НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2026 Том 26 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2026 Vol. 26 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-1-165-176

Deep learning for author gender and sex identification in natural language text

Aleksandr S. Romanov[1], Anastasiia M. Fedotova[2]✉, Anna V. Kurtukova[3]

[1,2,3] Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634034, Russian Federation

[1] alexx.romanov@gmail.com, https://orcid.org/0000-0002-2587-2222
[2] afedotowaa@yandex.ru✉, https://orcid.org/0000-0001-7844-4363
[3] av.kurtukova@gmail.com, https://orcid.org/0000-0001-5619-1836

Abstract
With the development of digital interaction, there is a need to control the information distributed by users so that it complies with legal requirements. The study proposes a technique for classifying Russian-language texts by considering both the biological sex of the authors (male and female) and gender-related distinctions, including heterosexual and homosexual groups as well as specific LGBT categories such as gay men, lesbian women, bisexual, and transgender authors. The article proposes a technique for identifying the sex and gender of authors of Russian-language texts by utilizing an ensemble of methods, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and the Russian Bidirectional Encoder Representations from Transformers (RuBERT). This research is introducing a technique for gender identification in Russian texts, as prior studies have only addressed the automatic determination of an author's sex. To conduct the study, custom datasets of user comments from social networks were created and annotated with both sex and gender labels. In tasks related to identifying the author's sex, the proposed technique achieved an accuracy exceeding 90 %. This includes the classification of heterosexual men and women as well as a mixed dataset comprising both heterosexual and homosexual individuals. When distinguishing individuals of the same sex based on sexual orientation, the results showed that homosexual and heterosexual women display uncommon writing styles in contrast to men of different orientations, with accuracy rates of 93 % for women and 85 % for men. Additionally, an experiment focused on identifying LGBT individuals and their gender identities based on writing style achieved an accuracy of 93 %. The key takeaway from this study is that combining CNN, RuBERT, and SVM leads to a more robust model for gender classification. The method has been tested on varying numbers of samples. In the basic experiments for determining sex, the method achieves an accuracy of 92 %. When addressing the task of detecting the author's LGBT affiliation, the method achieves 93 % accuracy. The proposed technique can be applied for automated monitoring of social networks to detect and analyze the gender identity of text authors. It is also promising for use in content moderation systems as well as in sociological and linguistic research.

Keywords
gender identification, author profiling, natural language processing, machine learning, neural networks

Acknowledgements
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).

For citation: Romanov A.S., Fedotova A.M., Kurtukova A.V. Deep learning for author gender and sex identification in natural language text. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 1, pp. 165–176. doi: 10.17586/2226-1494-2026-26-1-165-176

УДК 004.89

Глубокое обучение в задачах идентификации пола и гендера автора естественно-языкового текста

Александр Сергеевич Романов[1], Анастасия Михайловна Федотова[2]✉, Анна Владимировна Куртукова[3]

[1,2,3] Томский государственный университет систем управления и радиоэлектроники, Томск, 634034, Российская Федерация

[1] alexx.romanov@gmail.com, https://orcid.org/0000-0002-2587-2222
[2] afedotowaa@yandex.ru✉, https://orcid.org/0000-0001-7844-4363
[3] av.kurtukova@gmail.com, https://orcid.org/0000-0001-5619-1836

Аннотация

Введение. Развитие способов взаимодействия в цифровой среде приводит к необходимости контроля распространяемой пользователями информации на соответствие ее требованиям законодательства. В работе предлагается методика классификации русскоязычных текстов с учетом биологического пола (мужчины и женщины) и гендерных различий, включая гомо- и гетеросексуальные группы, а также отдельные категории ЛГБТ (геи, лесбиянки, бисексуальные и трансгендерные авторы). **Метод.** Предлагаемый подход основан на ансамбле методов, включающих Support Vector Machine (SVM), Convolutional Neural Networks (CNN) и Russian Bidirectional Encoder Representations from Transformers (RuBERT). Представленная методика позволяет определять не только пол, но и гендер автора русскоязычного текста, в то время как известные исследования рассматривают только автоматическое определение пола автора. Для проведения исследования собраны пользовательские наборы комментариев пользователей в социальных сетях, аннотированные по полу и гендеру. **Основные результаты.** В задачах определения пола автора предложенный метод показал точность, превышающую 90 %. Рассматривалась классификация как гетеросексуальных мужчин и женщин, так и смешанных групп, включавших гетеро- и гомосексуальных людей. При различении сексуальной ориентации среди представителей одного пола точность составила 93 % для женщин (различение лесбиянок и гетеросексуальных женщин) и 85 % для мужчин (классификация геев и гетеросексуальных мужчин). При исследовании по определению одновременно принадлежности к ЛГБТ и гендерной идентичности достигнутая точность составила 93 %. **Обсуждение.** Показано, что сочетание CNN, RuBERT и SVM может быть применено в качестве основы методики идентификации пола и гендера авторов текста. Предложенный подход может использоваться для автоматизированного мониторинга социальных сетей с целью выявления и анализа гендерной принадлежности автора текста. Разработанная методика перспективна для применения в системах контент-модерации, социологических и лингвистических исследований.

Ключевые слова

идентификация пола, определение автора текста, обработка естественного языка, машинное обучение, нейронные сети

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках базовой части государственного задания ТУСУРа на 2023–2025 гг. (проект № FEWM-2023-0015).

Ссылка для цитирования: Романов А.С., Федотова А.М., Куртукова А.В. Глубокое обучение в задачах идентификации пола и гендера автора естественно-языкового текста // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 1. С. 165–176 (на англ. яз.). doi: 10.17586/2226-1494-2026-26-1-165-176

Introduction

Changes in the structure of social interaction are closely related to the development of digital platforms that provide new opportunities for communication and self-expression. Today, people can not only maintain connections with their immediate environment, but also interact with global communities. However, with the development of digital interaction, there is a need to control the information distributed by users so that it complies with legal requirements [1]. Thus, in connection with the decision of the Supreme Court of the Russian Federation of November 30, 2023, according to which the LGBT movement was recognized as an extremist organization and banned in Russia, special attention should be paid to monitoring activity on social networks. Modern automated solutions are required to support these measures, ensuring compliance with legal requirements and preventing the dissemination of prohibited content.

This study presents a novel technique in the Russian language for determining the gender of a text author. A Convolutional Neural Network (CNN), a Support Vector Machine (SVM), BERT (Bidirectional Encoder Representations from Transformers) is used together in an ensemble. Semantic clustering is also used to select features. Additionally, the study introduces original methods for automatic gender annotation based on the user's name and filtering through the analysis of user reactions to the published content. This study is the first in the Russian language to address the determination of both the biological sex and gender of a text author. The method integrates the strengths of deep neural networks, including transformer models, with classical machine learning techniques. The approach is based on a carefully designed feature space, incorporating character trigram frequencies, semantic keywords extracted using BERTopic, Katz smoothing, and original data filtering methods.

Although the concept of gender encompasses a wide and continuously expanding range of identities: with some sources identifying between 46 and 78 genders — this study focuses on four primary LGBT-related categories: gay men, lesbian women, bisexual individuals, and transgender individuals. These categories were selected because they are both well-represented in the Russian-speaking digital space and feasible to reliably annotate based on publicly available data. Expanding the dataset to include less common gender identities is limited by the availability of open, annotated, and demographically diverse text samples. Including such underrepresented groups would undermine the reliability and statistical validity of the resulting models. Thus, the current scope ensures methodological rigor while addressing a relevant and socially significant subset of gender identities.

Review

As of early 2025, the analysis of LGBT culture and its followers in Russia remains in its infancy, with only a few sociological studies [1–3]. However, no automated solutions have yet been proposed for the issue. In contrast, researchers working with English and other languages have made more significant advancements in addressing this issue.

