information

MDPI

Article

Determining the Age of the Author of the Text Based on Deep Neural Network Models

Aleksandr Sergeevich Romanov, Anna Vladimirovna Kurtukova *, Artem Alexandrovich Sobolev, Alexander Alexandrovich Shelupanov and Anastasia Mikhailovna Fedotova

Faculty of Security, Tomsk State University of Control Systems and Radioelectronics, 634050 Tomsk, Russia; alexx.romanov@gmail.com (A.S.R.); bingjo-ya@yandex.ru (A.A.S.); saa@tusur.ru (A.A.S.); afedotowaa@icloud.com (A.M.F.)

* Correspondence: av.kurtukova@gmail.com

Received: 11 November 2020; Accepted: 18 December 2020; Published: 21 December 2020

check for updates

Abstract: This paper is devoted to solving the problem of determining the age of the author of the text based on models of deep neural networks. The article presents an analysis of methods for determining the age of the author of a text and approaches to determining the age of a user by a photo. This could be a solution to the problem of inaccurate data for training by filtering out incorrect user-specified age data. A detailed description of the author's technique based on deep neural network models and the interpretation of the results is also presented. The study found that the proposed technique achieved 82% accuracy in determining the age of the author from Russian-language text, which makes it competitive in comparison with approaches for other languages.

Keywords: authorship; age; text analysis; computer vision; social networks; FastText; VGG-Face; CRNN

1. Introduction

Text messages are the most popular form of communication and an important part of the daily life of a modern person. The unique style of such messages is directly dependent on gender, age, education and other factors, which makes it possible to clearly identify their author. The features that distinguish the author of a particular message are expressed through the structure of sentences, vocabulary, the use of certain linguistic structures and speech patterns. These described features make it possible to differentiate people into groups.

The relevance of research on this topic is associated with the auxiliary function for solving the problems of text mining [1–7], particularly its attribution [8,9]. The sentiment of the text, as well as the gender and age of the author, are the most informative features used in attribution. Their use can significantly improve the generalization ability of the model for authorship attribution. The obtained solutions of these problems find their application in such important areas of life as information security and forensics, commerce (for example, as a way to optimize targeted advertising), and science (as a tool for linguistic research). They are especially important in forensic expertise, where it is necessary to solve such problems as the following:

- Differentiating users of online platforms by age to combat pedophilia and prevent children from accessing adult content;
- Authorship attribution of an anonymous note with threats;
- Authorship attribution of a suicide note.

Everyone has a unique writing style. This is expressed in the sentence structure, vocabulary, use of certain language structures, and figures of speech. The described features help to differentiate people into groups by age.

The purpose of the study is to determine the age of the author of the text. The assigned tasks included the following:

1. Collecting data from a social network;
2. Pre-processing of texts;
3. Analysis of text classification methods;
4. Analysis of methods for determining age from a photo;
5. Data filtering using computer vision (CV) algorithms;
6. Classification of the original and processed data;
7. Evaluation and analysis of results.

2. Literature Review

A large number of works are devoted to the problem of determining the age of the author of a text. In [10], authors used four machine learning (ML) algorithms for age determination: logistic regression, naive Bayesian classifier (NB), gradient boosting, and multilayer perceptron (MLP). The dataset included 350 texts divided into 3 age classes (15–19, 20–24, and 25 years and older). The best result was obtained using MLP, with an accuracy of 67.18%. The average accuracy of the ensemble using classifiers was 51%.

The authors of [11] presented their TweetGenie system for determining the age of a user. The system is based on the idea of sociolinguistics, and used data obtained from users' social network pages. The corpus was based on the tweets of 3000 users. The authors identified three age categories of users (under 20, 20–40, and over 40). The mean absolute error (MAE) for this system was 4.844, and the Pearson coefficient was 0.866. The results showed the limitations of current text-based age determination approaches. The disadvantage of the system is the tendency to underestimate the age of the elderly. Specifically, Twitter users are perceived to be younger than they are.

The work in [12] was devoted to the determining the age of the text author with the support vector machine (SVM). The study used texts from social networks. The experiments used short texts containing 12 tokens or more. For the 16+ years class, the accuracy was 71.3%; for 18+ years, 80.8%; and for 25+ years, 88.2%. The best criteria for classifying age, according to the research results, were *n*-grams of words.

The authors of [13] proposed several classification models using deep learning and classical ML methods. These methods determined the age of a Twitter user based on the textual content of tweets. The age classifier was trained on a dataset of 11,160 users, which included combinations of words and symbol *n*-grams as attributes obtained using latent semantic analysis (LSA). When solving the problem of author profiling in the framework of PAN-2018 (an international competition on digital text forensics and stylometry tasks), this model achieved the best accuracy (82.21%) for the English language.

The analysis conducted in [14] was devoted to the comparison of SVM and a Bayesian neural network (NN). The texts for the experiments were taken from online diaries and divided into four age categories: under 18, 20–27, 30–37, and 40 and older. The accuracy achieved by the SVM model was 50.1%, and the accuracy of the Bayesian NN was 48.2%.

