IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 February 2024, accepted 10 March 2024, date of publication 13 March 2024, date of current version 20 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377231

RESEARCH ARTICLE

Semantic Clustering and Transfer Learning in Social Media Texts Authorship Attribution

ANASTASIA FEDOTOVA, ANNA KURTUKOVA, ALEKSANDR ROMANOV, AND ALEXANDER SHELUPANOV, (Senior Member, IEEE)
Department of Security, Tomsk State University of Control Systems and Radioelectronics, 634050 Tomsk, Russia

Corresponding author: Anastasia Fedotova (afedotowaaa@gmail.com)

This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2023–2025, project no. FEWM-2023-0015 (TUSUR).

ABSTRACT This paper is the fourth part of a research series that focuses on determining the authorship of Russian-language texts by analyzing short social media comments, including those from mass media and communities associated with destructive content. Semantic text clustering was used to analyze content and employed a transfer learning technique based on a pre-trained model to identify sensitive topics. Authorship attribution is implemented as a classical classification task with a closed set of authors and a more challenging open-set task. In the latter case, multiple experiments were conducted, incorporating the identification of destructive content with known authors and artificially generated texts. For open attribution, a method combining One-Class SVM and fastText was proposed. Results demonstrate high accuracy (92% or higher) for cases with 2 and 5 authors, regardless of comment length and the additional task of identifying authors of destructive text. Mixed-data experiments involving 10 or more authors yielded results comparable to or more accurate (84% or higher) than previous studies.

INDEX TERMS Authorship attribution, machine learning, natural language processing, semantic clustering, transfer learning.

I. INTRODUCTION

Text authorship attribution holds significant importance, especially when countering content that contains destructive elements. The term ''destructive texts'' refers to materials that either harm the audience or encourage illegal actions. In our contemporary, information-driven society, social networks and online platforms have transformed into arenas for the dissemination of various ideas and opinions. For a number of reasons, this position poses complex dangers to society security:

1) Identifying malicious authors. Authorship attribution is crucial for identifying individuals or groups who spread destructive content online. By determining the origin of such content, authorities can take appropriate actions to mitigate the harm caused by these authors.
2) Preventing cyberbullying. In the digital age, cyberbullying has become an issue, leading to severe mental and emotional distress for victims. Authorship attribution helps in tracing and holding accountable those responsible for online harassment and bullying.
3) Protecting vulnerable audiences. Destructive texts frequently target vulnerable groups, including children and individuals facing significant health challenges. Attribution can aid in monitoring and curtailing the spread of destructive content, safeguarding these at-risk populations.
4) Maintaining online discourse. By identifying the source of destructive or disruptive content, platforms can enforce community guidelines and ensure that discussions remain constructive and respectful.
5) Data-driven decision-making. Authorship attribution offers crucial insights into the dynamics of online harm, facilitating the development of more effective countermeasures. This information can inform the

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi.

development of more effective strategies to counteract destructive content.

6) Public safety. Online platforms have the potential to amplify destructive ideologies and threats to public safety. Authorship attribution helps in identifying and addressing such threats before they escalate into real-world violence or harm.
7) Counteracting disinformation. Destructive texts often overlap with disinformation campaigns. Knowing the source of disinformation can aid in countering false narratives and protecting the integrity of information online.

Authorship attribution is increasingly vital, especially in the context of social networks linked to communities that propagate destructive content. Recent advancements and implementations of specialized methodologies and algorithms have enabled more precise identification of the origins of destructive content. This, in turn, facilitates the recognition of authors and the implementation of essential measures to safeguard society's safety, mental health, and security.

This study aims to determine the authorship of Russian-language comments by social media users, including those from destructive communities. The research focuses on the development of the development of a modified authorship attribution methodology, building upon the foundations established in previously presented works [1], [2], [3], by using semantic text clustering and transfer learning. The work includes experiments with destructive content detection as well as the use of open attribution.

The scientific novelty of this study lies in its unique utilization of a methodology that combines transfer learning with semantic clustering.

The transfer learning process is based on a pre-trained model to identify destructive texts within the realm of short comment authorship identification. This advancement enhances the precision of classification and the ability to discern destructive text.

Using semantic clustering to identify destructive content offers a more efficient method for analyzing crucial semantic clusters within the text.

The combination of transfer learning and semantic text clustering introduces a novel methodology for establishing the authorship of Russian-language texts. This approach enables a more precise and well-established analysis of social media comments, particularly those containing destructive content. This has significance in uncovering potential threats and, in turn, deterring the spread of damaging content.

A. PROBLEM STATEMENT

Presently, the most prevalent issue in attribution pertains to the challenge of a closed set of authors. The proposed solution involves utilizing a collection of valid document samples from a finite group of potential authors. The objective is to discern the most probable author of a novel document whose authorship is unidentified, representing a novel case not previously encountered. It is crucial to underline that the actual author of an anonymous text might not necessarily be listed in candidates, particularly within the practical context. This complexity introduces a more intricate scenario: authorship attribution with an open set of candidates.

The peculiarity of this study lies in the specificity of the data used. The combination of destructive texts with ordinary user messages creates an additional task: the selection of negative influences on the content of society and the authors who create them. Therefore, the standard mathematical formulation of the problem in this study was modified. Let us generalize the mathematical formulation of the problem for any thematic texts.

There are several datasets: a dataset of thematic texts $\mathbf{T}$, a reference dataset of thematic texts (where all texts contain a given topic for certain) $\mathbf{T}_{\text{reference}}$. It is necessary to filter the dataset $\mathbf{T}$ based on the reference dataset. The transformation is performed by the semantic clustering of texts. It is a process of partitioning objects into number of clusters $\mathbf{C}$ based on the similarity of their features using formula 1:

$$clustering(\mathbf{T}_{\text{reference}}, \mathbf{C}) = \{c_1, c_2, \ldots, c_n\} \tag{1}$$

The result of such a process is a set of clusters, where c_i is the i-th cluster containing the most similar objects (texts). Partitioning into clusters is carried out to obtain the most frequent tokens $\mathbf{W}$ of the reference dataset for certain categories of texts: $N(\mathbf{W}, c_i)$ – number of times tokens occur in texts from cluster c_i and $N(c_i)$ – total number of texts in cluster c_i.

Then frequency of tokens $\mathbf{W}$ in cluster c_i is calculated using the following formula (2):

$$f(\mathbf{W}, c_i) = \frac{N(\mathbf{W}, c_i)}{N(c_i)} \tag{2}$$

So, filtered dataset $\mathbf{T}$ is formed by the tokens from clusters (3):

$$\mathbf{T} = \{t | t \epsilon \mathbf{T}, \exists \mathbf{W} : f(\mathbf{W}, c_i)\} \tag{3}$$

With the generated dataset, the classification problem is then solved directly.

In the mathematical problem of text authorship identification considered in the context of closed attribution, there are three sets of elements: $\mathbf{A} = \{a_1, a_2, \ldots, a_m\}$, $\mathbf{T} = \{t_1, t_2, \ldots, t_k\}$ and $\mathbf{T}' = \{t'_1, t'_2, \ldots, t'_s\}$, which represent authors, texts with known authorship and anonymous texts respectively. Each text, including anonymous texts, is matched with a feature vector. In the case of open attribution, it is noted that some anonymous texts were not written by any of the candidate authors. In other words, the number of authors in the testing sample is greater than the number of authors in the training set ($|\mathbf{A}_{\text{test}}| > |\mathbf{A}_{\text{train}}|$) ^($|\mathbf{A}_{\text{test}}| \bigcap |\mathbf{A}_{\text{train}}|$). In this case, all new authors are designated as a distinctive class, labeled as -1 (4), (5):

$$\mathbf{A}_{\text{test}} = \mathbf{A}_{\text{test}} \cup \{a''\} \tag{4}$$

$$\mathbf{T}' = \mathbf{T}' \cup \mathbf{T}'' \tag{5}$$

where a'' is authors belonged to -1 class, $\mathbf{T}'' = \{t''_1, t''_2, \ldots, t''_p\}$ – a'' authors' texts.

Then on the Cartesian product of the set of texts $\mathbf{T}$ and the set of authors $\mathbf{A}$ is a binary relation $R \subset \mathbf{T} \times \mathbf{A}$, in which tRa is fulfilled if some text t_i corresponds to author a_j. The condition that some anonymous text $t \in \mathbf{T}$ corresponds to author $a \in \mathbf{A}$. The relation tRa is fulfilled if the values of textual characteristics of the investigated text t_i correspond or are approximated to a certain degree to the values of features of the author a_j. In this case, the degree of approximation of values should be justified by experimental data.

Due to the thematic content of texts, it was decided to apply the transfer learning technique to solve the problem and utilize the knowledge of the model $\mathbf{M}_{\text{source}}$, which was originally trained for a related task, in the target model $\mathbf{M}_{\text{target}}$ to identify the author of the text:

$$\mathbf{M}_{\text{target}} = transfer(\mathbf{M}_{\text{source}}, \mathbf{T}, \mathbf{A}) \tag{6}$$

Thus, the solution of the problem is reduced to the calculation of the target function (7):

$$\mathbf{M}_{\text{target}}(a_j) : f(t_i') = [p(a_1), p(a_2), \ldots, p(a_m)] \tag{7}$$

where t_1' is an anonymous text, $p(a_1), p(a_2), \ldots, p(a_m)$ – probabilities of the text t_i' belonging to authors. The final answer in the authorship attribution question on an anonymous text is to select the author with the maximum likelihood value $max(p_i)$ based on the computation of the target function $f(t_i')$:

$$\mathbf{M}_{\text{target}}(a_j) = argmax(p_i) \tag{8}$$

Texts with certain authorship are used as training samples. An equal number of non-anonymous texts for each candidate author is provided. All texts used for the study are written in Russian.

The article presents a joint solution to the problems of determining destructive content and the authorship of a text due to the connection between these problems. Both tasks are related to information security areas. The dissemination of destructive content can harm mental health, influence the formation of opinions, and promote dangerous ideas. Therefore, it is important to identify not only destructive content, but also the authors who publish it. Including, distinguishing between authors of destructive content. This is the subject of a series of experiments in the article.

B. OUR RESEARCH BACKGROUND

This article aims to modernize the methodology described in our previous study, as illustrated in Figure 1. The methodology from our previous work, as detailed in [2], encompassed data preprocessing, selection of informative features, training of One-Class SVM and fastText models, and evaluation of the outcomes. Brief summaries of findings from past studies are provided below.

In previous research [1], [2], [3], the authors explored authorship attribution for classical literary texts, fanfiction works, and short comments by users on social networks in Russian. The findings revealed that in the context of classical machine learning (ML) methods, the support vector method (SVM) demonstrated optimal performance. It was trained on a selected feature space. For neural networks, fastText displayed favorable results with classical literary texts. However, in some instances, its accuracy lagged behind that of deep neural networks such as convolutional (CNN), long short-term memory (LSTM), their hybrids, bidirectional LSTM (BiLSTM), and two adaptations of Bidirectional Encoder Representations from Transformers – RuBERT [4] and MultiBERT [5]. Although the latter models showcased accuracy improvements of up to 5%, fastText prevailed in terms of training time when dealing with a maximum number of classes compared to all the aforementioned models.

Based on the results obtained, the experimentally proven effectiveness of fastText and SVM with feature selection has been established, due to which the open attribution problem has already been solved in the past work [2] by jointly using One-Class SVM for anomaly detection and fastText, which effectiveness has been experimentally proven for closed attribution in this and earlier works [1], [2], [3].

II. BACKGROUND AND RELATED WORK

Most of the works centered around text attribution have primarily focused on literary texts, where the text samples are significantly longer than the comments found on social media platforms. Consequently, methods that demonstrate high effectiveness for literary works might not yield equally meaningful results for generated texts. Literary texts typically exhibit distinct stylistic features. However, within social media, users frequently employ informal and unrestrained language, incorporating abbreviations, emojis, and other unconventional elements of style and grammar. This intricate dynamic complicates the task of authorship attribution, as stylistic features in comments may be less consistent and more susceptible to external influences. Using natural language processing techniques, analyzing the tone and emotional voice of the text can help identify the author of a comment more accurately. For effective authorship identification, a dataset containing information about the characteristics of each user should be used, which will be updated and extended to reflect new data.

Moreover, this paper takes into account texts produced by thematic organizations and their adherents who publish destructive content. In order to comprehensively assess the landscape of text author identification advancements, a meticulous analysis of contemporary works [1], [2] was supplemented.