Among studies focused on the Russian language, we provide an overview of the most notable approaches to determining an author's gender from text. T. Litvinova [4] proposed a method for identifying the gender of an internet blog author using Latent Dirichlet Allocation (LDA) and Random Forest (RF). Their study involved a dataset of 2,362 bloggers, with 41 % identified as women and 59 % as men. Classification was performed using RF. For a balanced dataset (966 samples per class), the classification accuracy reached 80 %.

The article [5] investigates various machine learning methods, including SVM, Naive Bayes (NB), K-nearest Neighbors (KNN), and RF, applied to Russian literary prose from the 1960s to 2000s. The analysis used features such as parts of speech (e.g., nouns, verbs, prepositions, pronouns, conjunctions, and adjectives) and punctuation marks. The dataset consisted of 417,000 words usages. The most effective methods were SVM and NB, each achieving an accuracy of 71.43 %. KNN performed the poorest, with an accuracy of 54.45 %.

Another study [6] analyzed a dataset of 114,046 words from 38 users of the social network VK, aged 18 to 20. Features for classification included the frequency of personal pronouns to represent egocentric, addressee-oriented, and collective positions, respectively. Various machine learning techniques were applied, including LDA, RF, SVM, Decision Trees (DT), NB, Logistic Regression (LR), Long Short-Term Memory (LSTM), as well as statistical methods like correlation analysis and the Generalized Linear Model. Among these, SVM achieved the highest accuracy of 86 %.

A study [7] examined two text corpora, each containing 120 texts written by 60 men and 60 by women. Classification was performed using a set-theoretic collocation model and a DT. Features included the frequency of noun-verb pairs, the use of Latin characters, word length, and punctuation marks. Additional features, such as the number of adjectives, frequency of adverbs, number of *n*-grams, and average word length, were also considered. The collocation model achieved an impressive accuracy of 93.3 % in determining the author's gender.

Bondarenko and Krivov [8] proposed a technique for determining the gender of a text author using SVM. The study utilized Russian-language texts, including news articles and interviews. The training dataset consisted of approximately 25,000 words, while 500 messages were used for testing. Prior to analysis, the texts underwent pre-processing steps, such as tokenization, stop-word removal, lemmatization, and normalization. The selected features were word frequencies within the dataset. The method performance varied depending on the type of training data: training on news articles resulted in an accuracy of 64 %, while training on interview texts achieved an accuracy of 81 %.

Another study [9] employed several Russian-language text corpora, including the Gender Imitation Crowdsource dataset which contained texts from 800 authors. Each author wrote three texts on the same topic, one in their natural style, one imitating the opposite gender, and another with a different style unrelated to gender. The average text length was approximately 300 words. The RusPersonality corpus, comprising 1,145 authors and over 1,850 texts with an average length of 230 words, included topics such as letters to friends and image descriptions. The RusProfiling corpus contained 1000 authors, with 641 men and 392 women. The study employed a wide range of machine learning models and neural networks, including CNN, Multilayer Perceptron (MLP), LSTM, character-based models with single-character encoding, and deep neural networks leveraging morpho-syntactic features. The most effective model was gradient boosting with character-based *n*-gram text representation, the best results in testing reached an F1-score of 58 %.

The goal of one of the PAN-2017 competitions was to identify the gender of a Russian-language text author. The competition provided participants with four datasets: user posts from X (formerly Twitter) and Facebook, thematic essays, and reviews from marketplaces.

The team led by I. Markov [10] achieved the highest results for the X, Facebook, and reviews datasets. Their approach combined SVM with mathematical statistical methods. The SVM-based approach achieved an accuracy of 93 % on the Facebook dataset and 62 % for the review dataset. The statistical method performed best on the X dataset, with an accuracy of 68 %.

Another notable approach was developed by R. Bhargava [11] who achieved the highest performance in classifying thematic essays. Their methodology combined LSTM and Bidirectional LSTM, reaching an accuracy of 78 %. While their preprocessing steps were similar to those of Markov's team, they also included stemming and lemmatization to enhance data preparation.

In the study [12], the authors created several of their own datasets and conducted classification on the same datasets provided to PAN competition participants. They compiled a dataset of essays involving 125 men and 269 women. Each participant was asked to write an essay on a given topic, then write two more essays on the same topic: one from the perspective of a person of the opposite gender and another from the perspective of a person of their own gender. The authors did not provide details on the data preprocessing methods. Feature extraction involved TF-IDF (Term Frequency-Inverse Document Frequency) text representation, character *n*-grams, and Word2Vec. A DT achieved the highest accuracy of 70 % on the authors' dataset with writing style imitation, while the hybrid CNN + LSTM model achieved an accuracy of 64 % on the PAN 2017 dataset.

The study by P. Panicheva et al. [13] aims to evaluate the effectiveness of semantic features in the task of gender classification of Russian-language social media posts. The dataset comprised 1,879 users (48 % male and 52 % female), with 130 texts per user. For semantic feature extraction, the authors employed LDA, Author-Topic Model (ATM), and Distributional Semantic Clusters (DSC). LDA was implemented in a multithreaded version with 10 iterations, treating each text as a separate document and extracting key terms. In the case of ATM, the approach was conceptually similar to merging texts written by the same author into a single, longer document. This allowed the model to exploit a larger portion of the available data and better capture traits inherent to the aggregated texts of different authors. Regarding DSC, the Skip-Gram Word2Vec model, trained on the Russian National Corpus, was applied. The features obtained by the three methods were used both as independent feature spaces and in the form of a unified vector. For the classification of author gender, an SVM model was selected. The results indicate the effectiveness of ATM-based features (69.2 % accuracy).

The approach to determining the gender of the author of Russian-language texts was presented by D.A. Galkina [14]. It is noted that the author's gender may be inferred from the verb ending in the past tense singular form. However, this method is not applicable to texts written in the third person or to formal texts that employ impersonal constructions. Therefore, the author relied on the findings of [15] which indicated that an abundance of pronouns in a text may serve as a characteristic feature of female-authored writing. Among the classifiers applied, Random Forest demonstrated the best performance for non-fiction texts, achieving 94 % accuracy, while for fiction texts AdaBoost performed best, with an accuracy of 85 %.

The authors of [16] addressed the task of gender identification based on short texts authored by men and women as well as by generative models designed to imitate male and female writing styles. The dataset consisted of 6,000 tweets, including those generated by AI. The methodology involved tweet preprocessing, TF-IDF feature extraction, and feature selection using the Chi-squared test. The classifiers employed were SVM, DT, RF, and MLP. The best performance was achieved with MLP, yielding an accuracy of 90 %.

The study [17] focuses on gender identification of authors of literary texts. The dataset employed, BookSCE, comprises 8,222 books. The methodology combines classical machine learning approaches (LR, XGBoost, and SVM) with deep learning models (BERT, GPT-2, XLNet, Robustly Optimized BERT Approach (RoBERTa), and T5). The models were trained on preprocessed textual data, using word *n*-grams as features for the classical methods and contextual embeddings for the deep learning models. The findings indicate that GPT-2 outperforms all other models, achieving an accuracy of 92.5 %, followed by XLNet with 90.7 %.

Among the studies focused on gender determination, notable contributions include the works of A. Karami, D. Manikandan, and others.

The article [18] aims to automate the categorization of LGBT user profiles on X (formerly Twitter), including both individual and organizational accounts, using machine learning methods applied to user texts and biographies. During data collection, profiles containing LGBT-related keywords were searched. A total of 42,644 profiles were initially collected which were reduced to 16,241 accounts after duplicate removal and manual annotation. The features were divided into bibliographic features (e.g., word frequency in profile descriptions and word count) and profile features. Feature selection was performed using Analysis of Variance (ANOVA). The optimal combination of profile and bibliographic features achieved an accuracy of 88 % using Naive Bayes.