The authors of [15] explored a method based on word taxonomies. A parallel algorithm, tax2vec, was used for constructing taxonomy-based features. The results of the work were demonstrated by solving six short text classification problems: determining the author's gender (PAN 2017), personality type (MBTI), age (PAN 2016), and news feed topics (BBC), identifying the side effects of drugs, and identifying their effectiveness. Tax2vec first compared words from individual documents with their synonyms, which were considered according to TF-IDF. For each task, at least 400 texts were used. Selected semantic functions were tested based on a hierarchical attention NN. It was found that enabling semantic functions could improve performance when working with short texts such as tweets. The following results were obtained: 81.7% for author gender (PAN 2017), 68.2% for personality type (MBTI), 51.2% for age (PAN 2016), 99.6% for news feed topics (BBC), 52.3% for drug side effects, and 50% for their effectiveness.

The authors of [16] proposed a joint learning approach for age prediction. In particular, the long short-term memory (LSTM) auxiliary layer was used to study the text representation and simultaneously combine the auxiliary representation into the main LSTM layer to adjust the age regression. During training, the auxiliary classification LSTM model and the main LSTM regression model were used together. The texts were collected from weibo.com, a well-known microblogging platform in China. The data included 2000 samples of texts written by people between 19 and 28 years old. These texts were divided into 10 age categories. The results showed that this approach significantly improved the effectiveness of age prediction using either an individual classification or a regression model. The accuracy of the dataset used was 57.3%.

The work in [17] presents the idea of using a multi-task convolutional NN (MTCNN) for determining the age and gender of the authors from weibo.com. Users with less than 40 subscribers were removed. The resulting dataset consisted of 263,460 entries of 136,072 users. In addition, information on gender and age was considered. Experiments showed that the proposed method demonstrated an improved result, compared with the SVM and CNN methods for a small dataset, and that in comparison with similar work for English texts, in the case of the Chinese language, as of 2016, the complexity of the problem increased. The experimental results showed accuracies of 68.7% and 71.4% for age and gender, respectively.

The research conducted in [18] presented an automated tool with a unique set of features for analyzing a given text. These parameters were unigrams, parts of speech, and production rules. The proposed method consisted of several steps. The first step was data entry. The second step was tokenization and the extraction of parameter sets. The third step was text cleaning. The fourth step was to apply feature selection to the data. The fifth step involved applying the classifier using different algorithms (SVM, NB algorithm, decision tree, logistic regression, random forest (RF), and multiclass classifier). The last step was to create an output class and evaluate the model performance. The following classes were used in the experiment: gender (male, female), age (younger—no more than 35 years old—and older, or 35 years old and older). The best results were an accuracy of 82.81% for gender determination using SVM and 83.2% accuracy for age determination using the SVM classifier.

In [19], one of the tasks was to classify Russian-language texts by gender and age. The authors used the SVM and RF approaches. The prediction was performed on the corpus of 15,000 LiveJournal posts. Age groups from 20 to 50 years old participated in the study, and texts containing less than 1000 characters were not considered. The best results were achieved using the RF classifier, resulting in a 49.5% accuracy for age without considering gender and 49.3% with it.

It should be noted that most of the methods considered for determining the age of the author of the text have low accuracy, but this is not the only drawback. All research was focused either on short (posts from social networks and blogs, comments, and so on) or long (articles, posts, reviews, and the like) texts; that is, they were strictly dependent on the length. This approach is incorrect since it does not allow us to assess the universality of the model when solving real problems. In addition, author profiling on social networks is especially difficult, since the slang used by users is often unstructured and contains noise. The problem is also the illiteracy of users, which can be either intentional or unintentional.

The main part of the described drawbacks can be eliminated at the stage of text preprocessing. However, converting the data into an appropriate form does not guarantee high accuracy. There are many factors that affect the accuracy of a model. The most significant of these is the raw data. Training the model will not be effective enough if the data are unreliable, noisy, or unbalanced. For example, real social network users may specify the wrong age in their profile. This requires filtering of the data. The solution to this problem is the implementation of CV models intended for the related problem of determining the age of a user from a photo. The overwhelming majority of the approaches to this solution are based on convolutional NN (CNN) architectures and their more complex modifications. Let us consider in more detail the most effective of these modifications.

Scientific work [20] was devoted to the deep expectation (DEX) method based on the VGG-16 architecture. The essence of this method is that a face is detected on the image, and then age prediction is performed using an ensemble of 20 different architectures. Photos of celebrities from IMDB and Wikipedia sites were used as experimental data. The DEX method showed the best result at the ChaLearn Looking at People competition in 2015, with an ε error of only 0.264975.

The authors of [21] used the dropout SVM approach. The approach was inspired by the success of a deep NN [22,23], where the problem of retraining can be solved by adding a dropout regularization function that turns off some part of the neurons during training. This approach improved the adaptation of neurons to input data. The accuracy of face images collected by the authors reached 70%.