In their paper [6], authors delve into text attribution within the English-language forum dedicated to discussing matters pertaining to destructive groups. To build the initial dataset, they extracted forum posts spanning from 2004 to 2010. This dataset encompasses 91,874 posts, comprising 13,995 discussions, and involving 2,082 active users. Three distinct datasets were formulated. The first dataset features three users selected based on their higher post count and destructivity scores. The second dataset includes five users chosen using the same criteria, and the third dataset comprises

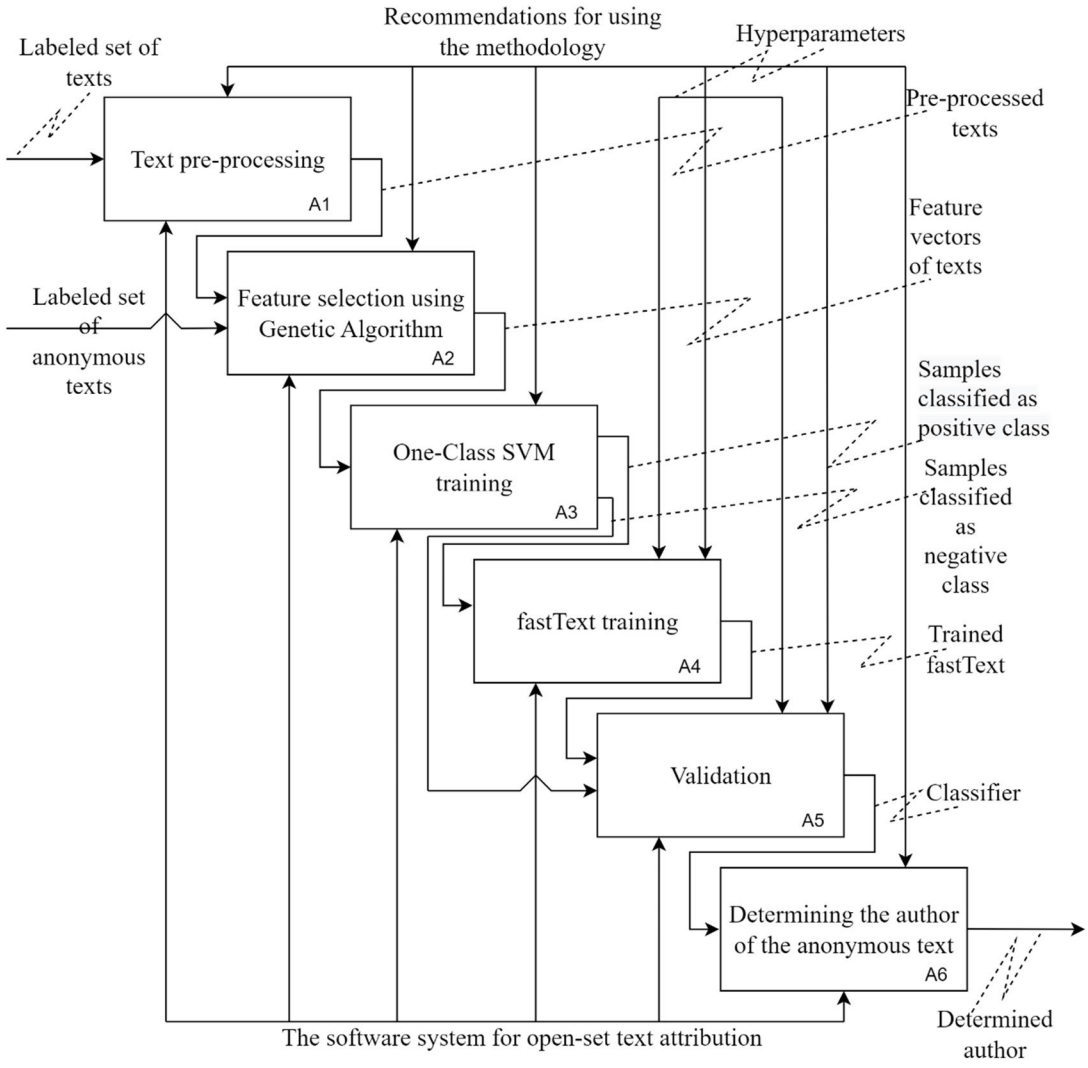

FIGURE 1. IDEF0 diagram of the authorship attribution process (previous methodology).

ten users meeting these criteria. Each dataset comprises 1400 posts per user. Subsequently, the experiments involved extracting the most active users, identified by their texts having the highest scores. The methodology is presented in two variations: 1) feature selection based on stylistics and vocabulary; and 2) a transformer-based approach. A Term frequency–inverse document frequency (TF-IDF) representation is computed for each text, and lexical and stylistic features are derived using the kernel-inspired encoder with a recursive mechanism for interpretable trees (KERMIT) [7]. These features are then fed into the SVM input. Consequently, when employing SVM with a pre-trained feature space, the results for 3, 5, and 10 authors were 85%, 75%, and 59% accuracy, respectively.

The article [8] delves into the application of authorship analysis techniques in examining posts on web forums linked to illegal groups. The researchers employed C4.5 and SVM models to analyze Arabic posts found on forums associated with recognized illegal entities. For each text, a feature space was generated, incorporating 301 features. This encompassed 87 lexical features (word and character-based), 158 syntactic elements (word and sentence structure, punctuation, word roots), 45 structural components (word-based), and 11 content-specific characteristics (author's race, nationality, degree of aggression). Through a series of experiments involving five authors, 20 messages per author, and cross-validation, the results showcased a remarkable level of author's pattern identification. The SVM classification achieved an accuracy of 90.8%, while C4.5 achieved 88.3%.

In their paper [9], T. Litvinova delves into the application of authorship analysis in the case of online security. T. Litvinova explores authorship analysis in the context of online security using real data from Russian forums. Their research focuses on authorship attribution in destructive con-

tent threads. The choice of 20 active participants from a vast dataset of over 1000 authors yielded an accuracy of 48.8% using LinearSVC with an l2 penalty. The authors undertake a study utilizing actual data sourced from Russian forums, employing diverse n-gram characteristics. This investigation centers around the authorship attribution of posts on the thematic forums containing destructive content. The study harnesses data from the forum dump, made available through the AZSecure-data [10]. A significant aspect of the research involves analyzing an extensive array of threads within the forum. To facilitate experimentation in attributing authorship across both similar and distinct topics, the authors compiled a dataset featuring the posts of 20 active participants. Selecting these 20 active participants posed a distinct challenge, as the original dataset spans more than 1000 authors and encompasses 699,000 posts, many of which pertain to mundane topics unrelated to destructive topics. To model this, LinearSVC with an l2 penalty was employed. The resulting accuracy was 48.8% for the dataset of 20 authors.

The paper [11] introduces an approach, leveraging Transformer architecture to merge stylistic and semantic analyses. The approach seamlessly merges two distinct types of analyses: stylistic, which centers on the writing style of the author, and semantic, which delves into the contextual meaning of the texts. In the PAN 2018 dataset, the proposed approach achieved an F1-score of 86%, surpassing the best-performing baseline method. Similarly, on the Amazon reviews dataset, the proposed approach achieved an accuracy of 98%.

The paper [12] introduces an approach based on the Graph-Based Siamese Network architecture. Subsequently, these feature vectors are compared using a graph-based methodology, utilizing a similarity matrix to calculate a weighted sum of the feature differences. The authors assessed the effectiveness on two datasets, encompassing the PAN 2018 dataset and a collection of Enron employee emails. In specific terms, the proposed method achieved an impressive F1-score of 95% on the PAN 2018 and 96% on the Enron.

In the paper [13], the authors collected data from the Enron email corpus, the HUMAINE dataset, and the Twitter message collection. They then harnessed various author profiling features, such as word frequency, punctuation usage, and sentiment analysis, to represent the unique traits of each author. The Decision Tree classifier boasts an accuracy of 95% on the Enron email corpus, 87% on the HUMAINE dataset, and 73% on the Twitter dataset.

V. A. Minaev, A. V. Simonov and others conducted a number of studies aimed at identifying destructive content on social networks [14], [15]. By destructive content, the authors understand texts with calls to incite national and religious hatred, dissemination of information about narcotic and other substances harmful to humans, especially young people, materials containing pornography, including those involving minors, as well as propaganda of terrorist, extremist and other criminal acts.

The first study [14] discusses models for classifying text content and methods for its preprocessing in order to identify destructive influences in social media. The dataset includes comments from users of the social network VK, further divided into three samples: publications without a specific topic (1), publications about Islam that do not have an extremist orientation (2), publications about radical Islam (3). The main methods of text vectorization were studied and applied: Bag of Words, TF-IDF, Word2vec. The highest accuracy (97%) when solving the problem of recognizing destructive content is provided by the system integration of the Bag of Words vectorization algorithm, the principal component method for reducing the feature space, logistic regression or random forest as learning models. Among the shortcomings of the work, it can be noted that the authors do not describe according to which method the publications were determined as radical, moreover, the social network VK does not allow the publication of pornographic or extremist content. The step of dividing the dataset into samples is also unclear, since only general results are given. In continuation of the study [15], the authors provide conclusions about the feasibility of further application of BERT, but the implementation and application of the methods themselves are planned to be carried out as part of further work.

Mashechkin et. al. published works [16], [17] devoted to the detection of destructive information posted on the Internet.

The method of automatic annotation and selection of keywords [16] for searching information of destructive content in text message flows was tested on the dataset [10]. Since not all texts in this dataset relate to destructive topics, a hierarchical clustering algorithm was initially applied to form 10 clusters (military activities, religious topics, politics in the Russian Federation, cooking, information technology, cars, terrorism, etc.). Analysis of the dataset using clustering shows that most discussion threads contain a political component, including (in many clusters) potentially containing extremist information. Next, a method for identifying keywords with removing information noise was used, based on the use of non-negative matrix factorization for the matrix of terms of text message branches. Thus, for each branch, a set of 15 keywords was formed, a total of three datasets: NMF30 – a set of annotations – the original set, in which each annotation contains sentences covering 30% of the total relevance of all sentences, NMF10 – each annotation contains sentences covering 10% of the total relevance of all proposals, KWORDS – a set of keywords for each branch (about 30 words).

Based on the obtained keywords, machine learning methods were applied to counter destructive using information from the Internet [17]. A new approach based on two-stage pattern search is proposed. In this approach, instead of a traditional search query, a sample document is used, and the goal of the search process is to find documents and messages relevant to this sample. Keywords are extracted from the sample document based on orthonormal non-negative matrix

factorization and representation of words in the form of n-grams. The found keywords form a search query, and the returned results, which contain a lot of noise and errors, are already ranked by the proposed original method. The ranking is carried out in such a way that those documents in the search results in which the weights of the hidden topics of the sample document are maximum are more relevant. Based on the proposed filtering method, the dataset [10] was enriched with comments from VK users in which the selected keywords occur. Classification of texts into three groups (dangerous, non-dangerous, unknown) was carried out using logistic regression with L2 regularization, a CART-type decision tree, random forest and ensembles of decision trees based on gradient boosting XGBoost and LightGBM. During training, the optimal parameters of each algorithm were selected using stratified cross-validation. For logistic regression and gradient boosting, the optimal L2 feature regularization parameter was selected on a grid of possible parameters. For algorithms based on decision trees, the maximum depth of the trees and the minimum number of leaves at the top of each tree were selected; for ensembles, the optimal size of the ensemble was selected. In order to avoid enumerating all possible combinations of parameters and effectively select the next point in the hyperparameter space, the sequential optimization approach SMBO (Sequential Model-Based Optimization) was used. Accuracy has been found to vary from 88% to 94% depending on the model. The maximum result was obtained by XGBoost, the minimum result was obtained by the decision tree.

The authors of the article [18] presented a system for identifying malicious and offensive messages. The system is based on machine learning and converting texts into vector form based on TF-IDF with a combination of unigrams and bigrams. The developed corpus consisted of 7000 Bengali text comments collected from Facebook. All texts were manually marked by expert linguists. After labeling, the authors received a balanced dataset, where 5600 texts were used for training, and 1400 for testing. Logistic regression, decision trees, random forest, naive Bayes classifier and SGD were used to test the system. Using SGD with parameters loss = ''log'', penalty = ''l2'', learning_rate = ''optimal'', max_iter = 40, random_state = 0, which allowed us to achieve the highest accuracy of 85%.

Article [19] is focused on conducting automated analysis of destructive texts in order to automate the analysis process and reduce the time of text examination. The following machine learning algorithms were considered: SVM, KNN, naive Bayes classifier and recurrent NN. The dataset contains unstructured texts with previously known classes: destructive and neutral texts. The search was based on materials included in the prohibited list. The logs were discovered in Telegram messenger communities. In total, about 300 destructive texts were discovered. All texts were reduced to lower case and cleared of punctuation. Texts were also vectorized, but the coding method was not specified by the authors. The accuracy of the methods was 87%, 53%, 53% and 45% for recurrent neural network, KNN, naive Bayes classifier and SVM, respectively.