D. Manikandan et al. [19] also addressed the problem of detecting hate speech against the LGBTQ+ community. Their technique was based on deep learning models utilizing transformer architectures, specifically BERT and XLM-RoBERTa. The dataset comprised comments collected from the YouTube platform in English, including 3,001 non-offensive comments, 157 homophobic comments, and 6 transphobic comments. The test set included 990 unlabeled comments. All texts were processed using WordPiece tokenizers (for BERT) and byte-level BPE tokenizers (for RoBERTa). The XLM-RoBERTa model was used with default parameters. BERT had an accuracy of 91 %, whereas XLM-RoBERTa had an accuracy of 93 %.

The task of creating annotated datasets and comprehensive dictionaries to train lexicon-based approaches is the focus of the article [20]. To accomplish this, the authors utilized datasets of gendered pronouns in English. They employed template queries such as "A occupation * pronoun" to automatically extract sentences with potential bias. Nine individuals collected and subsequently annotated a total of 700 sentences. These sentences were further augmented to 3,510 sentences using pronoun replacement techniques. The second dataset encompassed exclusionary term dictionaries, utilizing and expanding biased terms from the United Nations guidelines on gender-sensitive language via Word2Vec. The initial dictionary contained 86 words which were expanded to 145 terms. A separate dictionary of gendered neologisms was created using gender-biased terms from Urban Dictionary. The final dictionary contained approximately 500 new terms.

The paper [21] addresses the problem of gender identification in Arabic text, specifically in Egyptian dialect, using tweets. Existing methods have struggled with identiying gender from short sentences, like individual tweets. To solve this, the authors introduced the AraEGI dataset, consisting of 42,000 Egyptian dialect tweets annotated for both speaker and listener gender. The dataset includes three subsets: AraEGI-E, AraEGI-P, and AraEGI-S, each derived from publicly available datasets. The authors used transformer-based models to classify gender in these tweets and benchmarked their results. They found that these models performed well, overcoming the limitations of previous methods which were less effective for short texts.

The paper [22] addresses the critical problem of bias detection in language processing systems where biased data can lead to unfair outcomes in various applications, such as social media analysis, healthcare, and recruitment. The study introduces Nbias, a comprehensive framework for detecting both explicit and implicit biases in textual data. The authors designed a novel dataset and proposed a method for detecting bias using a transformer-based token classification model which introduces a new entity type called BIAS to specifically identify biased words or phrases. The model was evaluated using both quantitative and qualitative measures, showing accuracy improvements of 1 % to 8 % over baseline models.

In study [23], the dataset consisted of pages of LGBT activists and political figures in Russian, English, and Spanish, comprising more than 30,000 samples (10,000 texts for each language). The task was to identify LGBT activists. LR was employed as the classification model, yielding accuracy rates above 70 % across all languages. The highest result was obtained for English, with an accuracy of 83 %.

The authors of [24] address the task of identifying categories of prohibited or harmful content, such as LGBT-related material, online crime, violence, armed conflict, pornography distribution, toxic text, and others. The dataset was generated using generative models and comprised 124,597 Russian-language samples. All texts were manually annotated into 21 categories. For classification, a Russian-language implementation of BERT was employed. The results for detecting prohibited content ranged from 40 % to 85 %, depending on the category.

Recent research has shown the potential of Large Language Models (LLMs), such as GPT-3, GPT-4[1], RoBERTa, and XLM-R, for author profiling and gender classification tasks. These models benefit from extensive pretraining and can generalize well even with limited labeled data, achieving competitive results in zero-shot and few-shot settings [25, 26]. However, despite their high capacity, LLMs come with considerable computational costs and may introduce challenges in fine-tuning and deployment.

In this study, we propose an ensemble method combining SVM, CNN, and Russian Bidirectional Encoder Representations from Transformers (RuBERT). This approach balances the strengths of deep contextual modeling (RuBERT), local feature extraction (CNN), and robust classification (SVM). Compared to LLMs, our ensemble requires significantly fewer computational resources and performs effectively on balanced datasets. Furthermore, it provides greater interpretability and control which are essential in sensitive domains such as gender and identity classification. While LLMs represent a promising direction, we demonstrate that traditional and transformer-based ensembles can still yield high performance in practical applications.

These LLM-based methods outperform traditional approaches in contexts requiring deep semantic understanding, disambiguation, or long-range dependencies. However, they often require more computational resources and may inherit societal biases present in the training data.

The analysis of the reviewed sources demonstrates that there is a notable lack of approaches for identifying LGBT-related gender categories in Russian-language texts. Most existing studies on Russian material focus only on determining the author's biological sex and predominantly employ SVM, achieving a maximum reported accuracy of 93 %. In contrast, research on English-language texts addressing gender identity increasingly relies on transformer-based models, particularly BERT. For Russian, the RusProfiling dataset by T. Litvinova [4], tested within the PAN conferences, remains one of the few resources; however, it provides only sex annotation and is unsuitable for addressing more complex gender distinctions. Features traditionally used for sex identification are often insufficient for capturing the nuanced stylistic and identity-related patterns required for gender classification. Until now, the task of automated gender identification in Russian-language texts has not been systematically addressed. Therefore, this study aims to develop a unified feature space and a classification approach applicable to both sex and gender identification. While certain distinctions, such as those between homosexual and heterosexual men and women, are closely associated with biological sex, other gender-related differences require more advanced stylistic analysis and semantic modeling, which the proposed ensemble method is designed to capture.

Method

Fig. 1 illustrates the technique for determining the sex and gender of an anonymous text author. The subsections in this section provide a detailed description of each stage of the process.

Telegram is one of the most popular messaging platforms among Russian speakers, offering thematic channels that feature original posts and discussions in the form of comments.

Telegram attracts a broad user base, including individuals with diverse gender identities, who use the platform to discuss various topics and share experiences that may not always align with their identity or orientation. The dataset for this study was collected using the Telegram API. The Telegram dataset is divided into two parts: "Telegram channels" and "Telegram comments". The "Telegram channels" subset consists of original posts made by the authors of thematic Telegram channels. In contrast, the "Telegram comments" subset includes user-generated responses to these posts which are often more informal and expressive in nature. This separation allows for the analysis of linguistic features in both controlled (channel posts) and

[1] OpenAI. GPT-4 Technical Report. Available at: https://openai.com/research/gpt-4 (accessed: 10.04.2025).

