S., Shelupanov A. A., Meshheryakov R. V. (2011). Development and research of mathematical models, techniques and software of information processes in identifying the author of the text. 188 p. (In Rus.) = Romanov A. S., Shelupanov A. A., Meshheryakov R. V. Razrabotka i issledovanie matematicheskih modelej, metodik i programmnyh sredstv informacionnyh processov pri identifikacii avtora teksta: Tomsk: V-Spektr, 2011. - 188 s.

[Sharov] Sharov S. A. Frequency dictionary. URL: http://www.artint.ru/projects/frqlist.asp.

[Wang et al., 2019] Wang C., Li M. (2019) Language Models with Transformer. ArXiv:1904.09408.

[Vaswani et al., 2019] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I.(2019) Attention Is All You Need. ArXiv:1706.03762.

[Sun et al., 2019] Sun, Chi, Xipeng Qiu, Yige Xu and Xuanjing Huang.(2019) How to Fine-Tune BERT for Text Classification? ArXiv:1905.05583.

[Devlin et al., 2018] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. ArXiv:1810.04805.

[Zihang et al., 2019] Zihang D., Yang Z., Yang Y., Carbonell J. G., Le Q. V., Salakhutdinov R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. 2019. ArXiv:1901.02860.

[Chollet, 2017] Chollet F. Deep Learning with Python. Learning with Python: Manual. Manning Publications. 2017. 386 p.

[Korobov] Korobov M. Pymorphy2. URL:https://pymorphy2.readthedocs.io/en/latest/misc/citing.html .

[Synonyms dictionary] Synonyms dictionary URL: http://slovonline.ru/slovarsinonimov/.

[Tensor2Tensor Documentation] Tensor2Tensor Documentation. URL: https://github.com/tensorflow/tensor2tensor.

[Moshkov Library] Moshkov Library. URL: http://www.lib.ru.