The definition of destructive ideas in textual content in the Kazakh language is discussed in [20]. The authors have developed their own corpus, which includes a sample of destructive texts of 3000 words and 15,000 words of non-destructive or news texts. The source of the texts is the social network VK. Manual marking of texts made it possible to divide the collected comments into 2 classes: texts containing and not containing destructive ideas. Then tokenization, stemming and clearing of stop words were carried out. In order to classify texts, experiments were carried out with various machine learning models (random forest and gradient boosting) and vectorization methods (word2vec and TF-IDF). The obtained accuracies were 89%, 87%, 85% and 83% for the combinations gradient boosting + word2vec, random forest + word2vec, gradient boosting + TF-IDF and random forest + TF-IDF, respectively.

III. METHODOLOGY

The IDEF0 diagram of the presented experimental methodology process is presented in Figure 2.

The highlighting in red has been added to explicitly reflect the differences of the methodology compared to the previously developed methodology [2].

A. DATA COLLECTION AND TEXT PRE-PROCESSING

In the context of scientific analysis, it is crucial to underscore the fundamental differences between English and Russian, which significantly affect the process and quality of machine translation. English is classified as an analytical language, meaning that semantic relationships between words in a sentence are primarily expressed through a strict word order and the use of auxiliary words. In contrast, Russian is a synthetic language, where semantic relationships are expressed through changes in word forms (declension, conjugation), allowing for a more flexible arrangement of words within a sentence.

These linguistic features introduce certain difficulties into the translation process, especially when it comes to texts with non-standard spelling and grammar, such as user comments. Such texts often retain unique authorial characteristics, which may include specific terminology, wordplay, or even errors intentionally left by the author to convey a certain mood or emotion. In machine translation, these nuances can be lost or distorted as translation algorithms strive for grammatical and semantic correctness, which can lead to the loss of the author's style and individuality of the text. Therefore, we collected texts originally written in Russian.

Datasets of user comments were obtained from messengers Telegram and VK, as these platforms are the most popular for communication among Russian speaking citizens, cover all age categories, contain a variety of messages on different topics, and are therefore suitable for the task at hand, ensuring a certain level of symmetry in representation across different demographics and topics. The collection of a dataset of

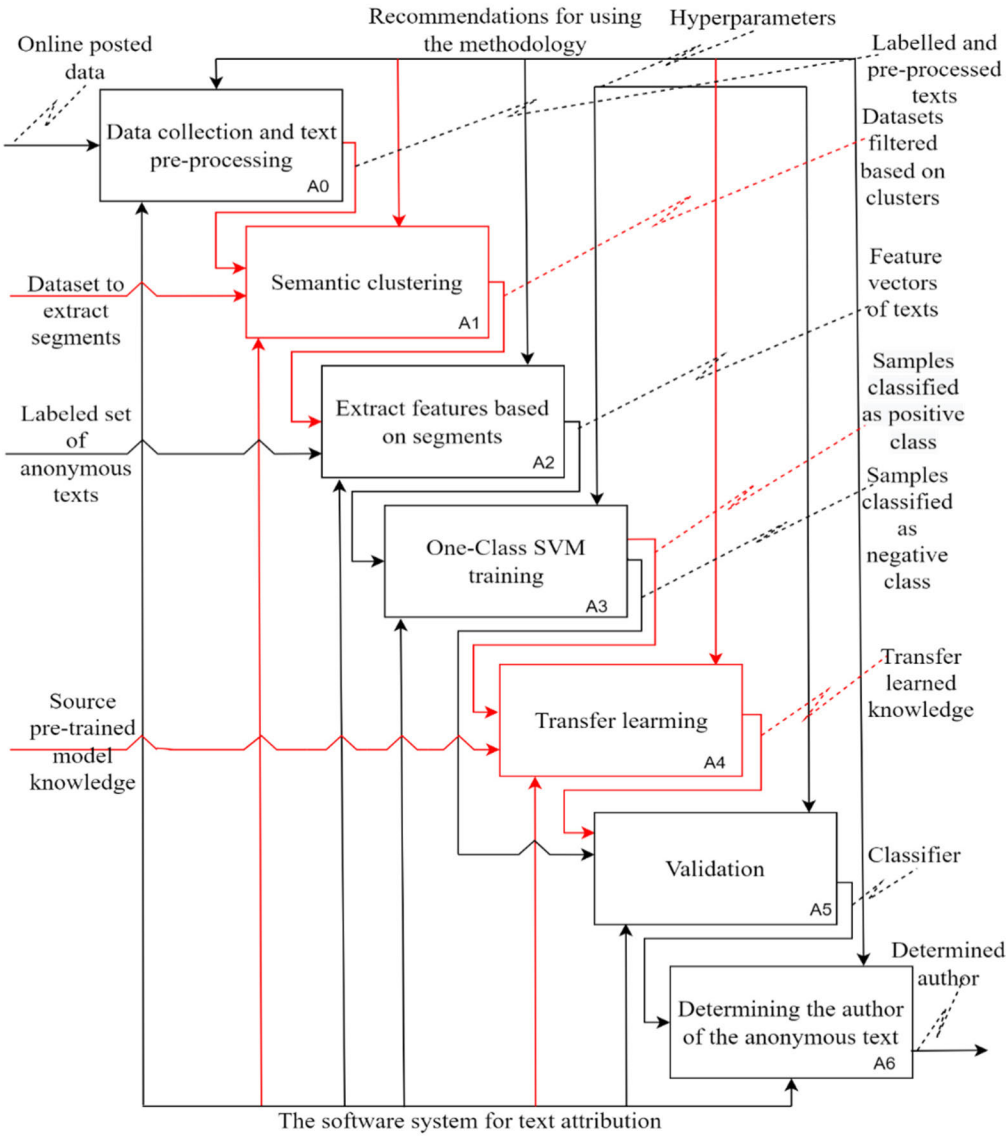

FIGURE 2. IDEF0 diagram of the authorship attribution process.

comments from Telegram containing destructive content included the following steps:

1) Channel identification. The first step is to identify and collect information about groups that publish both destructive and news content.
2) Creating a list of channels. Based on the gathered data, it is necessary to compile a list of groups to be included in the dataset.
3) Collecting comments. The API of the selected messaging platforms was used to collect user comments.

A dataset of comments from thematic forums containing destructive content was also used [10]. This forum includes various thematic pages where standard household issues such as repair, cooking, leisure, entertainment, and travel, as well as topics prohibited by law.

For research purposes, an additional dataset of short texts containing artificially generated samples based on the original dataset of VK comments was created. The generation was performed using the gpt-3.5-turbo model. To make the generated texts suitable for further work, the following constraints were imposed on the generation process:

1) Only texts of 50 or more characters were used as training samples.
2) Only texts in Russian were used.
3) The use of emoticons in the generated samples was allowed if they were present in the source texts.
4) Prohibition on distortion of the author's writing (fixing spelling and punctuation errors), i.e., if the original text contained such distortions, the generated sample must also contain them.
5) If in the original text, the author does not put dots at the end of the sentence, then the generated samples should not contain them either, and vice versa. Add the prefix ''ai'' to the author's identifier so that in

the future, artificial texts can be mixed with natural-language texts.

Through generation, 2062 texts by 216 authors were obtained.

It is important to emphasize that when using generated texts for authorship identification, it is essential to ensure their appropriateness and alignment with the authentic author's style. Among the commonly used similarity metrics for text, cosine similarity remains widely recognized. This metric assesses the similarity between two vectors within a multidimensional space. In the context of text, each text cluster can be represented as a vector, with each dimension corresponding to the frequency of a specific term in the document. To enable a comparison between the original and generated messages based on user ID, the following steps were taken:

1) Create a list of sources and artificial messages for each user ID.
2) Transformation of each message list into a TF-IDF matrix.
3) Calculation of the measure between every pair of source and generated messages for each user ID.
4) The final step entails computing the average cosine similarity for all message pairs within each user ID. This measurement provides an overall indication of message similarity for each user.

Cosine similarity (9) measures the angle between the representation vectors of texts and can serve as an indicator of their semantic similarity. It is worth noting that cosine similarity is only one of many possible similarity measures, and its effectiveness depends on the nature of the data being compared. However, such a solution is widely used in natural language processing and has proven to be effective in many text-processing applications [21], [22], so this method is chosen:

$$cos(\theta) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{||\boldsymbol{A}||\,||\boldsymbol{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (9)$$

where **A** and **B** are vectors that need to be compared.

To more fully assess the quality of text generation, a text syntax evaluation approach was used: analyzing sentence structure in the original and generated texts using metrics that assess grammatical correctness, structural integrity, and fluency of expression in the text, such as *BLEU* (Bilingual Evaluation Understudy) (10) [23] or *ROUGE* (Recall-Oriented Understudy for Gisting Evaluation) (11) [24].

$$BLEU = min(1, \frac{output\_lenght}{reference\_lenght}(\prod_{i=1}^{n} precision)^{\frac{1}{n}} \quad (10)$$

$$ROUGE = \frac{count_{match}(gram_n)}{count(gram_n)} \quad (11)$$

where *output_length* and *reference_length* — lengths of origin and generated samples, $count(gram_n)$ is the number of n-grams in samples, and $count_{match}(gram_n)$ is the number of n-grams in generated reference.

TABLE 1. Statistical characteristics of all described datasets.

Statistic	Telegram destructive	Telegram mass media	VK	AZSecure-data [10]
Number of authors	3765	101,163	2962	7125
Number of texts	43,364	437,532	202,892	699,981
Symbols in the dataset	1,604,468	2,688,428	30,652,109	73,467,890
Words in the dataset	401,117	1,119,238	4,708,619	11,294,114
Sentences in the dataset	80,222	352,456	767,134	2,388,431
Average length of a text in symbols	37	29	151.1	180.2.
Average length of a sentence in words	4.7	6.3	23.7	47.6
Average number of author's texts	44	61	102	86

TABLE 2. Assessment of similarity between original and generated samples.

Metric	Value
Cosine similarity	0.7
BLEU	0.58
ROUGE	0.61

The metrics were counted per author for his generated and original samples, and then the average across all authors was obtained (Table 2).

The value of metrics equal to 1 indicates that the texts are identical, while 0 indicates that the texts have no similarity. In our case, the generated texts contain synonymized content based on the original messages. The obtained values of *BLEU* 0.58 and *ROUGE* 0.61 are acceptable and indicate the similarity between the original and generated texts. The cosine similarity value of 0.7 indicates a high degree of similarity.

The result is a dataset containing 102 authors with 10 and more texts by each author (see Table 3 for more details).

The process of text preprocessing is minimal and includes lower case conversion, formatting of whitespace characters, removal of letters that do not belong to the Russian alphabet.

B. SEMANTIC CLUSTERING

The process of text clustering represents a method for categorizing texts into distinct groups. This method yields clusters, which are collections of texts that exhibit the highest degree of similarity. Unlike classification, which assigns new, previously unclassified texts to predefined groups based on characteristics learned from training data, clustering operates as an unsupervised method, lacking predefined labels for grouping. In contrast, classification relies on predetermined

TABLE 3. Statistical characteristics of the generated dataset.

Dataset characteristic	Value
Number of authors	102
Number of texts	1742
Symbols in the dataset	155,690
Words in the dataset	31,138
Sentences in the dataset	4448
Average length of a text in symbols	66
Average length of a sentence in words	6.1
Average number of author's texts	43
Number of authors	102
Number of texts	1742

groups and the number of these groups is fixed in advance. Clustering, therefore, allows for the dynamic identification of text groups based on inherent similarities without prior knowledge of group definitions.

Semantic clustering, a specific approach within text clustering, enhances this process by focusing on the meaning and contextual relevance of the texts. By incorporating semantic understanding, clustering goes beyond mere lexical analysis, allowing for a more nuanced and insightful grouping of texts. This approach is especially useful in fields such as sentiment analysis, topic modeling, and document summarization, where the deep semantic relationships between texts are crucial for accurate analysis. It should be noted that clustering algorithms are diverse and can be used not only as a stand-alone solution to the problem of dividing texts into clusters but also as an auxiliary tool to improve the efficiency of classification.

Many clustering methods are effective. These include the well-known k-means methods [25], latent Dirichlet algorithm (LDA) [26], Brown's clustering [27], and many others. However, within the scope of this study, more complex neural network methods are of most interest.

It was decided to use BERTopic [28] as a clustering tool. BERTopic is well established as an optimal solution for clustering texts based on their semantic proximity. Distinctive features of BERTopic are the ability to obtain vector representations, token frequencies, and weights, manage clusters-segments dimensions, and pre-train on custom data. Figure 3 shows a UML activity diagram illustrating the process of semantic clustering using BERTopic. A detailed description of the actions described in the diagram is given below the figure.

The first stage of BERTopic's operation is to create vector representations of the input texts. These representations are generated using SentenceTransformer [29]. It learns from a large corpus of text and considers the context of sentences when calculating these representations. By employing SentenceTransformer, BERTopic ensures a high level of semantic similarity between sentences.