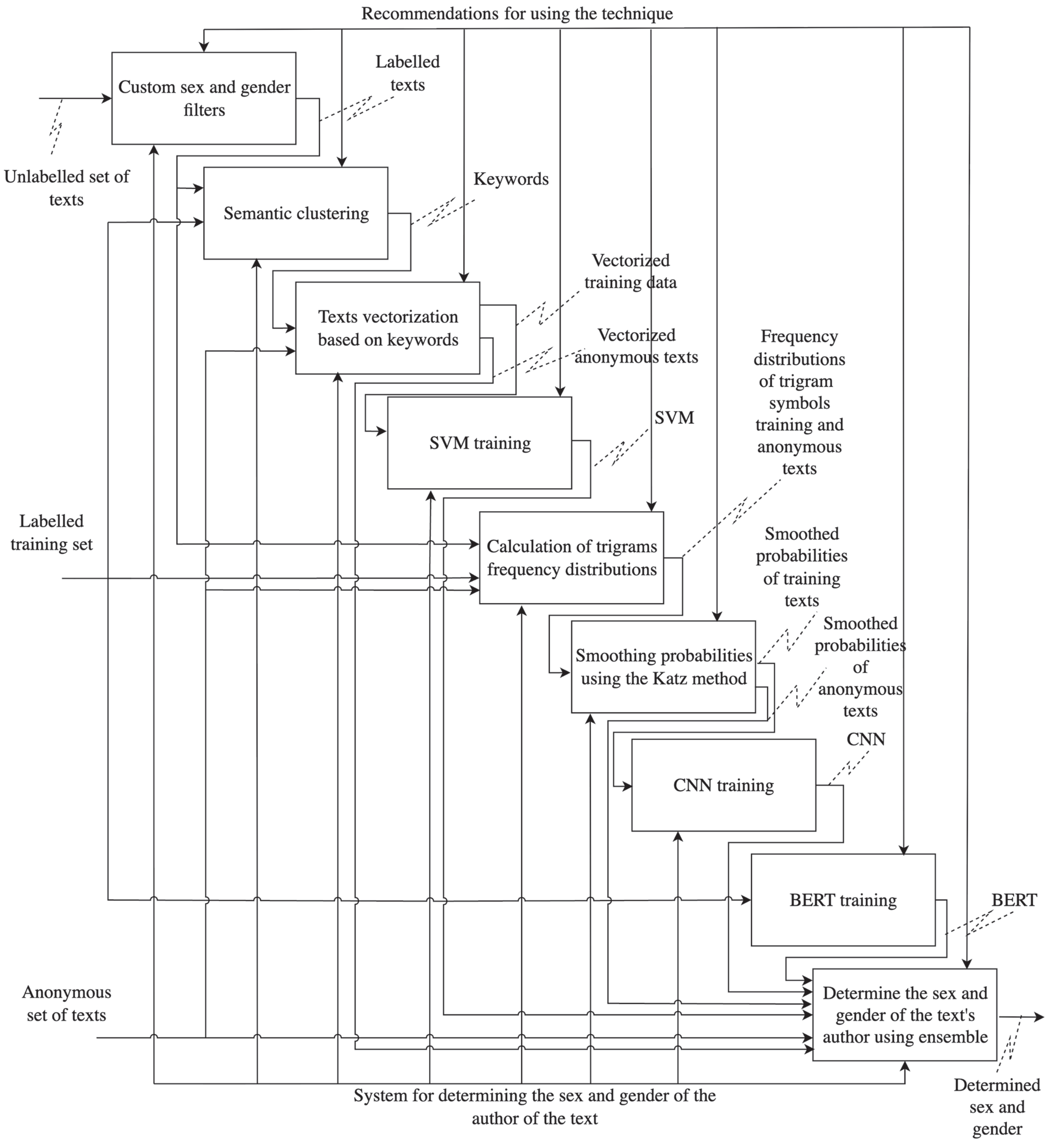

Fig. 1. Technique for determining the sex and gender of an anonymous text author

spontaneous (comments) writing contexts. A data cleaning process was implemented to remove noise, including spam, advertisements, and bot-generated content: we excluded bot-generated content, removed content that was marked as advertisements, also removed any URLs or links from the texts and messages, as they could potentially skew the data or introduce irrelevant content. Additionally, we filtered out messages that were flagged as harmful by other users such as spam or abusive content.

Authors of Telegram channels may allow subscribers to leave comments which often reflect unfiltered user reactions to posts. The absence of influence from Social Media Marketing specialists further enhances the value of these comments for linguistic analysis.

Since Telegram does not provide user gender data, we implemented a custom mechanism for automatic gender annotation. This approach utilizes available account information (e.g., first name, last name, username, and biography), a dictionary of Russian names (including transliterated variants)[1], and regular expressions for surnames.

[1] Online dictionary of Russian personal names by N.A. Petrovsky. Available at: https://1794.slovaronline.com (accessed: 27.10.2025).

To minimize errors in gender annotation, the following steps were implemented.

Step 1: The author's name was analyzed using a dictionary of Russian names and their transliterated forms, focusing on names commonly used by native Russian speakers.

Step 2: Gender was assigned based on both the first name and surname, as Russian surnames often have gender-specific endings.

Step 3: Additional information provided in the author's biography was considered. In LGBT+ channels, authors frequently include pronouns (e.g., she/her, he/his, they/them) or other details about their gender identity in their profile or channel description.

Step 4: The combination of these actions minimizes errors.

The profiles and channels used for data collection were manually curated, with careful consideration of the authors' biographies, available information, and their texts.

Information about the labeled datasets of comments after filtering is given in Table 1.

To create the second dataset, the social network X (formerly Twitter) was selected due to its allowance of content related to LGBTQ + topics. A manual keyword search was conducted to identify accounts actively discussing LGBTQ + topics, forming the basis of the dataset. Table 2 provides detailed information about the datasets.

The process of semantic clustering follows the approach outlined in our previous work [27]. In this study, we used the translated LGBT dataset[1] as the reference dataset. Function Flow Diagram of semantic clustering is presented in Fig. 2.

Stage descriptions: F1 — text dataset construction; F2 — embedding generation using the LaBSE model; F3 — embedding generation using the Paraphrase-multilingual-MiniLM model; F4 — embedding generation using the Multi-qa-mpnet-base-dot-v1 model; F5 — dimensionality reduction using UMAP; F6 — clustering using the HDBSCAN algorithm; F7 — clustering using the K-Means algorithm; F8 — clustering using the Birch algorithm; F9 — hyperparameter optimization via greedy search; F10 — clustering quality evaluation using standard quality metrics.

For each text, a feature vector was generated, incorporating keyword frequencies as SVM features.

The SVM was trained using 5-fold cross-validation with the following configuration: training algorithm — sequential optimization method, kernel — linear, regularization parameter — 1, permissible error level — 0.00001, normalization and compression heuristic — enabled. The SVM classifier was trained using keyword frequency vectors derived from the semantic clustering process illustrated in Fig. 2. The keywords were extracted automatically from each semantic cluster, and their normalized frequencies were used as features. In total, 3,000 features were selected based on their frequency and semantic relevance. This method ensures that the feature space captures semantically meaningful distinctions between texts. A full description of the semantic clustering process is provided in our previous work [27].

To train the models, it is necessary to obtain male and female feature vectors. Texts written by men tend to be more problem-oriented, concise, and informative, whereas women tend to write texts that are emotionally rich and expressive [28].

Methods based on *n*-gram frequency may prove to be the most effective for identifying gender differences in texts. This is due to the linguistic features of the Russian language [29] where the endings and suffixes of adjectives, verbs, and other parts of speech can differ depending on the subject's gender.

Katz's method [29] smoothed the trigrams in the dataset to generate the features for the CNN. The use of lexical-grammatical word classes is minimally reflective of the

[1] LGBT Tweets. Available at: https://www.kaggle.com/datasets/vencerlanz09/lgbt-tweets/ (accessed: 27.10.2025).

Table 1. Telegram comments dataset

Characteristic	Men	Women	Gays	Lesbians	Bisexuals	Transsexuals	LGBT in general
Telegram comments							
Number of texts	60,354	63,397	42,594	20,212	2,913	5,935	96,486
Number of characters	9,939,081	9,494,575	6,748,182	4,974,809	749,093	1,003,100	17,785,887
Number of users	8,497	14,298	3,429	1,998	655	1,009	13,253
Telegram channels							
Number of texts	48,491	76,034	1,681	2,527	3,192	5,381	26,272
Number of characters	28,532,043	28,759,251	1,105,241	2,873,749	2,473,842	3,235,497	3,007,826
Number of channels	25	101	9	5	5	10	11

Table 2. X dataset

Characteristic	Men	Women	Gays	Lesbians	Bisexuals	Transsexuals	LGBT in general
Number of texts	44,470	43,570	31,782	36,857	30,061	30,931	33,038
Number of characters	5,932,698	5,377,996	3,988,324	4,486,219	3,868,860	4,704,609	4,567,291
Number of users	69	64	81	65	106	80	63