In [24], a new CNN model called the soft stagewise regression network (SSR-Net) was presented. Age determination using a multiclass classifier is carried out in several stages. Each stage is responsible for refining the results of the previous one, which leads to a more accurate assessment. To solve the classification problem, SSR-Net assigns a dynamic range to each age class, allowing it to shift and scale according to the input image. The main advantage of the resulting model is compactness; it is only 0.32 MB in size. Despite its compactness, the performance of SSR-Net is close to modern methods, whose model sizes are often 1500 times larger. The MAE of the model was 2.52.

The study in [25] was based on the hyperplane ranking algorithm. The proposed approach uses age tags to predict rank. The age rank was obtained by aggregating a series of binary classification results. The FG-NET dataset was used for the experiments. The results showed that the learning strategy chosen by the authors exceeded the traditional approaches of classification, regression, and ranking. The system shows the best results for images with a neutral facial expression, as facial emotions negatively affect the results of the age assessment. The MAE value for this approach was 3.82.

The research presented in [26] demonstrated the VGG-Face model, developed on the basis of the well-known VGG-Very-Deep-16 architecture. The performance of the model was evaluated on Labeled Faces in the Wild [27] and YouTube Faces [28]. The resulting accuracy was 98%. In this work, it was proposed to refer to the experience of foreign researchers and use the advantages of natural language processing (NLP) and CV models to solve the problem of determining the age of the author of a Russian-language text.

Table 1 shows the results of the study of methods using feature extraction from photos on similar datasets.

Table 1. Comparison of computer vision (CV) methods.

Method	Public Time	Corpus	Accuracy ± Std (%)
DeepFace [29]	2014	Facebook (4.4M, 4K)	97.35 ± 0.25
DeepID2 [30]	2014	CelebFaces+ (0.2M, 10K)	99.15 ± 0.13
DeepID3 [31]	2015	CelebFaces+ (0.2M, 10K)	99.53 ± 0.10
FaceNet [32]	2015	CelebFaces+ (0.2M, 10K)	99.63 ± 0.09
Baidu [33]	2015	Baidu (1.2M, 18K)	99.77
VGGface [34]	2015	VGGface (2.6M, 2.6K)	98.95
light-CNN [35]	2015	MS-Celeb-1M (8.4M, 100K)	98.8
Center Loss [36]	2016	CASIA-WebFace, CACD2000, Celebrity+ (0.7M, 17K)	99.28
L-softmax [37]	2016	CASIA-WebFace (0.49M, 10K)	98.71
Range Loss [38]	2016	MS-Celeb-1M, CASIA-WebFace (5M, 100K)	99.52
L2-softmax [39]	2017	MS-Celeb-1M (3.7M, 58K)	99.78
Normface [40]	2017	CASIA-WebFace (0.49M, 10K)	99.19
CoCo loss [41]	2017	MS-Celeb-1M (3M, 80K)	99.86
vMF loss [42]	2017	MS-Celeb-1M (4.6M, 60K)	99.58
Marginal Loss [43]	2017	MS-Celeb-1M (4M, 80K)	99.48
SphereFace [44]	2017	CASIA-WebFace (0.49M, 10K)	99.42

Table 1. *Cont.*

Method	Public Time	Corpus	Accuracy ± Std (%)
CCL [45]	2018	CASIA-WebFace (0.49M, 10K)	99.12
AMS loss [46]	2018	CASIA-WebFace (0.49M, 10K)	99.12
Cosface [47]	2018	CASIA-WebFace (0.49M, 10K)	99.33
Arcface [48]	2018	MS-Celeb-1M (3.8M, 85K)	99.83
Ring loss [49]	2018	MS-Celeb-1M (3.5M, 31K)	99.50
MLP [15]	2018	350 texts	67.18
SVM [17]	2011	Flemish Dutch Netlog posts (1.5M)	71.3, 80.8, 88.2
CNN [18]	2019	Twitter (11K posts)	82.21
SVM, Bayesian NN [19]	2011	Russian-language texts (100)	50.1, 48.2
SVM [21]	2014	Labeled Faces in the Wild (LFW) (13K)	70
VGG-Very-Deep-16 [26]	2015	2.6M images	98

3. Methodology

The method for determining the age of the author of the text presented in Figure 1 includes several stages.