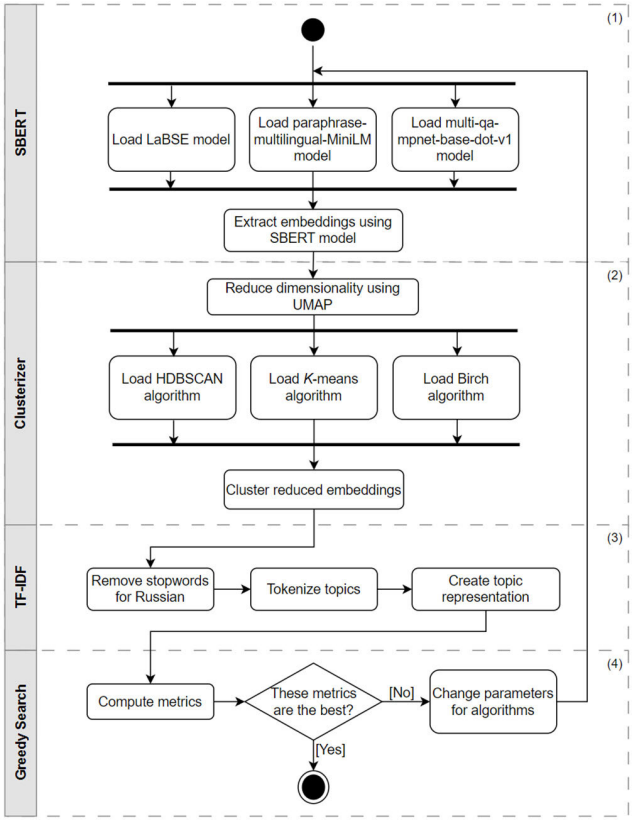

FIGURE 3. Clustering using BERTopic.

In the second stage, clustering takes place. After generating vector representations using SentenceTransformer, the dimensionality of these representations is reduced using the Uniform Manifold Approximation and Projection (UMAP) algorithm [30]. UMAP enables BERTopic to maintain the global data structure while reducing dimensionality, which is valuable for subsequent clustering. Following this dimensionality reduction, a specific clustering algorithm is applied to the UMAP-derived representations [31]. BERTopic supports various algorithms, including HDBSCAN, K-means, Birch, and others, for grouping sentences into topics.

To select the representations of each topic and rank them in the third stage, BERTopic uses TF-IDF and MMR (Maximal Marginal Relevance) methods. TF-IDF evaluates the importance of words in a sentence, while MMR takes into account both the topic relevance of a sentence and its diversity relative to already selected sentences. Since BERTopic is a flexible approach in all aspects, an important task is the correct selection of algorithms and their parameters at each of the described stages of work. The greedy search algorithm was used to automatically select optimal algorithms and their parameters. As a dataset for evaluating the performance of various combinations of BERTopic algorithms, a Russian-translated dataset from the Kaggle [32], content that encourages violence or illegal activity was used.

The quality of clustering was determined by two metrics that do not depend on ground truth.

The silhouette_score [33] (12), (13) metric is used to calculate the distance between an object and the nearest cluster of which the object is not a part. The best value of the metric is 1, and the worst value is −1. Values close to 0 are an indication of overlapping clusters, while negative values indicate that the sample has been assigned to the wrong cluster:

$$silhouette\_score = \frac{\sum_{i=1}^{n} silhouette\_score(i)}{n} \quad (12)$$

$$silhouette\_score(i) = \frac{max(a_i b_i)}{b_i - a_i} \quad (13)$$

where a_i — the average distance of i to all other data points in the same cluster (intra-cluster distance), b_i — the average distance of i to all data points in the nearest cluster (inter-cluster distance).

The davies_bouldin_score metric [34] (14), (15), (16) defines an average measure of the similarity of each cluster to its most cluster segment, where similarity is the ratio of the distances within a cluster to the distance between clusters. The minimum value of the metric is 0, with lower values indicating better clustering:

$$davies\_bouldin\_score = \frac{1}{n} \sum_{i=1}^{n_i} R_i \quad (14)$$

$$R_i = max\left(R_{ij}\right), i, j = 1..n \quad (15)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (16)$$

where s_i — the within cluster scatter for cluster i, which has to be as low as possible, s_j — the within cluster scatter for cluster j, which has to be as low as possible, d_{ij} — the separation between the i-th and the j-th cluster, which ideally has to be as large as possible.

It was decided to try several combinations of algorithms included in BERTopic. Multilingual multi-qa-mpnet-base-dot-v1, paraphrase-multilingual-MiniLM-L12-v2 and LaBSE [35] were selected as models used by Sentence-Transformer to build representations. Clustering algorithms include K-means with Elbow method, Birch and HDBSCAN, which allow us to determine the optimal number of clusters automatically. The algorithm parameters were selected using greedy search. The experimental results are presented in Table 4.

The best result was achieved by combining LaBSE for obtaining vector representations and the HDBSCAN algorithm for clustering. Parameters selected for this combination using greedy search: UMAP: n_neighbors=15, n_components=3, metric=jaccard; HDBSCAN: metric=Euclidean, cluster_selection_method=eom.

The silhouette_score metric for the experimentally selected settings was 0.76. The davies_bouldin_score metric was 0.32. This result indicates that the clusters do not overlap and have a sufficient level of separation. A visualization of UMAP clusters obtained by BERTopic is shown in Figure 4, and examples of word clouds corresponding to the top 10 words for the selected clusters are shown in Figure 5.

TABLE 4. Results of experiments with BERTOPIC.

Metric	Sentence transformer	Clustering algorithm		
		K-means	Birch	HDBSCAN
silhouette score	multi-qa-mpnet-base-dot-v1	0.22	0.42	0.59
	paraphrase-multilingual-MiniLM-L12-v2	0.36	0.55	0.74
	LaBSE	0.33	0.54	**0.75**
davies-bouldin score	multi-qa-mpnet-base-dot-v1	0.71	0.58	0.69
	paraphrase-multilingual-MiniLM-L12-v2	0.6	0.49	**0.32**
	LaBSE	0.55	0.43	**0.32**

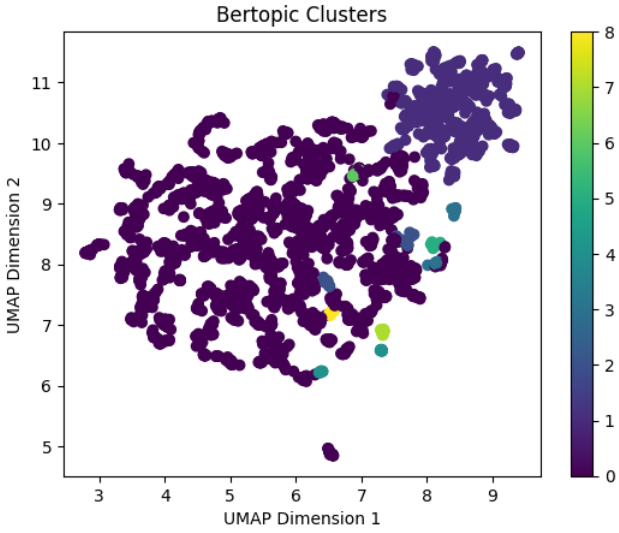

FIGURE 4. UMAP visualization.

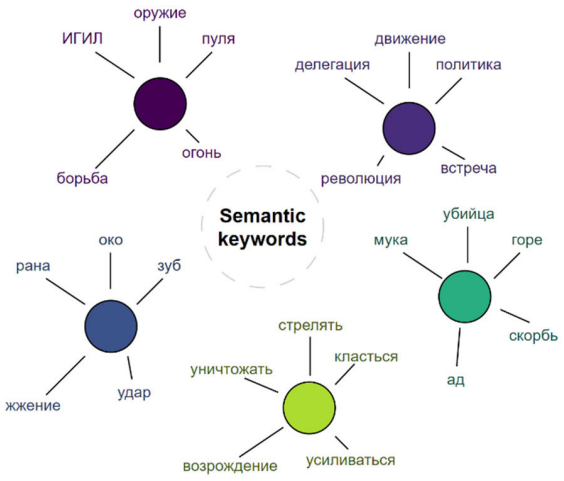

FIGURE 5. Examples of text clusters.

It is worth noting that the use of semantic clustering makes it possible to obtain informative keywords for Russian-language destructive texts without manual marking and the involvement of linguistic experts.

C. EXTRACT FEATURES BASED ON CLUSTERS

Based on the obtained clusters, thematic texts' authors and their texts from the $\mathbf{T}$ dataset were selected. Let us denote $\mathbf{K} = \{k_1, k_2, \ldots, k_v\}$ is the set of keywords obtained from the clusters. For each text, the presence of the extracted clusters was checked. In case of presence of 3 or more keywords from the clusters in the text, such text was included in the sample forming the final dataset of destructive texts $\mathbf{T}'$(17, 18):

$$count\_keys(\boldsymbol{T}, \boldsymbol{K}) = count(t|k \in \boldsymbol{K}, t \in \boldsymbol{T}) \tag{17}$$

$$\boldsymbol{T}' = text|text \in \boldsymbol{T}, count\_keys(text, k) \geq 3\} \tag{18}$$

Therefore, semantic clustering enhances the quality of thematic datasets, particularly those that are destructive, by selecting texts that contain keywords from a reference dataset. Moreover, this technique proves beneficial for filtering large volumes of data and identifying texts appropriate for training models. Through the focused selection of relevant texts, semantic clustering ensures that datasets are not only more coherent but also more aligned with the specific themes of interest. This precision is crucial for developing models that are both accurate and efficient in processing and analyzing text-based information.

D. ONE CLASS SVM

There are two steps in this method. The formation of a feature vector and the selection of informative features in accordance with the characteristics of the Russian language is given in [2]. The texts are initially supplied as a feature vector into the input of a one-class support vector machine (SVM). This approach was chosen because to the One-Class SVM's capability to identify anomalies. The authors refer to texts in the negative class as oddities. When the test set contains anomalies (negative class samples) and the training set's data has a normal distribution, anomaly detection is possible. One-Class SVM constructs a non-linear space based on regular and additional observations, with the anomalous data being cut off by a boundary. Decide whether the input data are from one of the well-known authors whose texts were utilized for both training and testing (positive classes) by performing the step.

E. TRANSFER LEARNING

Since the conduct of this study differs from the standard task of determining authorship (because of using destructive texts), an additional stage of transfer learning was introduced into the methodology. The goal of applying transfer learning is to use a source model trained for the task of analyzing destructive content. The effectiveness of using transfer learning has been proven in many natural language processing tasks [36], [37], [38]. The hypothesis for implementing transfer learning is that using the knowledge of the source model when training the target one (to determine authorship) will improve the accuracy of the study when working with destructive content.

At the stage of training the model, a transfer learning technique based on a classifier previously trained to identify sensitive topics (propagation of crime, drugs. radical politics, suicide) was applied.

When searching for a model, the limitation was the use of the Russian language. The second constraint was the ability of the model to distinguish negative connotations of texts (psychological pressure from calls to illegal actions). The following models were identified as fitting the described criteria [39], [40], [41].

Among the options presented, the model [42], described by the authors in the article [43], because in addition to the standard division according to the emotional tone of the text, the authors carried out a more in-depth work related to highlighting sensitive topics in the text, e.g., terrorism, religious, crime, politics. shaming. drags, racism, sexism.

Such topics are found in destructive texts, making the model applicable to this task in the transfer learning step. The authors trained the model on a dataset of 82,000 manually labeled Russian-language short comments. The use of such data in training the model can also be attributed to its advantages to the destructive text detection or authorship attribution task since the datasets that were used also contain short texts (see Section IV). When training the model, the authors tuned the pre-trained RuBERT 2019 model on their own data.

RuBERT is a language model developed by DeepPavlov, and it is based on the Google BERT architecture. BERT, which stands for ''Bidirectional Encoder Representations from Transformers'' is a transformer-based model designed for natural language understanding tasks. Google initially introduced BERT as a pre-trained model for English text, but it has since been adapted and fine-tuned for various languages, including Russian. RuBERT is specifically trained and fine-tuned on Russian text data, making it a useful tool for a wide range of natural language processing tasks in the Russian language, such as text classification, sentiment analysis, named entity recognition, and more. DeepPavlov is a research initiative that focuses on natural language processing and conversational AI, and they have contributed to the development of RuBERT to serve the Russian-speaking NLP community. The model was used in Deeppavlov's implementation. Authors did not provide any information about setting specific values for the hyperparameters of the neural network. Our implementation of transfer learning is presented in Figure 6. Here is a brief description of each step:

1) Load pre-trained BERT model. Load a pre-trained BERT model that has been fine-tuned on a relevant task: classification sensitive topics.
2) Extract BERT embeddings for text samples. Use the pre-trained BERT model to extract embeddings for the text samples in target dataset.