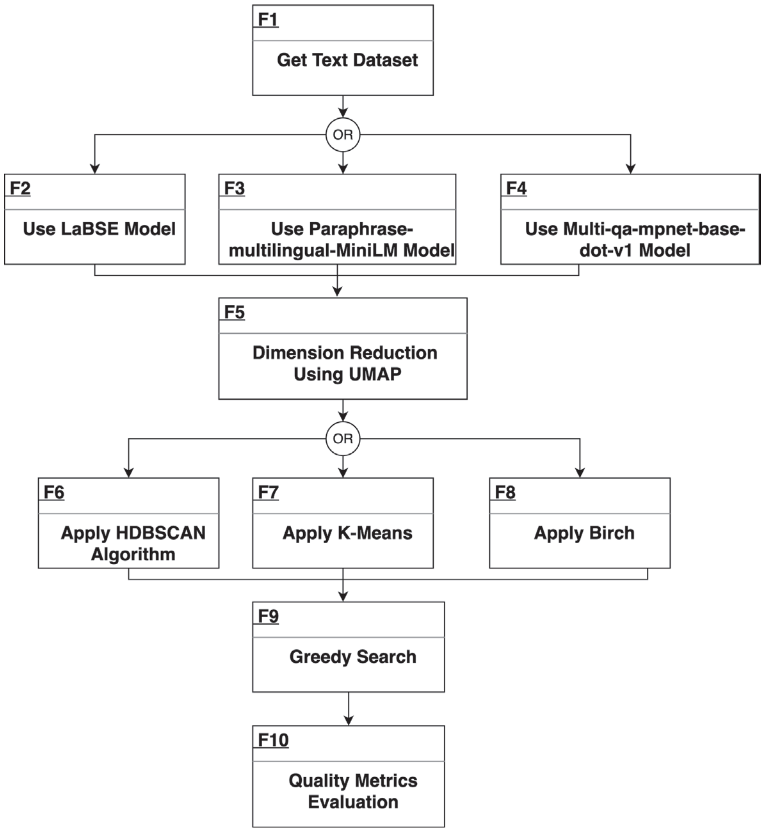

Fig. 2. Semantic clustering

author's intentions and is less likely to be consciously altered for stylistic transformations.

The CNN [30, 31] was trained using 5-fold cross-validation with the following configuration: embedding layer: dimensionality — 300, convolutional layer: layer size — 2,048, kernel size — 3, activation function — ReLU, BatchNormalization, GlobalMaxPooling1D, dense layer: dimensionality — number of classes, activation function — SoftMax, number of epochs — 10, batch size — 16. The model consists of: Embedding Layer, SpatialDropout1D Layer, Conv1D Layer, GlobalMaxPooling1D Layer, Dense Layer. The model uses binary cross-entropy as the loss function since this is a binary classification problem and categorical for multi-class. The Conv1D layer has filters, and each filter has a kernel of size 3. The Dense layer has weights connecting the output of the GlobalMaxPooling1D to the output units (number of classes).

The BERT model [32, 33] was used in the implementation RuBERT. Text vectorization was performed using the built-in tokenizer, and applied with the following parameters: activation function — ReLU, dropout rate — 0.2, learning rate — 0.00001, number of epochs — 10, batch size — 16, number of warm-up steps — 500, weight decay constant — 0.01. Also, an early stopping mechanism during the training of BERT and RuBERT was employed.

The final decision is made using an ensemble of SVM, CNN, and RuBERT classifiers. The models are assigned weights of 0.2, 0.4, and 0.4 for SVM, CNN, and RuBERT, respectively. Weights were chosen using greedy search.

The models were trained using 5-fold cross-validation, with 80 % of the data allocated for training and 20 % reserved for testing. This approach ensures a robust evaluation and reduces the risk of biases affecting the model's performance.

Results

The study included ten experimental setups. The first involved sex classification among heterosexual individuals, determining the author's biological sex when only heterosexual authors were considered. The second focused on detecting whether the text author belongs to the LGBT community. The third addressed distinguishing texts written by heterosexual men or women from those authored by LGBT representatives. The fourth involved joint classification of sex and orientation, identifying whether a text was written by a heterosexual or homosexual man or woman. The fifth experiment compared homosexual men and homosexual women. The sixth examined sex classification across orientation, determining the author's sex in a mixed group of heterosexual and homosexual individuals. The seventh targeted the distinction between heterosexual and homosexual men, while the eighth focused on differentiating heterosexual and homosexual women. The ninth aimed to classify texts authored by heterosexual men, heterosexual women, and homosexual women. Finally, the tenth experiment identified whether a text was written by a bisexual person, a heterosexual man, or a heterosexual woman.

Table 4. Results of classification experiments for all models on the X, Telegram posts, and Telegram comments datasets

Case	Dataset	SVM	RF	KNN	GRU	LSTM	fastText	RuBERT	CNN	SVM + CNN	CNN + RuBERT	SVM + CNN + RuBERT
		Accuracy, %										
1	X	76 ± 1	68 ± 3	60 ± 5	72 ± 3	77 ± 4	70 ± 2	91 ± 3	90 ± 2	88 ± 3	**92 ± 4**	91 ± 3
	T. posts	81 ± 1	76 ± 2	62 ± 2	79 ± 2	85 ± 2	68 ± 3	88 ± 4	89 ± 2	85 ± 4	**90 ± 3**	**90 ± 3**
	T. comm.	67 ± 3	66 ± 3	51 ± 3	75 ± 3	83 ± 4	64 ± 4	**92 ± 3**	85 ± 3	78 ± 3	90 ± 2	90 ± 3
2	X	86 ± 3	69 ± 2	57 ± 4	80 ± 4	83 ± 3	76 ± 2	88 ± 2	88 ± 2	89 ± 3	88 ± 2	**89 ± 4**
	T. posts	83 ± 1	84 ± 2	55 ± 6	82 ± 4	86 ± 2	75 ± 1	86 ± 2	86 ± 2	84 ± 3	**92 ± 4**	90 ± 3
	T. comm.	72 ± 3	72 ± 3	56 ± 3	78 ± 2	85 ± 1	70 ± 2	91 ± 4	91 ± 4	86 ± 2	**93 ± 3**	**93 ± 4**
3	X	63 ± 4	55 ± 4	38 ± 3	71 ± 3	76 ± 2	67 ± 3	82 ± 3	79 ± 3	76 ± 4	**84 ± 4**	82 ± 3
	T. posts	79 ± 7	82 ± 3	54 ± 1	79 ± 2	78 ± 3	76 ± 2	86 ± 3	79 ± 2	83 ± 2	86 ± 3	**87 ± 3**
	T. comm.	82 ± 6	78 ± 2	56 ± 4	66 ± 3	60 ± 5	78 ± 3	82 ± 4	84 ± 2	85 ± 3	83 ± 4	**87 ± 4**
4	X	62 ± 4	27 ± 3	38 ± 5	65 ± 5	67 ± 2	55 ± 5	79 ± 2	78 ± 4	79 ± 3	**80 ± 3**	79 ± 5
	T. posts	71 ± 1	66 ± 1	60 ± 8	73 ± 2	75 ± 2	64 ± 2	85 ± 2	81 ± 2	83 ± 4	**87 ± 3**	85 ± 4
	T. comm.	69 ± 4	48+6	25+2	79 ± 2	88 ± 2	68 ± 1	90 ± 3	89 ± 2	78 ± 3	**91 ± 3**	**91 ± 2**
5	X	75 ± 2	60 ± 4	66 ± 2	73 ± 3	73 ± 3	75 ± 2	88 ± 2	83 ± 3	82 ± 4	**89 ± 3**	88 ± 3
	T. posts	89 ± 1	83 ± 1	85 ± 6	88 ± 2	84 ± 1	86 ± 3	88 ± 4	90 ± 2	90 ± 3	**92 ± 3**	**92 ± 2**
	T. comm.	74 ± 4	61 ± 4	63 ± 1	72 ± 2	71 ± 2	66 ± 4	84 ± 3	75 ± 2	80 ± 3	**87 ± 3**	86 ± 3
6	X	69 ± 2	54 ± 4	62 ± 2	67 ± 3	69 ± 3	72 ± 3	**79 ± 3**	73 ± 3	**79 ± 2**	73 ± 3	76 ± 3
	T. posts	70 ± 4	66 ± 2	69 ± 4	82 ± 3	83 ± 2	69 ± 2	86 ± 2	85 ± 2	80 ± 3	**91 ± 2**	85 ± 3
	T. comm.	80 ± 2	72 ± 3	66 ± 3	80 ± 3	78 ± 3	76 ± 2	90 ± 3	89 ± 2	86 ± 2	**92 ± 4**	90 ± 2
7	X	82 ± 4	71 ± 4	73 ± 3	83 ± 2	75 ± 2	82 ± 2	88 ± 3	72 ± 5	84 ± 4	**89 ± 3**	88 ± 3
	T. posts	89 ± 1	85 ± 2	81 ± 3	85 ± 1	86 ± 3	85 ± 3	90 ± 1	90 ± 6	90 ± 3	**93 ± 2**	93 ± 3
	T. comm.	71 ± 3	66 ± 6	51 ± 2	66 ± 4	68 ± 3	68 ± 3	79 ± 3	78 ± 2	75 ± 2	**81 ± 2**	80 ± 3
8	X	76 ± 2	71 ± 2	49 ± 4	77 ± 3	84 ± 3	80 ± 1	89 ± 3	86 ± 2	83 ± 4	**92 ± 3**	89 ± 2
	T. posts	91 ± 2	88 ± 2	84 ± 2	89 ± 3	88 ± 2	86 ± 3	95 ± 2	92 ± 2	95 ± 3	**96 ± 2**	**96 ± 2**
	T. comm.	78 ± 2	69 ± 4	65 ± 2	70 ± 2	76 ± 3	70 ± 2	92 ± 2	88 ± 3	85 ± 2	**92 ± 4**	90 ± 4
9	X	64 ± 3	57 ± 4	35 ± 5	68 ± 4	74 ± 3	78 ± 2	78 ± 2	78 ± 2	75 ± 3	**81 ± 2**	80 ± 2
	T. posts	78 ± 8	72 ± 1	69 ± 1	74 ± 3	76 ± 2	72 ± 2	83 ± 3	79 ± 1	83 ± 2	**85 ± 4**	**85 ± 2**
	T. comm.	58 ± 2	53 ± 2	47 ± 2	65 ± 4	66 ± 2	54 ± 3	75 ± 3	**81 ± 4**	69 ± 2	78 ± 3	72 ± 2
10	X	66 ± 3	58 ± 4	44 ± 5	69 ± 3	76 ± 3	73 ± 4	76 ± 3	78 ± 2	72 ± 3	**80 ± 2**	78 ± 3
	T. posts	75 ± 2	65 ± 2	62 ± 2	71 ± 3	70 ± 3	64 ± 3	88 ± 2	85 ± 2	84 ± 3	**90 ± 2**	88 ± 3
	T. comm.	57 ± 2	54 ± 1	49 ± 1	60 ± 1	62 ± 2	55 ± 4	69 ± 2	69 ± 1	67 ± 3	**70 ± 2**	69 ± 3