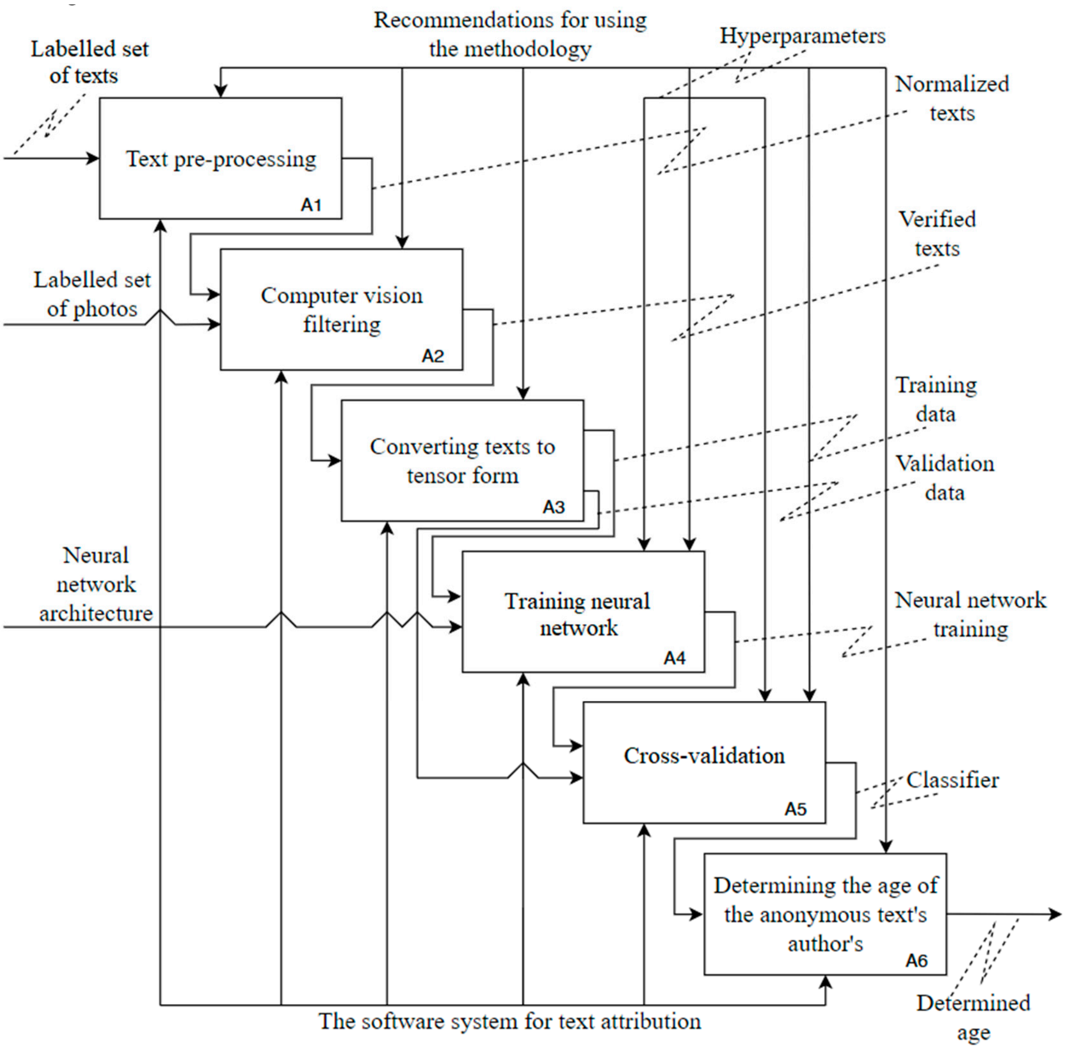

Figure 1. The method for determining the age of the author of the text.

Step 1: The pre-processing stage involves clearing texts from noisy data. Such data negatively affects the classification procedure. One of the problems of social networks is the large amount of spam messages left by users. Therefore, the set of texts is cleared of duplicate messages and comments containing such spam words as asset, subscribe, like, hack, and mutual subscription, among others. Another characteristic of texts from online platforms is the frequent use of emoticons, particularly by audiences up to 18 years old. Short comments consisting mainly of emoticons and including less than five Russian words should also be deleted, as they are not informative enough. All emoticons are replaced with the @emoji tag.

Step 2: The data filtering stage is intended for additional validation of the age data. Often, social media users specify the wrong age on their profile. This may be due to various reasons; however, a frequent purpose is the aim of accessing adult (18+ years old) content or directly registering for an online platform. To solve this problem, it was decided to use photos from users' pages and the CV model as a tool for effectively filtering inaccurate data.

According to the results obtained by other researchers, the VGG-Face model is the most suitable model for solving the problem of determining age from a user's photo. Therefore, it was decided to use it as the basis for filtering inaccurate data. This method also includes several additional steps:

- The age of the user is determined by the photo. Then, 2 years are added or subtracted to or from it.
- If the age specified by the user falls within the interval, then the counter increases. Otherwise, it remains unchanged.
- The age of the user is considered correct if the counter is equal to or more than half the number of the author's photos.

Such actions allow the selection of only reliable data, based on the coincidence of the age specified in the profile and determined by the photo. Messages from users who publish mostly old photos are not included in the dataset.

Step 3: Data conversion into tensor form is carried out using hashing. A dictionary of the most common words is used, and each word from the text is compared with its index from the dictionary when encoding. Hashing is especially effective with significantly sparse data.

Step 4 assumes the use of the filtered data from the previous stage for training a deep NN. The task of dividing texts into age groups is nontrivial. Therefore, research [50] was devoted to identifying the most effective architecture of NN. It was found that the most effective architectures for determining the age of the author of the text were FastText and CRNN [51], a hybrid of a CNN and a recurrent NN (RNN).