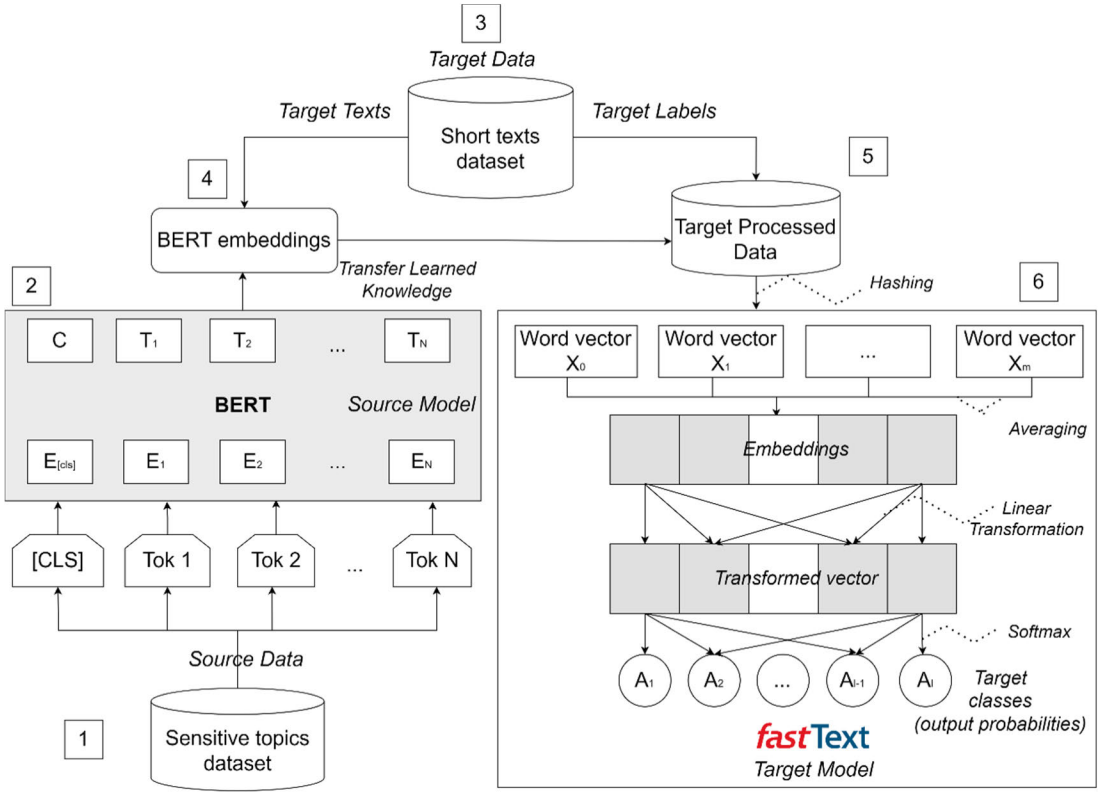

FIGURE 6. Transfer learning process.

3) Load and preprocess authorship attribution dataset. Import the authorship attribution dataset, which encompasses columns for ''label'' (the author's name or identifier) and ''text'' (the text samples). Preprocess this dataset to extract labels and texts.
4) Convert text samples into BERT embeddings. Transform the text samples into BERT embeddings to facilitate further analysis.
5) Combine BERT embeddings with labels. Combine the BERT embeddings with their corresponding author labels. Format the data for fastText, including labels with the __label__ prefix.
6) Train a fastText classifier previously recognized as the optimal solution for authorship detection among other deep neural network models and classical machine learning methods considered with hyperparameters selected in our previous study [1] on labeled BERT embeddings. The model predicts the author of the test text based on the learned embeddings.

This structured approach ensures a systematic application of transfer learning to enhance the accuracy and efficiency of authorship attribution, leveraging the strengths of both BERT and fastText technologies.

In the output layer, while addressing the task of authorship classification, the count of outputs within this layer will align with the number of authors (classes), which could be 2, 5, 10, or 20. The selected activation function, suitable for multi-class classification, is Softmax.

On the other hand, for the destructive content detection task, the output layer will consist of two neurons representing the categories of destructive or non-destructive text. The activation function employed in this context will be sigmoidal.

Here is a pseudocode of the transfer learning process (Algorithm 1):

F. VALIDATION AND DETERMINING AN ANONYMOUS AUTHOR

When evaluating results in this context, accuracy measures the proportion of correctly classified samples (texts) out of the total number of samples in the test sample. Cross-validation on 5 folds was also applied in the experiments to avoid excessive variation in accuracies.

For destructive content detection, the classification was binary (destructive/non-destructive text) and the experimental setup encompassed the utilization of 10 to 1000 samples per class. When determining authorship, the experiments were conducted for 2, 5, 10, 20 and 50 authors in the case of closed attribution and 2+1, 4+1, 9+1, 19+1 and 49+1 in the case of open attribution. The number of samples per class was 50.

IV. RESULTS

Subsection IV-A contains statistical tests of data quality; subsection IV-B contains the results of experiments on destructive content detection; subsection IV-C contains the results of experiments on text authorship attribution in the

Algorithm 1 Transfer Learning

```
Set dataset_path: > path_to_dataset
Set model_path: > path_to_bert_model
Set embedddings: > []
Set fastText_model: > FastText
Set data: > []
Set number_of_ngrams: > 3
Set learning_rate: > 0.6
Set dimensions: > 500
Set loss_function: > ''ova''
 Set max_number_of_segments: > 2,000,000
 Set cross_validation_results : > []

 procedure Load_pretrained_bert_model(model_path):
    BERT_model=Load_model(model_path)
    Freeze_Layers (BERT_model)
    return BERT_model

 procedure Load_and_preprocess_dataset(dataset_path):
    dataset=Load_dataset(dataset_path)
    labels=dataset[''labels'']
    texts.remove_extra_spaces()
    texts.lower()
    return labels, texts

 procedure Get_bert_embeddings (texts):
    bert_embeddings=embeddings
    for each in texts:
       tok_texts=bert_tokenizer(texts)
       Bert_embeddings=bert_embeddings(tok_texts)
    return bert_embeddings

 procedureCombine_label_embedding(label, embedding):
    embedding_string=string(embedding)
    formatted_example=label + '' '' + embedding_string

 procedure Save_data_to_fasttext_format(data, file_path):
    file_path.open()
    for each formatted_example in data:
       file_path.write(formatted_example)
    file_path.close()

 procedure Set_fastText_parameters(**parameters):
    fastText_classifier(parameters=**parameters)
    return fastText_classifier

 Set BERT_model : > Load_pretrained_bert_model()
 Set labels, texts : > Load_and_preprocess_texts(dataset)
 Set bert_embeddings: > Get_bert_embeddings(texts)
 begin
 for each_fold in folds:
    for each_label, embedding pair in zip(training.labels,
training.embeddings):
       formatted_example=Combine_label_and_embedding
(label, embedding)
       data.append(formatted_example)
    parameters=learning_rate, dimensions,
max_number_of_segments
fastText_cls=Set_fastText_training_parameters(**parameters)
    fastText_cls.train(data)
    validation_accuracy=fastText_cls.evaluate(validation_data)
    cross_validation_results.append(validation_accuracy)
 end
```

case of a closed set of candidates (subsection IV-C1) and in an open set (subsection IV-C2); subsection IV-D contains the performance evaluation of semantic clustering and transfer learning in comparison with the previously developed methodology [2].

A. DATA QUALITY

The first step in collecting datasets is to ensure that the sources are reliable and of good quality, and that there is sufficient data available. The study used comments from users of the VK social network, considered the most popular in Russia and the CIS. VK takes 2nd place in the ranking of the most popular social networks in terms of the number of Russian audiences. The VK social network is used by 85% of Russian Internet users, with 52% of them using the Russian social network every day. Therefore, such a source contains a sufficient amount of data.

The second source was the Telegram messenger. In the ranking of the most popular Android applications in Russia, Telegram is in 2nd place, in the App Store in 4th place.

The third source is the AZSecure-data dataset used in many studies of Russian-language content [10]. His choice was based on the ability to compare the results with other studies.

For generated texts, quality control included calculating the BLEU, ROUGE and cosine similarity metrics to ensure that the generated data was sufficiently similar to the original samples (Section III-C).

To assess the data quality, it was decided to initially conduct some checks to ensure there were no duplicates or blank values. Samples containing only emoji or containing text other than Russian were removed. Subsequently, the Chi-square test was employed to examine the homogeneity of the data. The null hypothesis (H0) posits that the data distribution is homogeneous (with no significant difference), while the alternative hypothesis (H1) suggests that the data distributions are heterogeneous (exhibiting a significant difference).

The results of the test for each dataset separately and their unions are summarized in Table 5. Highlighted p-values greater than the significance level of 0.05 in bold. The p-value represents the probability of obtaining the same chi-square statistic as calculated if the H0 were true. If the p-value is less than or equal to the significance level, H0 is rejected, indicating a significant difference in the distribution of texts or patterns between groups. If the p-value is greater than the significance level, H0 should not be rejected, suggesting homogeneity.

Thus, since the p-values are 0.31, 0.45, 0.58, 1.24. 0.53 for each of the datasets separately exceeds the significance level (0.05), H0 is accepted — the data distribution is homogeneous. Also, H0 is accepted for the case of combining the VK and VK generated, Telegram destructive texts and Telegram mass media datasets (p-values 0.37 and 0.33, respectively), which indicates the homogeneity of the data in the combined datasets. In all other cases, H0 is rejected and H1 is accepted, which means that the datasets should not be combined when conducting experiments into a single dataset.

TABLE 5. Chi-squared test results (*p*-values).

	VK	VK generated	Telegram destructive texts	Telegram mass media texts	AZSecure-data [10]
VK	0.31	0.37	**0.0007**	**0.001**	**0.02**
VK generated	0.37	0.45	**0.004**	**0.03**	**0.01**
Telegram destructive texts	**0.00 07**	**0.00 4**	0.58	0.33	**0.02**
Telegram mass media	**0.00 1**	**0.03**	0.33	1.24	**0.04**
AZSecure-data [10]	**0.02**	**0.01**	**0.02**	**0.04**	0.23

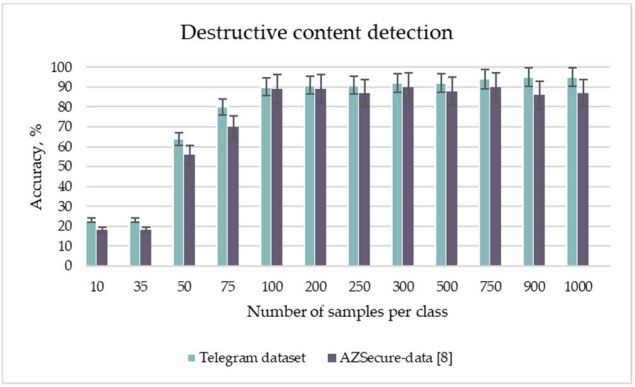

FIGURE 7. Results of the experiment aimed at destructive texts detection.

B. DESTRUCTIVE CONTENT DETECTION

In determining whether a text contains content that may encourage or promote destructive behavior, a dataset that combines the Telegram Mass Media and Telegram Destructive Texts datasets collected from Telegram was utilized. In this dataset, Telegram Destructive content samples are labeled as containing destructive content, while Telegram Mass Media samples are labeled as not containing it. VK comments dataset was not used, as doing so could potentially have had a negative impact on the data's homogeneity (as discussed in Section IV-A).

A similar experiment was also conducted using the AZSecure-data dataset [10]. With the assistance of semantic clustering, a group of authors who left comments with illegal content was identified within the dataset, forming the first class. The second class consisted of texts from authors writing about ordinary topics. The experiment followed a similar procedure to the one conducted with the dataset of destructive texts from Telegram. The results are presented in Figure 7.

Based on the results, the accuracy of destructive content detection improves when more samples per class are employed. An accuracy of 90% is achieved with 100 samples per class for the Telegram dataset and 300 samples for the AZSecure-data dataset. When using only 10-50 samples per class, the classification accuracy does not exceed 62%.

TABLE 6. Results of experiments aimed at authorship attribution.

Number of authors	Accuracy, %				Training time, sec.
	Destructive content	Mass media	VK	AZSecure-data [10]	
2	93±3	93±3	93±3	93±3	1190-1344
5	94±5	94±5	94±5	94±5	3713-4232
10	96±3	96±3	96±3	96±3	6,420-6,709
20	87±4	87±4	87±4	87±4	9,196-9,545

In general, the results obtained on the AZSecure-data [10] are lower than when conducting similar experiments for a dataset with destructive content due to the less homogeneity of the data contained in AZSecure-data, which is proven in Data Quality section (IVA).

C. AUTHORSHIP ATTRIBUTION

1) CLOSED SET ATTRIBUTION

This section deals with authorship detection among Telegram users commenting on posts in communities on different topics (mass media and material containing information advocating violence or illegal activities).

The first experiment aimed at determining authorship is a classification within a group of users commenting on destructive content.