Note. Values highlighted in bold indicate the best experimental results.

Table 4 presents the results of the experiments conducted on the Telegram posts (T. posts) and comments (T. comm.) and X datasets. It should be noted that, the fastText classifier was included as a baseline. fastText is an efficient end-to-end model that performs both word embedding and classification. In our setup, texts were preprocessed by steps declared in the Method section. These texts were then directly fed into fastText, which uses subword information to create embeddings and applies a linear classifier.

The use of ensemble methods outperforms classical approaches by up to 53 % and surpasses neural networks and deep neural networks by up to 25 % and 21 %, respectively, in specific analyses. Additionally, the ensemble approach performs up to 8 % better than individual NN models. The final experiment for determining the author's sex and gender was conducted on a combined dataset that included texts from both social networks. In this case, only the accuracy of ensembles that demonstrated the best performance on individual datasets was evaluated. Table 5 presents the results.

Since the datasets used in the experiments were balanced (i.e., equal number of samples per class), the accuracy metric was appropriate for evaluating model performance. However, we acknowledge that for imbalanced datasets, alternative metrics such as F1-score would be more suitable.

In tasks distinguishing based on sex, the accuracy of the method exceeds 92 % for both the classification

Table 5. Experimental evaluation of the technique on combined dataset

Case	Number of samples per class	SVM + CNN	SVM + CNN + RuBERT	CNN + RuBERT
		Accuracy, %		
1	153,315	91 ± 4	**92 ± 3**	**92 ± 4**
2	336,316	91 ± 3	**93 ± 3**	92 ± 3
3	153,315	74 ± 3	**82 ± 2**	**82 ± 3**
4	59,596	74 ± 3	84 ± 2	**85 ± 3**
5	59,596	84 ± 3	89 ± 3	**90 ± 3**
6	229,372	84 ± 2	86 ± 2	**88 ± 3**
7	76,057	83 ± 2	84 ± 4	**85 ± 2**
8	59,596	86 ± 3	**93 ± 3**	91 ± 2
9	59,596	74 ± 4	**80 ± 2**	79 ± 3
10	36,166	76 ± 3	75 ± 3	**77 ± 3**

Note. Values highlighted in bold indicate the best experimental results.

of heterosexual men and women and the use of a mixed dataset that includes both heterosexual and homosexual men and women. This highlights the identification of speech behavior patterns that are dependent on sex rather than orientation. When distinguishing individuals of the same sex by orientation, it was found that homosexual and heterosexual women exhibit less similar writing styles than men of different orientations (93 % accuracy for women and 85 % in the experiment with men). The experiment achieved an accuracy of 93 % in identifying LGBT individuals and gender identity (traditional and non-traditional) based on writing style. Additionally, in an experiment aimed at predicting gender groups which included four LGBT classes and two classes representing hetero users, the model accuracy reached 69 %.

Discussion and Conclusion

This article introduces a technique for identifying the sex and gender of a text author, distinguished by its integration of semantic clustering and an ensemble approach combining SVM, CNN, and RuBERT. This work advances the field of author profiling by providing a more accurate method for determining gender and biological sex in textual data.

The key takeaway from this study is that combining CNN, RuBERT, and SVM leads to a more robust model for gender classification. The method has been tested on varying numbers of samples. In the basic experiments for determining sex, the method achieves an accuracy of 92 %. When addressing the task of detecting the author's LGBT affiliation, the method achieves 93 % accuracy. In experiments distinguishing between homosexual and heterosexual men and women, the proposed approach demonstrates accuracy of 85 % and 93 %, respectively.

References

1. Buyantueva R. Cultural and geopolitical conflicts between the West and Russia: western NGOs and LGBT activism. *Connexe: Les Espaces Postcommunistes en Question (s)*, 2022, vol. 8, pp. 148–167. https://doi.org/10.5077/journals/connexe.2022.e1031
2. Stoltz P., Khlusova A. Sexual citizenship, nationalisms, and bordering: russian LGBT activists and the “Foreign agent law”. *The Palgrave Handbook of Gender and Citizenship*, 2024, pp. 425–447. https://doi.org/10.1007/978-3-031-57144-2_19
3. Simonova M.A. The problem of LGBTgbt propaganda: the experience of Russia and the Usa (theoretical and legal analysis). *Agrarian and Land Law*, 2023, no. 3 (219), pp. 162–166. (in Russian). https://doi.org/10.47643/1815-1329_2023_3_162
4. Litvinova T.A. Topic modeling of the corpus of blogs in Russian with respect to author's gender: text and context. *Ethnopsycholinguistics*, 2022, vol. 2 (9), pp. 7–23. (in Russian). https://doi.org/10.31249/epl/2022.02.01
5. Khazova A. Automatic detection of gender identity: the phenomenon of russian women's prose. *NSU Vestnik. Series: Linguistics and Intercultural Communication*, 2020, vol. 18, no. 1, pp. 22–32. (in Russian). https://doi.org/10.25205/1818-7935-2020-18-1-22-32
6. Stepanenko A.A., Rezanova Z.I. Pronouns as machine learning markers in genre and gender text attribution. *Tomsk State University Journal of Philology*, 2022. no. 79, pp. 131–154. (in Russian). https://doi.org/10.17223/19986645/79/7