FastText architecture is distinguished by some features of input data preprocessing. Skip-gram with negative sampling is used for the vector representation of words. Negative sampling provides negative examples to connect words that are not paired in context. Anywhere from 3 to 20 negative words are selected for each word. Skip-gram ignores word structure, so a model that breaks words into n-grams was added. Usually, the value of n can be from 3 to 6. The whole word is also added to the chain of n-grams. This approach allows working with words that the model has not previously encountered. Hashing is used since the features obtained by splitting into n-grams have an excessively large dimension. We used the following parameters for FastText training:

- Learning rate: 0.1;
- Rate of updates for the learning rate: 100;
- Size of the context window: 5;
- Number of negative samples: 5;
- Loss function: softmax.

The CRNN architecture is an efficient combination for analyzing text sequences. The convolutions of various dimensions allow selection of the most significant features for classification, regardless of their size. Recurrent layers respond to temporary changes and correct context dependencies. The recurrent part of the architecture is represented by LSTM. The choice of LSTM was due to its ability to store long-term dependencies in its memory cells. This is particularly valuable in solving problems of text analysis. The convolutional part is represented by convolutions with filter dimensions 1, 3, and 5. Thus, the CRNN architecture makes it possible to respond to both local and global informative features, as well as short-term and long-term dependencies.

As with the NN architecture, the training parameters also have a serious impact on the final result. Their choice was carried out experimentally, based on the experience of researchers in the NLP field [14–18]:

- Input layer dimension: 128;
- Optimization function: adaptive moment estimation (Adam);
- Loss function: binary cross entropy;
- Batch size: 32;
- Number of training epochs: 50.

In addition to the described architectures, variations of BERT [52] were considered. This is a well-known and effective architecture for text mining. BERT is a bidirectional model based on the classic transformer architecture which combines the advantages of convolutional and recurrent architectures. Classic multilingual BERT (bert-base-multilingual-cased), XLM (xlm-mlm-xnli15-1024) [53], and RoBERTa (xlm-roberta-base) [54] were applied in this study, as these models are well-adapted to the Russian language. We used the default parameters for all applied BERT modifications.

Step 5 involves validating the selected models. The procedure for validation was 10-fold cross-validation. This was used as a way to get reliable assessments and improve the learning process. The most accurate model based on the validation results was chosen as a decision-making tool.

Step 6: At the last step, the model obtained in Step 5 is used to predict the age of the author of the text.

4. Results

Effective text classification requires a large corpus of representative data. Often, there are marked-up texts in various languages in the public domain, but there are no such corpora for solving the problem of determining the age of the author of a Russian-language text. This situation creates an additional task: collecting and marking up a dataset.

For experimental purposes, it was decided to use real data from users of the vk.com online platform, since combating pedophilia in social networks is one of the key areas of application of solutions to this task. The choice of the resource was due to the rapidly growing number of public messages left by users. On average, for the most popular social network in Russia and the CIS, vk.com, this number reached 550,000 messages left by more than 30,000 users.

The data were collected from the platform's community pages using the API. Thus, more than 50,000 links and 70,000 photos from the social network communities were received. This process was automated; a special script retrieved the last 100 entries left on the community pages and their related comments. The total amount of data collected was over 2 million records. Each of these entries included a short comment, an age tag, and five photos from the user's account.

The data were preprocessed according to the author's method. They were cleared of spam, uninformative messages, duplicates, and so on. Then, the preprocessed data were filtered by the VGG-Face model, and 5500 examples were selected from 75,000 texts as a result of applying the data filtering method based on the VGG-Face model. For these texts, the user's age turned out to be reliable.

In this experiment, distributions of the authors and users of vk.com by age before and after filtering were obtained (Figure 2). Based on the resulting distributions, we can conclude that some users deliberately underestimated their ages in their profiles. The graph shows a drop in the area of users under 18 years old and an increase for people over 18.

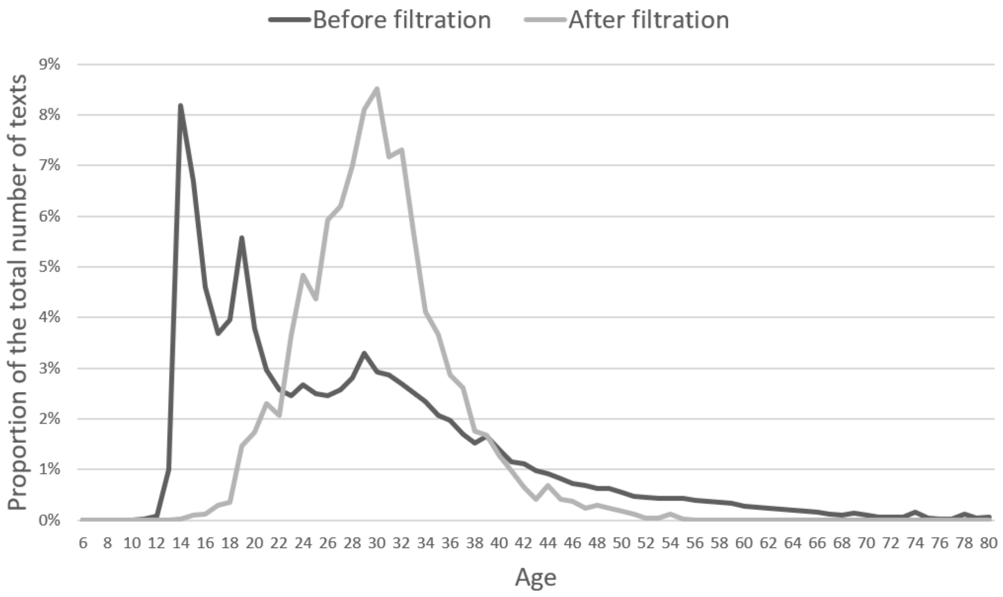

Figure 2. Distribution of users by age before and after filtering.