A similar experiment to the first one was conducted for authors commenting on Telegram news communities. This experiment is significant for two reasons: first, the need to compare the results obtained from datasets from different sources, and second, to test the hypothesis that the pre-trained models used for detecting sensitive topics in transfer learning can effectively detect commenters of individual authors expressing aggression. The second aspect arises from the fact that when commenting on news, people express opinions based on posted content, and these opinions often encompass different tones.

The results of this experiment are presented in Table 6.

When classifying 2 authors, the accuracy reaches 96%. With five authors, it is 89%, which indicates the model's ability to identify authors within a group of commentators in communities with similar topics.

With ten authors, the accuracy reaches 83%, indicating a high level of separability in the model. The lower results obtained for 20 authors can be attributed to the model's increased difficulty in identifying the true author within a larger number of classes, making the task more complex. Another factor contributing to these results is the presence of similar patterns in the writing style of the users used in the dataset. This could be due to the fact that the pre-trained model was designed to identify negatively toned texts, and in the case of destructive communities, most of the comments fall into that category. This similarity in tone can confuse the model when making a decision. Concerning the influence of

TABLE 7. Results of experiments aimed at open-set authorship.

Number of authors	Accuracy, %		Training time, sec
	Mass media and destructive texts	VK users' comments and generated comments	
2+1	92±8	95±7	1263-1921
4+1	84±8	91±7	3802-4419
9+1	71±9	79±6	6,844-7,450
19+1	51±9	61±5	9,327-10,227
49+1	21±8	18±5	19,154-20,133

thematic groups, it can be noted that the results vary by no more than 9% from one group to another with the same number of classes. This indicates that the developed methodology is independent of the choice of a specific thematic category. Based on the results obtained, it should be noted that there are no significant differences with the accuracies obtained for the dataset based on Telegram, which indicates the high generalization ability of the model.

2) OPEN SET ATTRIBUTION

This experiment on authorship attribution represents a more complex modification of the previous experiments. The complexity arises from the application of open attribution, where texts from authors commenting on destructive content are introduced only during the model testing phase. Consequently, for open attribution, training was conducted using the texts of N authors (where $N = 2, 4, 9, 19, 49$), and during testing, text samples from M authors were introduced. This means that the set of authors for training is N, while for testing, it becomes $N + M$. The division of the initial set N into training and test samples, as in the case of closed attribution, is carried out in an 80:20 ratio. Experiments were conducted for 2+1, 4+1, 9+1, 19+1 and 49+1 authors, where the second summand represents an additional negative class that includes authors whose texts are added during the testing stage. The number of added anonymous texts is 20% of the training sample.

The last experiment is related to the application of open attribution to identify artificially created content and bots in social networks, using VK as an example. At the testing stage, samples of generated comments were implemented. The frequencies of words that are elements of the clusters obtained in the semantic clustering step were used as features for One-Class SVM. Results are demonstrated in Table 7.

In the context of an open set attribution, the model exhibits a remarkable ability to identify the authorship of a given text with 96% accuracy. However, as the number of potential authors increases, the accuracy of author identification tends to decrease. This phenomenon can be attributed to the increasing complexity and diversity of writing styles and patterns within a larger set of candidates.

Notably, when it comes to mass media texts, the model demonstrates a relatively high level of accuracy, with a success rate of 95% (with a deviation of ±7%) in the case of 2+1 authors. Even in scenarios where there are 19+1 authors, the accuracy remains notably high at 61% (with a deviation of ±5%). This suggests that the model is particularly adept at recognizing and attributing authorship in texts created by professional media organizations and their contributors.

In contrast, when dealing with destructive texts, the model faces a greater challenge. In cases involving 2+1 authors, the accuracy drops to 84% (with a deviation of ±8%). For scenarios with 19+1 authors, the accuracy decreases even further to 51% (with a deviation of ±9%).

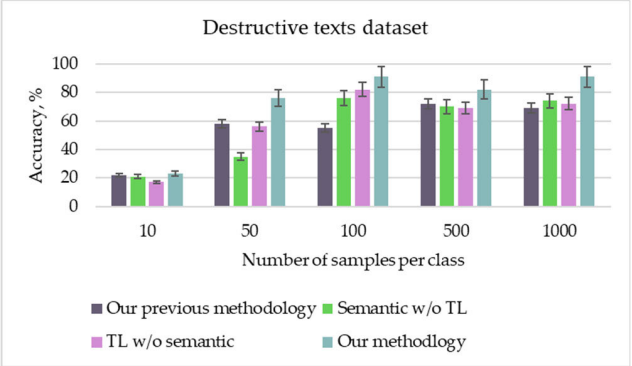

FIGURE 8. Destructive texts detection (Telegram destructive texts dataset).

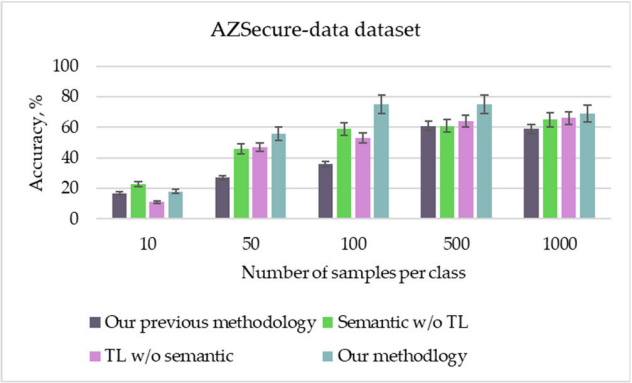

FIGURE 9. Destructive texts detection (AZSecure-data).

D. EVALUATION OF SEMANTIC CLUSTERING AND TRANSFER LEARNING IN COMPARISON WITH OUR PREVIOUS METHODOLOGY

The developed methodology is an evaluation of our previous methodology [2]. Let us demonstrate the advantages of semantic clustering and transfer learning, subsequently adding them to the source methodology and evaluating the results (see figures 8-15).

From the results, it can be observed that adding transfer learning enables us to achieve accuracy improvements of up to 10%, 4%, and 20% in the cases of closed and open author sets, and destructive content detection, respectively. The addition of semantic clustering results in improvements of 8%, 3%, and 16% for the same cases. The combined use of

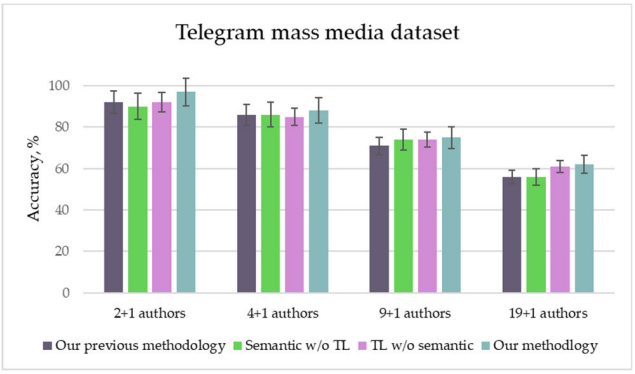

FIGURE 10. Open-set attribution (Telegram mass media texts dataset).

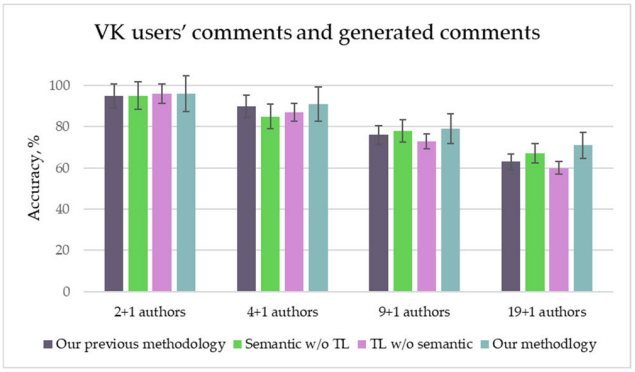

FIGURE 11. Open-set attribution (VK users' comments and generated comments).

both tools yields a combined accuracy improvement of 13%, 14%, and 22% for both authorship and destructive content detection cases.

V. DISCUSSION

A. COMPARISON OF THE RESULTS WITH RELATED WORKS

The most crucial aspect in evaluating the methodology is its efficiency, advantages, and limitations in comparison to analogs for a closed set of candidates (refer to Table 8) and for destructive content (refer to Table 9). It is worth noting that as of 2023, there are no studies available for the task of open attribution of Russian-language text, so a similar comparison cannot be conducted.

Most of the works dedicated to the problem of authorship attribution in the case of a closed set of candidates have primarily focused on a small number of authors. Our obtained accuracy values for such cases (2 and 5 authors) are 9-13% higher than the results presented by other researchers. In the case of 10 and 20 authors, our method demonstrates an accuracy improvement of up to 22% compared to other studies. For a more direct comparison, we used the same AZSecure-data dataset [10] as in the work of Litvinova, T. et al. [7] for experiments with 20 authors. The higher accuracy rate (62% vs. 49%) confirms the effectiveness of semantic clustering and transfer learning when contrasted with classical approaches that rely on manual selection of lemmas for dataset filtering.

TABLE 8. Comparison of the results of studies aimed at closet set attribution.

Authors	Dataset	Accuracy, %			
		2	5	10	20
Litvinova, T. et. al. [9]	AZSecure-data [10]	-	-	-	49
Abbasi, A.; Chen, H. [8]	Web radical forums	-	85	-	-
Ranaldi, L.; Ranaldi, F.; Fallucchi, F.; Zanzotto, F.M. [6]	DarkNet texts	85	75	59	-
Moshkin, M., Fadeev, D., Yarushkina, N. [43]	Materials prohibited for publication	90	-	-	-
Ours [1]	Short social media texts	93	87	**81**	62
Ours	Destructive content	96	87	80	59

TABLE 9. Comparison of research findings on identifying destructive content in text.

Authors	Dataset	Number of samples per class	Accuracy, %
Moshkin, M., Fadeev, D., Yarushkina, N. [43]	Materials prohibited for publication	150	84
Mussiraliyeva, S. et.al. [44]	Religious texts from prohibited forums	300	89
Kapitanov, A. et. al. [45]	Prohibited web forums	200	91
Ours	Telegram dataset	100	**92**
	AZSecure-data [10]	100	88

The developed methodology requires only 100 samples for each class to achieve an accuracy of 92%. In contrast, similar studies have achieved comparable results but with a significantly larger number of samples, ranging from 150 to 300. This makes our methodology more preferable for addressing real practical problems.

There are some advantages of the proposed methodology:

1. Fewer samples per class are needed to obtain results comparable to and superior to those reported in other studies.

2. When using the dataset [10], many scientists pre-filter the data before training in order to use only destructive texts. Litvinova et al. [9] use a number of templates for this purpose in order to subsequently identify destructive ones based on their presence in the text. The approach we propose is more flexible: thanks to semantic clustering, it is possible to obtain

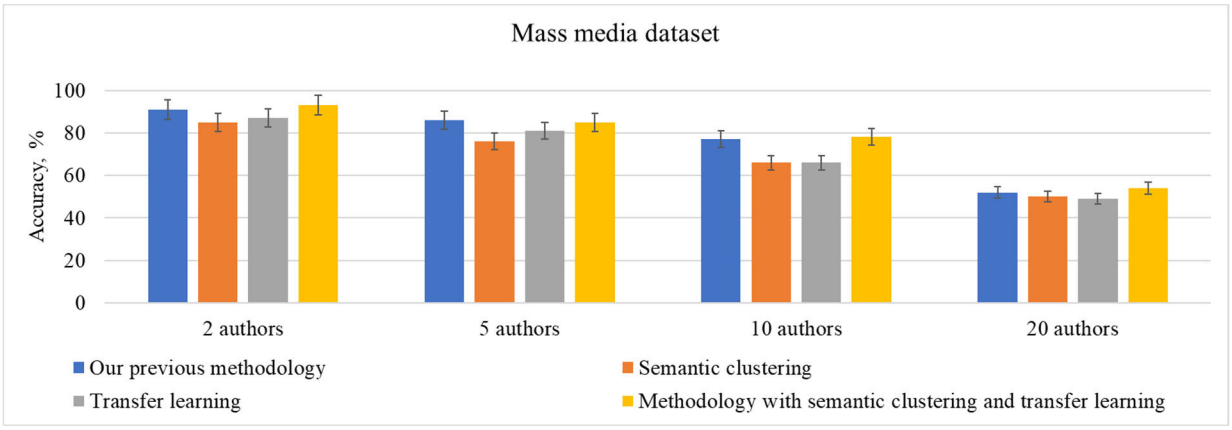

FIGURE 12. Closed-set attribution. Mass media dataset.