Литература

1. Buyantueva R. Cultural and geopolitical conflicts between the West and Russia: western NGOs and LGBT activism // Connexe: les espaces postcommunistes en question (s). 2022. V. 8. P. 148–167. https://doi.org/10.5077/journals/connexe.2022.e1031
2. Stoltz P., Khlusova A. Sexual citizenship, nationalisms, and bordering: russian LGBT activists and the “Foreign agent law” // The Palgrave Handbook of Gender and Citizenship. 2024. P. 425–447. https://doi.org/10.1007/978-3-031-57144-2_19
3. Симонова М.А. Проблема пропаганды ЛГБТ: опыт России и США (теоретико-правовой анализ) // Аграрное и земельное право. 2023. № 3 (219). С. 162–166. https://doi.org/10.47643/1815-1329_2023_3_162
4. Литвинова Т.А. Тематическое моделирование корпуса блогов на русском языке с учетом гендера автора: текст и контекст // Этнопсихолингвистика. 2022. № 2 (9). С. 7–23. https://doi.org/10.31249/epl/2022.02.01
5. Хазова А.Б. Автоматическое определение половой принадлежности автора текста: феномен русской женской прозы // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18. № 1. С. 22–32. https://doi.org/10.25205/1818-7935-2020-18-1-22-32
6. Степаненко А.А., Резанова З.И. Местоимения в автоматической жанровой и гендерной атрибуции текстов // Вестник Томского государственного университета. Филология. 2022. № 79. С. 131–154. https://doi.org/10.17223/19986645/79/7

7. Osochkin A., Piotrowska X., Fomin V. Automatic identification of authors' stylistics and gender on the basis of the corpus of russian fiction using extended set-theoretic model with collocation extraction. *Glottometrics*, 2021, vol. 50, pp. 76–89. https://doi.org/10.53482/2021_50_389
8. Bondarenko V.V., Krivov D.A. Profiling of the author of electronic communication. *Proceedings of Young Scientists and Specialists of Samara University*, 2023, no. 2 (23), pp. 46–52. (in Russian)
9. Sboev A., Gudovskikh D., Moloshnikov I., Rybka R. A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models. *AIP Conference Proceedings*, 2019, vol. 2116, no. 1, pp. 270006. https://doi.org/10.1063/1.5114280
10. Markov I., Gómez-Adorno H., Sidorov G., Gelbukh A. The winning approach to cross-genre gender identification in Russian at rusprofiling 2017. *Ceur Workshop Proceedings*, 2017, vol. 2036, pp. 20–24.
11. Bhargava R., Goel G., Shah A., Sharma Y. Gender Identification in Russian Texts. *Ceur Workshop Proceedings*, 2017, vol. 2036, pp. 13–16.
12. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception. *Procedia Computer Science*, 2018, vol. 123, pp. 417–423. https://doi.org/10.1016/j.procs.2018.01.064
13. Panicheva P., Mirzagitova A., Ledovaya Y. Semantic feature aggregation for gender identification in Russian Facebook. *Communications in Computer and Information Science*, 2018, vol. 789, pp. 3–15. https://doi.org/10.1007/978-3-319-71746-3_1
14. Galkina D.A. Automatic author profile detection for natural language texts. *Path to the Science: Applied Mathematics, Computer Science, and Information Technology*, 2021, pp. 26–28. (in Russian)
15. Argamon S., Koppel M., Fine J., Shimoni A. Gender, genre, and writing style in formal written texts. *Text*, 2003, vol. 23, no. 3, pp. 321–346. https://doi.org/10.1515/text.2003.014
16. Alowibdi J.S. Gender prediction of generated tweets using generative AI. *Information*, 2024, vol. 15, no. 8, pp. 452. https://doi.org/10.3390/info15080452
17. Younkin V., Litvak M., Rabaev I. Automatic gender identification from text. *Applied Sciences*, 2024, vol. 14, no. 24, pp. 12041. https://doi.org/10.3390/app142412041
18. Karami A., Lundy M., Webb F., Boyajieff H.R., Zhu M., Lee D. Automatic categorization of LGBT user profiles on twitter with machine learning. *Electronics*, 2021, vol. 10, no. 15, pp. 1822. https://doi.org/10.3390/electronics10151822
19. Manikandan D., Subramanian M., Shanmugavadivel K. A system for detecting abusive contents against LGBT community using deep learning based transformer models. *Ceur Workshop Proceedings*, 2022, vol. 3395, pp. 106–116.
20. Doughman J., Khreich W. Gender bias in text: Labeled datasets and lexicons. *arXiv*, 2022. arXiv:2201.08675. https://doi.org/10.48550/arXiv.2201.08675
21. Abdelmaguid A., Torki M., Khalafallah A. AraEGI: Arabic gender identification using transformers for Egyptian dialect. *Proc. of the IEEE Symposium on Computers and Communications (ISCC)*, 2024, pp. 1–7. https://doi.org/10.1109/ISCC61673.2024.10733708
22. Raza S., Garg M., Reji D.J., Bashir S.R., Ding C. NBIAS: A natural language processing framework for BIAS identification in text. *Expert Systems with Applications*, 2024, vol. 237, part B, pp. 121542. https://doi.org/10.1016/j.eswa.2023.121542
23. Park C.Y., Yan X., Field A., Tsvetkov Y. Multilingual contextual affective analysis of LGBT people portrayals in Wikipedia. *Proc. of the 15th International AAAI Conference on Web and Social Media*, 2021, vol. 15, no. 1, pp. 479–490. https://doi.org/10.1609/icwsm.v15i1.18077
24. Babakov N., Logacheva V., Panchenko A. Beyond plain toxic: detection of inappropriate statements on flammable topics for the Russian language. *arXiv*, 2022. arXiv:2203.02392. https://doi.org/10.48550/arXiv.2203.02392
25. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020, vol. 33.
26. Liang P., Bommasani R., Lee T., Tsipras D., Soylu D., Yasunaga M., et. al. Holistic evaluation of language models. *arXiv*, 2022. arXiv:2211.09110. https://doi.org/10.48550/arXiv.2211.09110
27. Fedotova A., Kurtukova A., Romanov A., Shelupanov A. Semantic clustering and transfer learning in social media texts authorship