The obtained distribution was uneven. This was due to the fact that social networks are not very popular among people over 40 in Russia. This fact was additionally confirmed by the official statistics of vk.com. Additionally, filtering negatively affected the number of messages from users under 18. Therefore, about 5000 additional messages by pupils from schools in Tomsk affiliated with Tomsk State University of Control Systems and Radio-electronics (TUSUR) were collected from social networks. The ages of these pupils were all known.

The training was conducted using two datasets. The first set included texts, age labels, and images that were not filtered by the VGG-Face method. The second set included both the filtered and additional messages of the Tomsk pupils.

It was decided to implement both binary and multiclass classification. For this purpose, the data contained in the dataset were divided into categories. In the first case, the sets included two categories: under 18 and over 21 years old. In the second, under 18, from 21 to 27, and over 30 were the categories.

The choice of categories was intended to distinguish between underage and adult users. It should be noted that texts written by authors between the ages of 18 and 21 were not considered when training the models, since the generalization ability of the models at this interval was unsatisfactory and would negatively affect the final result. The situation was similar for the age interval of 27–30 years. The main distinctive features of the selected groups were the level of education (school, university), writing style (official business, conversational, containing slang), as well as vocabulary.

The results of the experiments are presented in Table 2. It should be noted that the implemented approaches are difficult to compare with analogues because there are no existing solutions for determining the age of the author of a short Russian text using an NN. In addition, the accuracy of solutions based on traditional ML methods [19] is much lower than that of NN approaches (15% less on average). Thus, their consideration in this case would not be applicable. The Russian language has a distinctive feature of inflection, so comparing this to methods used for English is also incorrect. Inflection means more complex word formation and a high degree of morphological and syntactic homonymy. These aspects make it difficult to use the original methods. Their accuracy for the Russian language drops to 50%.

Table 2. Results of experiments.

Categories	**Accuracy for Two Categories (%)**		**Accuracy for Three Categories (%)**		**Average Epoch Time**	**Total Training Time**
Models	Raw Data	Filtered Data	Raw Data	Filtered Data		
FastText	66.5	82.1	44.5	62.4	60 s	3600 s
CRNN	63.8	81.8	43	61.1	114 s	5700 s
BERT	65.8	80.3	41.1	60.4	540 s	27,000 s
XLM	49.8	50.1	38	46.2	660 s	33,000 s
RoBERTa	50.3	75.3	40.7	60.1	530 s	26,500 s

The use of a verified dataset improved the initial accuracy of the models by an average of 17.5%. The obtained values of accuracy allow us to conclude that it is advisable to use CV models in order to improve the training set and clean it from inaccurate data.

In addition, the dataset was tested using the only publicly available tool [55] designed to determine gender and age based on quantitative parameters of texts, such as 3–8 *n*-grams. The result was 65% when determining the age.

5. Discussion and Conclusions

As part of the study, the technique for determining the age of the author of a Russian-language text based on the FastText model and the method for filtering user photos using the VGG-Face model was developed.

Great attention was paid to the experimental data. There are no Russian-language corpora to solve this problem, so our own dataset was collected from social networks. Experiments have shown that using raw social network data is inaccurate, since the real age does not coincide with the ages indicated in the profiles. Therefore, the results of methods evaluated on raw data from social networks cannot be trusted.

We proposed the CV algorithm to filter users' messages. Determining the age by photo with VGG-Face and comparing it to the age specified in the profile ensures that the user has a correctly specified age. In this case, users' messages can be used for model training. Experiments have shown that the application of a filtering procedure has a significant impact on the data distribution by age and the results obtained. Another conclusion drawn from experiments on verified data by photos is that users deliberately underestimate their age on social networks. Such actions may be carried out for illegal purposes, particularly for unimpeded communication with users under 18. The analysis confirms the importance of identifying the real age of social network users, mainly for the detection of pedophilia.

The results on the verified corpus are due to the fact that the wrong age can be easily predicted in social media. Therefore, the classification of the original raw data showed worse results. There is a possibility that the user will upload photos that are not his own, but our results show that the method is suitable for this case.

The best result for Russian-language text was obtained using the FastText model on verified data. The cross-validation accuracy was 82.1% for two categories. The obtained accuracy of the method is comparable to the approaches for English, Chinese, and other research, even considering the complexity of the Russian language.