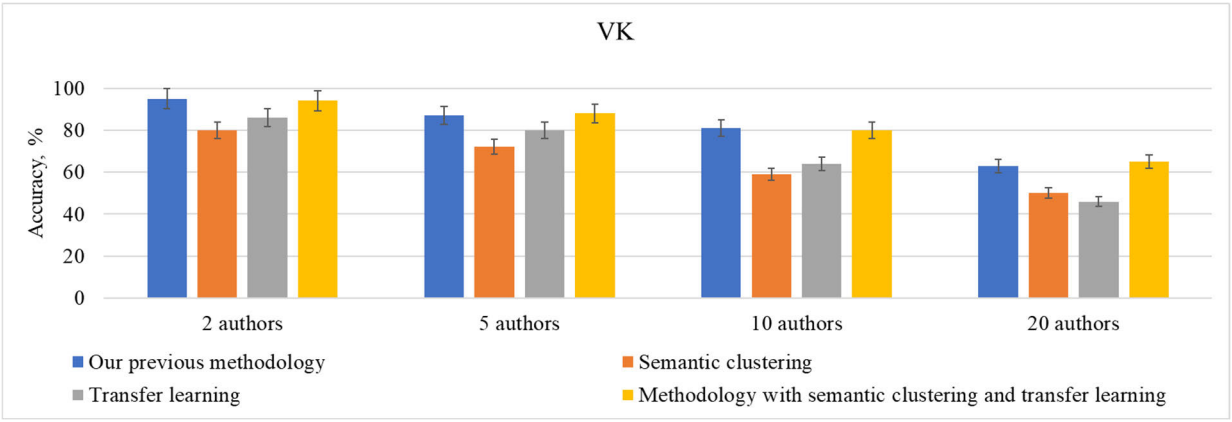

FIGURE 13. Closed-set attribution. VK dataset.

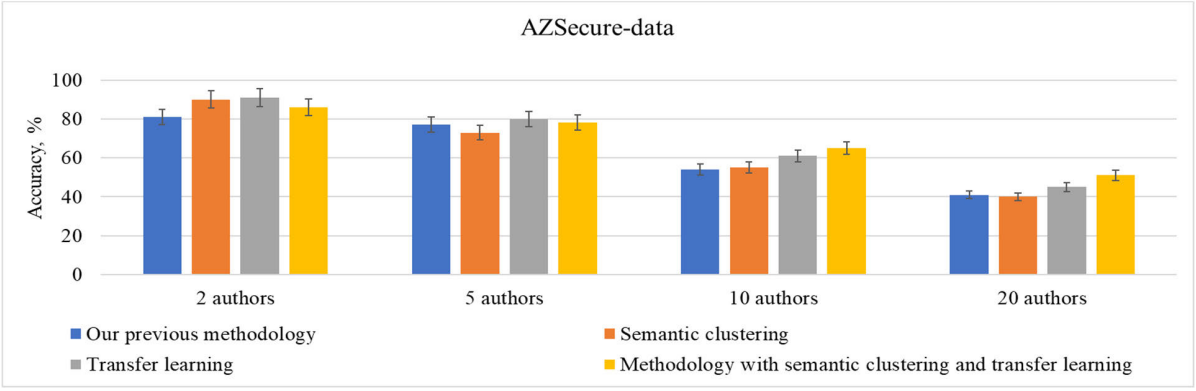

FIGURE 14. Closed-set attribution. AZSecure-data.

not only explicit lemmas for filtering, but also semantic keywords that may not be obvious during the initial analysis. In addition, this approach does not require manual marking.

3. Most studies generally do not include attribution in the case of 20 or more candidates. Our technique allows for similar experiments and is superior in accuracy to the results presented by other researchers.

4. The presented methodology allows conducting experiments for an open set of candidates, which is still a poorly studied area in Russian-language research.

5. The use of transfer learning makes the technique applicable for determining the author in the case of thematic texts due to the acquisition of knowledge using a pre-trained initial model.

B. COMPARISON WITH STATE-OF-THE ART AI MODELS

To evaluate the practical impact of the provided methodology, it was decided to provide a comparison of state-of-the-art AI models that are versatile and capable of handling various tasks (ChatGPT 4 [46], YandexGPT 2 [47]). These AI models

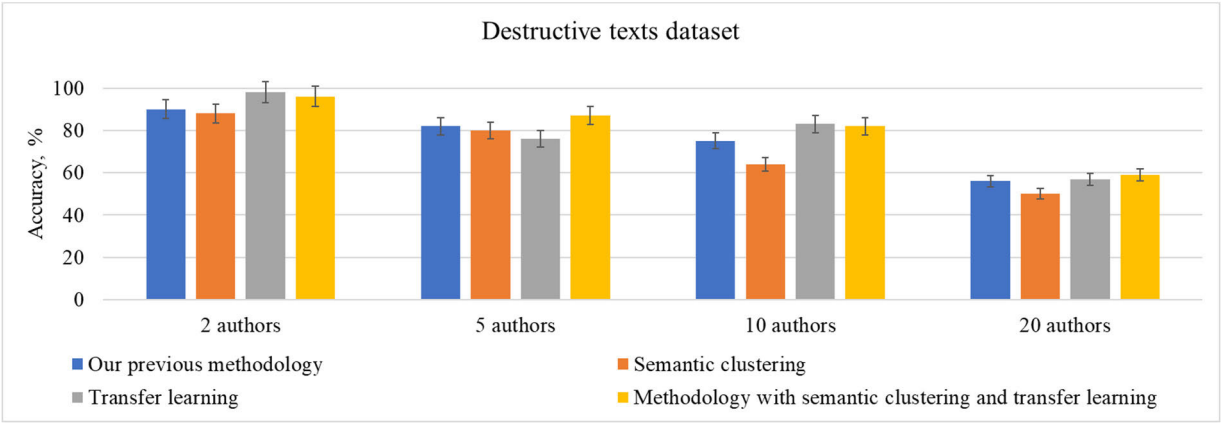

FIGURE 15. Closed-set attribution. Destructive texts dataset.

TABLE 10. Comparison with state-of-the-art models.

Model	Accuracy of identifying destructive texts, %	Accuracy of identifying non-destructive texts, %	Overall accuracy, %
Ours	76%	86%	83%
ChatGPT 4	64%	84%	74%
Yandex GPT 2	n/a (5 answers of 50 prompts)	n/a (4 answers of 50 prompts)	n/a

are generally designed for a broad range of applications, such as natural language processing, content generation, and data analysis. In Particular, the YandexGPT 2 model was made for solving tasks preferably in the Russian language. This comparison aims to assess the effectiveness of these models in scenarios of destructive text detection.

We did not compare these models for the authorship determination task, since the solution to such a problem strictly depends on the training data. Since the interaction with the models is done through queries of limited length, it is difficult to input a proper number of training texts for 5 or more authors.

The experiments used 100 texts, 50 containing destructive content and 50 containing no content. The task of the models was to determine whether the input text was destructive or not. The Prompt for the models was the same, except that for YandexGPT 2 it was translated into Russian, since this model is designed to work with Russian-language queries and responses.

Due to a number of limitations, the YandexGPT 2 model is not always able to give an unambiguous answer about the destructive content of the text, as the authors of the model imposed a number of restrictions on some potentially dangerous topics, so in most cases the model's answer was ≪I am not ready to talk about this topic so as not to offend anyone ≫. As a result, it is not possible to evaluate the accuracy of the model in this experiment.

The results are summarized in Table 10.

According to the results, the proportion of correct answers of the model presented in the article was 83%. These results are related to the specificity of the task and data. Since the problem is related to Russian-language text classification, it is reasonable that ChatGPT, which is perfectly adapted to many NLP tasks for English, performs somewhat worse than a model specially trained for the task. ChatGPT 4 model - 74%.

C. GENERAL GUIDELINES FOR USING THE METHODOLOGY

In these general guidelines for using the methodology, the concept of the balance and equality of sample sizes and class representation is critical for robust model training and performance:

1) When identifying a short text author, it is needed to have 50 samples of each class to train the model. Each sample should be between 50 and 150 characters in length.
2) When identifying destructive content, it is needed to have 100 samples of each class to train the model. The length of each sample should be between 50 and 200 characters.
3) If it is necessary to filter a dataset containing destructive content based on semantic clustering using a reference dataset.
4) When selecting an initial model for the transfer learning process, it is needed to use pre-trained models designed for analyzing Russian-language texts or develop your own model.
5) Semantic clustering should be implemented with the following parameters: UMAP: n_neighbors=15; n_components=3; metric=jaccard; HDBSCAN: metric=Euclidean; cluster_selection_method=eom.
6) The following fastText parameters should be used: the number of neurons in the output layer for authorship classification should be equal to the number of authors (classes), which can be 2, 5, 10, or 20; the activation function should be Softmax. When working on the task of determining destructive content, the number of outputs should be 2 (destructive or non-destructive text), and the activation function should be Sigmoid. Additionally, use number_of_ngrams=3; learning_rate=0.8; dimension=500.

7) For one-class SVM parameters, utilize the sequential optimization method, a linear kernel, a regularization parameter of 1, an acceptable error rate of 0.00001, and apply normalization and compression heuristics.

The methodology has some limitations:

1) The transfer learning process of identifying materials that promote potentially dangerous ideas can be context- and culture-dependent, making it difficult to develop a one-size-fits-all methodology.
2) The transfer learning process itself is also a non-trivial task. The need to select a suitable pre-trained model to solve a related problem whose knowledge is used by the target model, or even to train it, complicates the task.
3) The effectiveness of the technique directly depends on the use of semantic clustering. Clustering allows us to divide texts into semantic clusters according to their subject matter. Its inclusion in the authorship detection methodology contributes to more accurate content analysis and the identification of potentially dangerous materials by filtering out different semantic contexts.
4) It is important to keep in mind that in comments, users may express more than substantive opinions, which adds additional complexity to the analysis in the form of noise and superfluous information.

Regarding the available data, it is worth noting that there is only one open dataset containing similar Russian-language content. However, this dataset has significant shortcomings (lack of markup, many comments unrelated to destructive content), which may affect the quality and generalizability of the results obtained.

The combined use of semantic clustering and transfer learning is a complex and resource-intensive process, as it necessitates the use of several models. The first model contains knowledge related to solving a particular subject matter, while the second model is tailored to directly address the target tasks. When compared to neural networks and classical machine learning methods, the transfer learning process demands more meticulous fine-tuning, substantial computational resources, and experimentation for parameter selection.

Using transfer learning, specifically with a pre-trained BERT model, in the context of analysis of destructive content offers several advantages:

1. Transfer learning process includes a pre-trained BERT model. This model has already learned a wealth of language representations, which can be fine-tuned for specific task, speeding up the research process.

2. BERT's architecture is designed to understand the context and nuances of language. This capability is crucial in authorship attribution, where subtle variations in writing style are key indicators.

3. The approach of extracting BERT embeddings and then combining them with labels for training a fastText classifier demonstrates the adaptability of transfer learning.

4. The diversity and size of the dataset BERT was originally trained on help in generalizing the learning process.

D. LIMITATIONS AND POTENTIAL AREA FOR IMPROVEMENT

Limitations of the technique include:

1. Required amount of text. Each text must consist of at least 50 characters after pre-processing.

2. If it is necessary to filter a thematic dataset to select key information, you should have a reference dataset, based on which you can extract keywords using semantic clustering.

3. When determining the authorship of a text, all texts must be accurately written by the authors alone (without co-authors).

4. In the case of thematic texts, the choice of a transfer learning model should be justified by the subject area of the topic. Such a model must be pre-trained on a sufficient amount of data so that its knowledge can be used in solving the target problem.

5. Availability of sufficient computing resources to work with the neural network. In case of limited resources, classical machine learning methods should be considered instead of NN.

6. The concept of ''destructive content'' may differ in different studies. Some works understand destructive content as only legally prohibited actions, others include content containing psychological violence and immoral texts. This should be taken into account when forming a model, choosing a reference dataset and an initial model for transfer learning. Regarding ethical considerations, if the destructive content of a study contains open calls for hatred and incitement to hatred, then we do not recommend publishing such data in open sources, so as not to introduce thoughtless propaganda of destructive actions. To resolve ethical issues, we have adopted a broad and comprehensive definition of destructive content that covers not only activities prohibited by law, but also content that has the potential to cause psychological harm or promote immorality. This approach allows us to cover a wide range of destructive content, while recognizing the diversity of interpretations in different legal and cultural contexts. We openly declare the subjective nature of determining destructive content and the possibility of different interpretations. In our methodology section, we clearly describe the criteria used for classification, the sources of our reference dataset, and the rationale for selecting our initial model for transfer learning. We also do not encourage authors of destructive comments.

7. The technique is applicable to Russian-language texts.

8. When using One-Class SVM to solve authorship determination with an open set of candidates, a preliminary selection of informative features should be carried out.

9. It is necessary to carry out checks for data homogeneity in the case of combining datasets.

E. FUTURE PLANS

In our future studies, we plan to include more categories to classify different kinds of destructive content. This will help us better understand how these categories affect our

ability to correctly identify who wrote the text. By doing this, we hope to improve our method and make it more accurate in recognizing authors in a wide range of situations.