7. Osochkin A., Piotrowska X., Fomin V. Automatic identification of authors' stylistics and gender on the basis of the corpus of russian fiction using extended set-theoretic model with collocation extraction // Glottometrics. 2021. V. 50. P. 76–89. https://doi.org/10.53482/2021_50_389
8. Кривов Д.А., Бондаренко В.В. Профилирование автора электронного сообщения // Вестник молодых учёных и специалистов Самарского университета. 2023. № 2 (23). С. 46–52.
9. Sboev A., Gudovskikh D., Moloshnikov I., Rybka R. A gender identification of text author in mixture of Russian multi-genre texts with distortions on base of data-driven approach using machine learning models // AIP Conference Proceedings. 2019. V. 2116. N 1. P. 270006. https://doi.org/10.1063/1.5114280
10. Markov I., Gómez-Adorno H., Sidorov G., Gelbukh A. The winning approach to cross-genre gender identification in Russian at rusprofiling 2017 // Ceur Workshop Proceedings. 2017. V. 2036. P. 20–24.
11. Bhargava R., Goel G., Shah A., Sharma Y. Gender Identification in Russian Texts // Ceur Workshop Proceedings. 2017. V. 2036. P. 13–16.
12. Sboev A., Moloshnikov I., Gudovskikh D., Selivanov A., Rybka R., Litvinova T. Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception // Procedia Computer Science. 2018. V. 123. P. 417–423. https://doi.org/10.1016/j.procs.2018.01.064
13. Panicheva P., Mirzagitova A., Ledovaya Y. Semantic feature aggregation for gender identification in Russian Facebook // Communications in Computer and Information Science. 2018. V. 789. P. 3–15. https://doi.org/10.1007/978-3-319-71746-3_1
14. Галкина Д.А. Автоматическое определение профиля автора текста на естественном языке // Путь в науку: прикладная математика, информатика и информационные технологии. 2021. С. 26–28.
15. Argamon S., Koppel M., Fine J., Shimoni A. Gender, genre, and writing style in formal written texts // Text. 2003. V. 23. N 3. P. 321–346. https://doi.org/10.1515/text.2003.014
16. Alowibdi J.S. Gender prediction of generated tweets using generative AI // Information. 2024. V. 15. N 8. P. 452. https://doi.org/10.3390/info15080452
17. Younkin V., Litvak M., Rabaev I. Automatic gender identification from text // Applied Sciences. 2024. V. 14. N 24. P. 12041. https://doi.org/10.3390/app142412041
18. Karami A., Lundy M., Webb F., Boyajieff H.R., Zhu M., Lee D. Automatic categorization of LGBT user profiles on twitter with machine learning // Electronics. 2021. V. 10. N 15. P. 1822. https://doi.org/10.3390/electronics10151822
19. Manikandan D., Subramanian M., Shanmugavadivel K. A system for detecting abusive contents against LGBT community using deep learning based transformer models // Ceur Workshop Proceedings. 2022. V. 3395. P. 106–116.
20. Doughman J., Khreich W. Gender bias in text: Labeled datasets and lexicons // arXiv. 2022. arXiv:2201.08675. https://doi.org/10.48550/arXiv.2201.08675
21. Abdelmaguid A., Torki M., Khalafallah A. AraEGI: Arabic gender identification using transformers for Egyptian dialect // Proc. of the IEEE Symposium on Computers and Communications (ISCC). 2024. P. 1–7. https://doi.org/10.1109/ISCC61673.2024.10733708
22. Raza S., Garg M., Reji D.J., Bashir S.R., Ding C. NBIAS: A natural language processing framework for BIAS identification in text // Expert Systems with Applications. 2024. V. 237. Part B. P. 121542. https://doi.org/10.1016/j.eswa.2023.121542
23. Park C.Y., Yan X., Field A., Tsvetkov Y. Multilingual contextual affective analysis of LGBT people portrayals in Wikipedia // Proc. of the 15th International AAAI Conference on Web and Social Media. 2021. V. 15. N 1. P. 479–490. https://doi.org/10.1609/icwsm.v15i1.18077
24. Babakov N., Logacheva V., Panchenko A. Beyond plain toxic: detection of inappropriate statements on flammable topics for the Russian language // arXiv. 2022. arXiv:2203.02392. https://doi.org/10.48550/arXiv.2203.02392
25. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. 2020. V. 33.
26. Liang P., Bommasani R., Lee T., Tsipras D., Soylu D., Yasunaga M., et. al. Holistic evaluation of language models // arXiv. 2022. arXiv:2211.09110. https://doi.org/10.48550/arXiv.2211.09110
27. Fedotova A., Kurtukova A., Romanov A., Shelupanov A. Semantic clustering and transfer learning in social media texts authorship

attribution. *IEEE Access*, 2024, vol. 12, pp. 39783–39803. https://doi.org/10.1109/ACCESS.2024.3377231
28. Wahyuningsih S. Men and women differences in using language: A case study of students at STAIN Kudus. *EduLite: Journal of English Education, Literature and Culture*, 2018, vol. 3, no. 1, pp. 79–90. https://doi.org/10.30659/e.3.1.79-90
29. Katz S.M. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1987, vol. 35, no. 3, pp. 400–401. https://doi.org/10.1109/tassp.1987.1165125
30. Khatua A., Cambria E., Ghosh K., Chaki N., Khatua A. Tweeting in support of LGBT? A deep learning approach. *Proc. of the ACM India Joint International Conference on Data Science and Management of Data*, 2019, pp. 342–345. https://doi.org/10.1145/3297001.3297057
31. Duan M., Li K., Yang C., Li K. A hybrid deep learning CNN–ELM for age and gender classification. *Neurocomputing*, 2018, vol. 275, pp. 448–461. https://doi.org/10.1016/j.neucom.2017.08.062
32. Lavanya S.K., Navaneethakrishnan S.C. Building tamil text dataset on LGBTQIA and offensive language detection using multilingual BERT. *Proc. of the International Conference on Inventive Computation Technologies (ICICT)*, 2022, pp. 489–496. https://doi.org/10.1109/icict54344.2022.9850904
33. Bhardwaj R., Majumder N., Poria S. Investigating gender bias in BERT. *Cognitive Computation*, 2021, vol. 13, no. 4, pp. 1008–1018. https://doi.org/10.1007/s12559-021-09881-2

attribution // IEEE Access. 2024. V. 12. P. 39783–39803. https://doi.org/10.1109/ACCESS.2024.3377231
28. Wahyuningsih S. Men and women differences in using language: A case study of students at STAIN Kudus // EduLite: Journal of English Education, Literature and Culture. 2018. V. 3. N 1. P. 79–90. https://doi.org/10.30659/e.3.1.79-90
29. Katz S.M. Estimation of probabilities from sparse data for the language model component of a speech recognizer // IEEE Transactions on Acoustics, Speech, and Signal Processing. 1987. V. 35. N 3. P. 400–401. https://doi.org/10.1109/tassp.1987.1165125
30. Khatua A., Cambria E., Ghosh K., Chaki N., Khatua A. Tweeting in support of LGBT? A deep learning approach // Proc. of the ACM India Joint International Conference on Data Science and Management of Data. 2019. P. 342–345. https://doi.org/10.1145/3297001.3297057
31. Duan M., Li K., Yang C., Li K. A hybrid deep learning CNN–ELM for age and gender classification // Neurocomputing. 2018. V. 275. P. 448–461. https://doi.org/10.1016/j.neucom.2017.08.062
32. Lavanya S.K., Navaneethakrishnan S.C. Building tamil text dataset on LGBTQIA and offensive language detection using multilingual BERT // Proc. of the International Conference on Inventive Computation Technologies (ICICT). 2022. P. 489–496. https://doi.org/10.1109/icict54344.2022.9850904
33. Bhardwaj R., Majumder N., Poria S. Investigating gender bias in BERT // Cognitive Computation. 2021. V. 13. N 4. P. 1008–1018. https://doi.org/10.1007/s12559-021-09881-2

Authors

Aleksandr S. Romanov — D.Sc., Associate Professor, Senior Researcher, Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634034, Russian Federation, sc 57221288963, https://orcid.org/0000-0002-2587-2222, alexx.romanov@gmail.com

Anastasiia M. Fedotova — Junior Researcher, Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634034, Russian Federation, sc 57214763715, https://orcid.org/0000-0001-7844-4363, afedotowaa@yandex.ru

Anna V. Kurtukova — Junior Researcher, Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634034, Russian Federation, sc 57207346456, https://orcid.org/0000-0001-5619-1836, av.kurtukova@gmail.com

Авторы

Романов Александр Сергеевич — доктор технических наук, доцент, старший научный сотрудник, Томский государственный университет систем управления и радиоэлектроники, Томск, 634034, Российская Федерация, sc 57221288963, https://orcid.org/0000-0002-2587-2222, alexx.romanov@gmail.com

Федотова Анастасия Михайловна — младший научный сотрудник, Томский государственный университет систем управления и радиоэлектроники, Томск, 634034, Российская Федерация, sc 57214763715, https://orcid.org/0000-0001-7844-4363, afedotowaa@yandex.ru

Куртукова Анна Владимировна — младший научный сотрудник, Томский государственный университет систем управления и радиоэлектроники, Томск, 634034, Российская Федерация, sc 57207346456, https://orcid.org/0000-0001-5619-1836, av.kurtukova@gmail.com

Received 24.03.2025
Approved after reviewing 15.10.2025
Accepted 24.01.2026

Статья поступила в редакцию 24.03.2025
Одобрена после рецензирования 15.10.2025
Принята к печати 24.01.2026