Author Contributions: Conceptualization, A.S.R. and A.A.S. (Artem Alexandrovich Sobolev); methodology, A.V.K. and A.A.S. (Artem Alexandrovich Sobolev); software, A.A.S. (Artem Alexandrovich Sobolev); validation, A.S.R., A.V.K., A.M.F.; formal analysis, A.A.S. (Artem Alexandrovich Sobolev) and A.M.F.; investigation, A.V.K. and A.A.S. (Artem Alexandrovich Sobolev); resources, A.A.S. (Alexander Alexandrovich Shelupanov); data curation, A.S.R. and A.A.S. (Alexander Alexandrovich Shelupanov); writing—original draft preparation, A.V.K. and A.S.R.; writing—review and editing, A.A.S. (Alexander Alexandrovich Shelupanov); visualization, A.M.F.; supervision, A.A.S. (Alexander Alexandrovich Shelupanov); project administration, A.S.R.; funding acquisition, A.A.S. (Alexander Alexandrovich Shelupanov). All authors have read and agreed to the published version of the manuscript.

Funding: The article was prepared as part of the implantation of the Leading Research Center (LRC) Trusted Sensor Systems. Financial support provided by the Ministry of Digital Development, Communications and Mass Media of the Russian Federation and Russian Venture Company (RVC JSC) (Agreement №009/20 dated 04/10/2020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kurtukova, A.; Romanov, A. Identification Author of Source Code by Machine Learning Methods. *SPIIRAS Proc.* **2019**, *18*, 742–766. [CrossRef]
2. Kurtukova, A.; Romanov, A.; Fedotova, A. De-Anonymization of the Author of the Source Code Using Machine Learning Algorithms. In Proceedings of the 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), Yekaterinburg, Russia, 25–27 October 2019; pp. 612–617.
3. Romanov, A.; Kurtukova, A.; Fedotova, A.; Meshcheryakov, R. Natural Text Anonymization Using Universal Transformer with a Self-attention. In Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, 27 November 2019; pp. 22–37.
4. Romanov, A.S.; Vasilieva, M.I.; Kurtukova, A.V.; Meshcheryakov, R.V. Sentiment Analysis of Text Using Machine Learning Techniques. In Proceedings of the 2nd International Conference "R. Piotrowski's Readings LE & AL'2017", Saint Petersburg, Russia, 27 November 2017; pp. 86–95.
5. Kurtukova, A.; Romanov, A.; Shelupanov, A. Source Code Authorship Identification Using Deep Neural Networks. *Symmetry* **2020**, *12*, 2044. [CrossRef]
6. Bianchi, G.; Bruni, R.; Scalfati, F. Identifying e-Commerce in Enterprises by means of Text Mining and Classification Algorithms. *Math. Probl. Eng.* **2018**, *2018*, 1–8. [CrossRef]
7. Bruni, R.; Bianchi, G. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Syst. Appl.* **2020**, *142*, 113001. [CrossRef]
8. Rakhmanenko, I.; Shelupanov, A.; Kostyuchenko, E. Automatic text-independent speaker verification using convolutional deep belief network. *Comput. Opt.* **2020**, *44*, 596–605. [CrossRef]
9. Kostyuchenko, E.Y.; Viktorovich, I.; Renko, B.; Shelupanov, A.A. User Identification by the Free-Text Keystroke Dynamics. In Proceedings of the 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, Russia, 18–25 August 2018; pp. 1–4.
10. Nemati, A. Gender and Age Prediction Multilingual Author Profiles Based on Comment. *FIRE* **2018**, *2266*, 232–239.
11. Nguyen, D.-P.; Trieschnigg, R.B.; Dogruoz, A.S.; Gravel, R.; Theune, M.; Meder, T.; De Jong, F. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014, Dublin, Ireland, 23–29 August 2014; pp. 1950–1961.
12. Peersman, C.; Walter, D.; Vaerenbergh, L. Predicting age and gender in online social networks. In Proceedings of the International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 37–44.
13. Daneshvar, S. User Modeling in Social Media: Gender and Age Detection. Ph.D. Thesis, University of Ottawa, Ottawa, ON, Canada, 2019.
14. Tumanova, K.S. Algorithm for the Classification of Texts in Russian by Age and Gender of the Author. 2011. Available online: https://studylib.ru/doc/2366008/tumanova-kristina---text (accessed on 9 November 2020).
15. Škrlj, B.; Martinc, M.; Kralj, J.; Lavrač, N.; Pollak, S. tax2vec: Constructing Interpretable Features from Taxonomies for Short Text Classification. *Comput. Speech Lang.* **2020**, *65*, 101104. [CrossRef]
16. Chen, J.; Cheng, L.; Yang, X.; Liang, J.; Quan, B.; Li, S. Joint Learning with both Classification and Regression Models for Age Prediction. *J. Physics Conf. Ser.* **2019**, *1168*, 032016. [CrossRef]
17. Abdallah, E.E.; Alzghoul, J.R.; Alzghool, M. Age and Gender prediction in Open Domain Text. *Procedia Comput. Sci.* **2020**, *170*, 563–570. [CrossRef]
18. Wang, L. Multi-Task Learning for Gender and Age Prediction on Chinese Microblog. In Proceedings of the International Conference on Computer Processing of Oriental Languages, Kunming, China, 2–6 December 2016; pp. 189–200.
19. Ustalov, D.; Filchenkov, A.; Pivovarova, L.; Žižka, J. Artificial Intelligence and Natural Language. In Proceedings of the 6th Conference, AINL 2017, Saint Petersburg, Russia, 20–23 September 2017; pp. 170–177.