VI. CONCLUSION

The paper presents a methodology for determining the authorship of Russian-language texts, in particular, short comments of social network users to publications containing news and destructive content. The technique is based on the joint use of semantic text clustering to filter the dataset, as well as transfer learning based on a pre-trained model aimed at determining the types of sensitive topics. All datasets used were subjected to homogeneity checks. When conducting experiments on text authorship detection, the authors treated the problem as a classical classification problem with a closed set of authors and its complex modification - the problem of authorship detection with an open set of candidates. The introduction of texts generated by the generative model was carried out to complicate the problem. A method combining one-class SVM and fastText was proposed for open attribution. In the case of closed attribution, the proposed method achieves high accuracy for classifying 2 and 5 authors (92% and above), and in the case of 10 and more classes, the results are comparable to previous author studies, indicating the adaptability of the methodology to determine the authorship of disruptive content along with regular texts. A comparison with the work of other researchers dealing with this topic has been made. Our obtained accuracy values for such cases (2 and 5 authors) are 9-13% higher than the results presented by other researchers.

VII. ETHICAL CONSIDERATION

The data utilized in this research is exclusively sourced from public domains. This includes information that is freely accessible to anyone, without the need for special permissions or access rights. The public nature of this data inherently reduces certain privacy concerns, as the information is already in the public sphere and available for public consumption and analysis. The methods employed in the collection of this data strictly adhere to legal standards. This compliance ensures that the data gathering process does not infringe on any laws or regulations concerning data privacy or ethical research practices.

VIII. PUBLICLY AVAILABLE DATA

In our commitment to advancing research in the field of natural language processing and content analysis, we are pleased to announce the availability of a publicly accessible demo version of our Telegram Destructive Dataset. This dataset has been meticulously curated to facilitate the exploration and study of destructive content within social media texts, specifically focusing on the Telegram platform.

The Telegram Destructive Dataset encompasses a diverse range of texts, including examples of legally prohibited actions, psychological violence, and immoral texts, reflecting the complex and multifaceted nature of destructive content. The dataset is intended solely for academic and research purposes, and users are encouraged to adhere to ethical guidelines and considerations when utilizing this resource. The demo version of the Telegram Destructive Dataset is available for download at [48].

REFERENCES

[1] A. Fedotova, A. Romanov, A. Kurtukova, and A. Shelupanov, "Authorship attribution of social media and literary russian-language texts using machine learning methods and feature selection," *Future Internet*, vol. 14, no. 1, p. 4, Dec. 2021, doi: 10.3390/fi14010004.
[2] A. Fedotova, A. Romanov, A. Kurtukova, and A. Shelupanov, "Digital authorship attribution in russian-language fanfiction and classical literature," *Algorithms*, vol. 16, no. 1, p. 13, Dec. 2022, doi: 10.3390/a16010013.
[3] A. Romanov, A. Kurtukova, A. Shelupanov, A. Fedotova, and V. Goncharov, "Authorship identification of a russian-language text using support vector machine and deep neural networks," *Future Internet*, vol. 13, no. 1, p. 3, Dec. 2020, doi: 10.3390/fi13010003.
[4] *RuBERT Model*. Accessed: Jan. 13, 2024. [Online]. Available: https://huggingface.co/DeepPavlov/rubert-base-cased
[5] *MultiBERT Model*. Accessed: Jan. 13, 2024. [Online]. Available: https://huggingface.co/bert-base-multilingual-cased
[6] L. Ranaldi, F. Ranaldi, F. Fallucchi, and F. M. Zanzotto, "Shedding light on the dark web: Authorship attribution in radical forums," *Information*, vol. 13, no. 9, p. 435, Sep. 2022, doi: 10.3390/info13090435.
[7] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, and F. Fallucchi, "KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 256–267, doi: 10.18653/v1/2020.emnlp-main.18.
[8] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 67–75, Sep. 2005.
[9] T. Litvinova, O. Litvinova, and P. Panicheva, "Authorship attribution of Russian forum posts with different types of N-gram features," in *Proc. 3rd Int. Conf. Natural Lang. Process. Inf. Retr.*, Jun. 2019, pp. 9–14, doi: 10.1145/3342827.3342834.
[10] *Dark Web Forums*. Accessed: Oct. 13, 2023. [Online]. Available: https://www.azsecure-data.org/dark-web-forums.html
[11] M. Najafi and E. Tavan, "Text-to-text transformer in authorship verification via stylistic and semantical analysis," in *Proc. CLEF*, 2022, pp. 2607–2616. Accessed: Oct. 13, 2023.
[12] D. Embarcadero-Ruiz, H. Gómez-Adorno, A. Embarcadero-Ruiz, and G. Sierra, "Graph-based Siamese network for authorship verification," *Mathematics*, vol. 10, no. 2, p. 277, Jan. 2022, doi: 10.3390/math10020277.
[13] C. Deutsch and I. Paraboni, "Authorship attribution using author profiling classifiers," *Natural Lang. Eng.*, vol. 29, no. 1, pp. 110–137, Jan. 2023, doi: 10.1017/s1351324921000383.
[14] T. C. Nagavi and D. S. Aishwarya, "Automated detection of destructive content in social media," *4th Int. Conf. Recent Develop. Control, Automat. Power Eng. (RDCAPE)*, pp. 124–130, 2021.
[15] V. A. Minaev et al., "Monitoring and identifying destructive information influences in modern social media," [In Russ.], *Inf. Secur., Yesterday, Today, Tomorrow*, pp. 140-145, 2022. Accessed: Mar. 14, 2024. [Online]. Available: https://elibrary.ru/item.asp?id=48737476
[16] I. V. Mashechkin, "Methods of automatic annotation and selection of keywords in problems of detecting extremist information on the Internet," *Modern Inf. Technol. IT Educ.*, vol. 12, no. 1, pp. 188–198, 2016.
[17] I. V. Mashechkin et al., "Machine learning methods for the task of detecting and monitoring extremist information on the Internet," [In Russ.], *Programming*, no. 3, pp. 18–37, 2019. Accessed: Mar. 14, 2024. [Online]. Available: https://elibrary.ru/item.asp?id=37154657
[18] O. Sharif, M. M. Hoque, A. S. M. Kayes, R. Nowrozy, and I. H. Sarker, "Detecting suspicious texts using machine learning techniques," *Appl. Sci.*, vol. 10, no. 18, p. 6527, Sep. 2020, doi: 10.3390/app10186527.
[19] A. B. Goncharov, A. V. Rybakov, and I. M. Azhmukhamedov, "Automated analysis of extremist texts," [In Russ.], *Math. Methods Eng. Technol.*, no. 8, pp. 91–95, 2019.
[20] S. Mussiraliyeva, M. Bolatbek, B. Omarov, and K. Bagitova, "Detection of extremist ideation on social media using machine learning techniques," in *Proc. Int. Conf. Comput. Collective Intell.* Cham, Switzerland: Springer, 2020, pp. 743–752.

[21] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Inf. Sci.*, vol. 307, pp. 39–52, Jun. 2015, doi: 10.1016/j.ins.2015.02.024.

[22] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Proc. 14th Int. Conf.*, 2013, pp. 611–618, doi: 10.1007/978-3-642-41278-3_74.

[23] *BLEU Metric*. Accessed: Jan. 13, 2024. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/bleu

[24] *ROUGE Metric*. Accessed: Jan. 13, 2024. [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[25] *Scikit-learn Metrics*. Accessed: Oct. 13, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[26] *Latent Dirichlet Allocation*. Accessed: Oct. 13, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[27] L. Derczynski, S. Chester, and K. S. Bogh, "Tune your Brown clustering, please," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, vol. 2015, 2015, pp. 110–117.

[28] *BERTopic*. Accessed: Oct. 13, 2023. [Online]. Available: https://maartengr.github.io/BERTopic/index.html

[29] *SentenceTransformers Documentation*. Accessed: Oct. 13, 2023. [Online]. Available: https://www.sbert.net/

[30] *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. Accessed: Oct. 13, 2023. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/

[31] *How HDBSCAN Works*. Accessed: Oct. 13, 2023. [Online]. Available: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

[32] *How ISIS Uses Twitter*. Accessed: Oct. 13, 2023. [Online]. Available: https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter

[33] *Silhouette Score*. Accessed: Oct. 13, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[34] *Davies Bouldin Score*. Accessed: Oct. 13, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

[35] *LaBSE*. Accessed: Oct. 13, 2023. [Online]. Available: https://huggingface.co/sentence-transformers/LaBSE

[36] N. Houlsby, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learning.*, 2019, pp. 2790–2799.

[37] N. Durrani, H. Sajjad, and F. Dalvi, "How transfer learning impacts linguistic knowledge in deep NLP models?" 2021, *arXiv:2105.15179*.

[38] J. Bharadiya, "Transfer learning in natural language processing (NLP)," *Eur. J. Technol.*, vol. 7, no. 2, pp. 26–35, Jun. 2023, doi: 10.47672/ejt.1490.

[39] *Russian Toxicity Classifier*. Accessed: Oct. 13, 2023. [Online]. Available: https://huggingface.co/s-nlp/russian_toxicity_classifier

[40] *Emotion Detection*. Accessed: Oct. 13, 2023. [Online]. Available: https://huggingface.co/cointegrated/rubert-tiny2-cedr-emotion-detection

[41] *Russian Sensitive Topics*. Accessed: Oct. 13, 2023. [Online]. Available: https://huggingface.co/apanc/russian-sensitive-topics

[42] N. Babakov, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko, "Detecting inappropriate messages on sensitive topics that could harm a company's reputation," in *Proc. 8th Workshop Balto-Slavic Natural Lang. Process.*, 2021, pp. 26–36.

[43] V. Moshkin, D. Fadeev, and N. Yarushkina, "Development of a system for finding extremist texts," in *CEUR Workshop Proc.*, 2021, pp. 45–52.

[44] S. Mussiraliyeva, M. Bolatbek, B. Omarov, Z. Medetbek, G. Baispay, and R. Ospanov, "On detecting online radicalization and extremism using natural language processing," in *Proc. 21st Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2020, pp. 1–5, doi: 10.1109/ACIT50332.2020.9300086.

[45] A. I. Kapitanov, I. I. Kapitanova, V. M. Troyanovskiy, V. F. Shangin, and N. O. Krylikov, "Approach to automatic identification of terrorist and radical content in social networks messages," in *Proc. IEEE Conf. Russian Young Res. Electr. Electron. Eng. (EIConRus)*, Jan. 2018, pp. 1517–1520, doi: 10.1109/EICONRUS.2018.8317386.

[46] D. Trajanov, G. Lazarev, L. Chitkushev, and I. Vodenska, "Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks," in *Proc. E3S Web Conf.*, vol. 436, 2023, pp. 1–6, doi: 10.1051/e3sconf/202343602004.

[47] G. Ilieva, T. Yankova, S. Klisarova-Belcheva, A. Dimitrov, M. Bratkov, and D. Angelov, "Effects of generative chatbots in higher education," *Information*, vol. 14, no. 9, p. 492, Sep. 2023, doi: 10.3390/info14090492.

[48] *Destructive Dataset Demo Version*. Accessed: Feb. 27, 2024. [Online]. Available: https://github.com/alexxromanov/destructive_texts_corpus

ANASTASIA FEDOTOVA was born in Tomsk, Russia, in 1999. She received the Graduate degree with a specialization in information security from Tomsk State University of Control Systems and Radioelectronics, in 2023. She is the author of more than 20 scientific articles. Her main research interests include text analysis and information security.

ANNA KURTUKOVA was born in Kemerovo, in 1997. She received the Graduate degree with a specialization in information security from Tomsk State University of Control Systems and Radioelectronics, in 2021.

She is the author of more than 40 scientific articles and the winner of more than ten research competitions and conferences. Her research interests include information security, machine learning, natural language processing, and computer vision.

ALEKSANDR ROMANOV was born in Tomsk, in 1985. He received the Graduate degree with a specialization in information security from Tomsk State University of Control Systems and Radioelectronics (TUSUR), in 2007, and the Candidate of Engineering Sciences degree, in 2010.

He has been an Associate Professor with TUSUR, since 2012. He has published more than 100 articles. His research interests include information security, text processing, artificial intelligence, mathematical modeling, numerical methods, and program complexes.

ALEXANDER SHELUPANOV (Senior Member, IEEE) received the Graduate degree in applied mathematics and mechanical engineering from Tomsk State University, in 1976.

In 1988, he became the Head of Sector, the Head of Division, and the Deputy Director for Research. In 1993, he joined the team of Tomsk State University of Control Systems and Radioelectronics (TUSUR), in 1999, becoming the Head of the Department of Complex Information Security of Computer Systems. In 2008, he became the Director of the Institute of System Integration and Security, TUSUR. From 2010 to 2014, he was the Vice-Rector for Research, and from 2014 to 2019, he was a Rector with TUSUR. In 2019, he was elected as the University President. He has published more than 500 research articles. He is a Research Supervisor of the Tomsk IEEE Chapter. His research interests include the fundamental and applied bases in the field of design and development of complex systems for information security, and information security.