20. Rothe, R.; Timofte, R.; Van Gool, L. DEX: Deep EXpectation of Apparent Age from a Single Image. In Proceedings of the IEEE International Conference on Computer Vision Workshops 2015, Santiago, Chile, 11–12 December 2015; pp. 252–257.
21. Eidinger, E.; Enbar, R.; Hassner, T. Age and Gender Estimation of Unfiltered Faces. *IEEE Trans. Inf. Forensics Secur.* **2014**, *10*, 2170–2179. [CrossRef]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**. [CrossRef]
23. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
24. Yang, T.; Huang, Y.; Lin, Y.; Hsiu, P.; Chuang, Y. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 1078–1084.
25. Chang, K.-Y.; Chen, C.-S. A Learning Framework for Age Rank Estimation Based on Face Images with Scattering Transform. *IEEE Trans. Image Process.* **2015**, *24*, 785–798. [CrossRef] [PubMed]
26. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference 2015, Swansea, UK, 7–10 September 2015; Volume 1, pp. 41.1–41.12.
27. Huang, G.; Mattar, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report; University of Massachusetts: Amherst, MA, USA, 2007.
28. Wolf, L.; Hassner, T.; Maoz, I. Face Recognition in Unconstrained Videos with Matched Background Similarity. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; pp. 529–534.
29. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
30. Sun, Y.; Chen, Y.; Wang, X.; Tang, X. Deep Learning Face Representation by Joint Identification-Verification. In Proceedings of the NIPS 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 1988–1996.
31. Sun, Y.; Liang, D.; Wang, X.; Tang, X. Deepid3: Face recognition with very deep neural networks. *arXiv* **2015**, arXiv:1502.00873.
32. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
33. Liu, J.; Deng, Y.; Bai, T.; Wei, Z.; Huang, C. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv* **2015**, arXiv:1506.07310.
34. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
35. Wu, X.; He, R.; Sun, Z.; Tan, T. A light CNN for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2884–2896. [CrossRef]
36. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 499–515.
37. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. In Proceedings of the ICML 2016, New York, NY, USA, 19–24 June 2016; Volume 2, p. 7.
38. Zhang, X.; Fang, Z.; Wen, Y.; Li, Z.; Qiao, Y. Range Loss for Deep Face Recognition with Long-Tailed Training Data. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 5409–5418.
39. Ranjan, R.; Castillo, C.D.; Chellappa, R. L2-Constrained Softmax Loss for Discriminative Face Verification. *arXiv* **2017**, arXiv:1703.09507.
40. Wang, F.; Xiang, X.; Cheng, J.; Yuille, A.L. Normface: L2 Hypersphere Embedding for Face Verification. In Proceedings of the 25th ACM international Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1041–1049.
41. Liu, Y.; Li, H.; Wang, X. Rethinking feature discrimination and polymerization for large-scale recognition. *arXiv* **2017**, arXiv:1710.00870.

42. Hasnat, M.; Bohne, J.; Milgram, J.; Gentric, S.; Chen, L. Von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv* **2017**, arXiv:1706.04264.
43. Deng, J.; Zhou, Y.; Zafeiriou, S. Marginal Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 60–68.
44. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. Sphereface: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 212–220.
45. Qi, X.; Zhang, L. Face recognition via centralized coordinate learning. *arXiv* **2018**, arXiv:1801.05678.
46. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive Margin Softmax for Face Verification. *IEEE Signal Process. Lett.* **2018**, *25*, 926–930. [CrossRef]
47. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J. Cosface: Large Margin Cosine Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
48. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 4690–4699.
49. Zheng, Y.; Pal, D.K.; Savvides, M. Ring Loss: Convex Feature Normalization for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5089–5097.
50. Sobolev, A.A.; Kurtukova, A.V.; Romanov, A.S.; Vasilieva, M.I. Electronic Instrumentation and Control Systems. Determination of the Age of the Author of an Anonymous Text. In Proceedings of the XV International Scientific and Practical Conference 2019, Kyiv, Ukraine, 24–25 October 2019; Volume 2, pp. 128–131.
51. Lai, S.; Xu, L.; Liu, K. Recurrent Convolutional Neural Networks for Text Classification. In Proceedings of the 29 AAAI Conference on Artificial Intelligence 2015, Austin, TX, USA, 25–29 January 2015; pp. 2267–2273.
52. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
53. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. *arXiv* **2019**, arXiv:1901.07291.
54. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv* **2019**, arXiv:1911.02116.
55. Demo Versions of a Computer Program for Diagnosing the Gender and Age of a Participant in Internet Communication Based on the Quantitative Parameters of His Texts. Available online: https://github.com/sag111/author_gender_and_age_profiling_with_style_imitation_detection (accessed on 9 November 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.