Age Determination of the Social Media Post's Author Using Deep Neural Networks and Facial Processing Models

Aleksandr Romanov [a], Anna Kurtukova [a], Artem Sobolev [a] and Anastasia Fedotova [a]

[a] *Tomsk State University of Control Systems and Radioelectronics (TUSUR), 146, Red Army street Tomsk, 634045, Russia*

Abstract

This article describes an original approach based on deep neural network models to the problem of determining the age of the text author. A detailed description of the author's method and the results of its application is demonstrated. The analysis of tools for filtration of unreliable data in photo age determination, methods of text author's age determination are also present in the paper. This method can be considered a deserving competitor for other approaches due to the obtained result - 82% accuracy.

Keywords [1]

Authorship, age, text analysis, computer vision, social networks, FastText, VGG-Face, CRNN.

1. Introduction

This research is indirectly related to intellectual analysis of the text [1-4], establishing the authorship of the text [5, 6], and tonal analysis. This fact establishes the high relevance of the topic. The most informative attributes of attribution are the author's gender and age. Their use makes it possible to effectively separate the true / false authors of the text, as well as to increase the separating ability of the model in classification.

Possible applications include information security, science and commerce (e.g., as a way to optimize targeted advertising) and science (as a tool for linguistic research), and forensic science. In relation to the latter, it is important to differentiate users of social networks in order to restrict children from adult content and to prevent threats of pedophilia.

In addition, solving these issues is very important for the field of forensic science in particular.

- detection of the author of an anonymous note with threats;
- Age-based differentiation of social media users to counter pedophilia and prevent children from accessing adult content;
- identification of the author of the suicide note, if it is necessary to confirm the version of suicide.

Everyone has a unique writing style. The author's style consists in a variety of vocabulary, speech circulation, building a sentence structure, vocabulary, using certain linguistic structures. Such features make it possible to differentiate people into groups by age.

2. Purpose and objectives of the research

The purpose of the study is to determine the age of the author based on an anonymous text. Designated tasks: collecting data from a social network, pre-processing of texts, study and analysis of text's classification methods, determination of age from a photo, choice of methods and algorithms, data filtering using CV technologies, classification of the original and processed data, analysis of the results.

Ultimately, the use of CV technologies for data filtering gave a result, the classification accuracy increased by almost 20%

3. Literature review

Many studies are devoted to determining the age of the author of the text. In [7], four machine learning algorithms (ml): logistic regression, naive Bayesian classifier (NB), gradient boosting, and multi-layer perceptron (MLP) were used as classifiers. The data for the study consisted of 350 texts divided into 3 classes (15-19, 20-24, 25, and older). The best result was obtained when using MLP. The accuracy was 67.18%. The average accuracy of the ensemble with the used classifiers was 51%.

The TweetGenie system for determining the user's age is presented in the article [8]. The system combines sociolinguistics ideas and data obtained from social networks user's pages in social networks. The corpus was based on the tweets of 3,000 users. The authors have identified 3 categories of users (under 20, from 20 to 40 and over 40 years). The mean absolute error (MAE) for this system was 4.844, the Pearson coefficient was 0.866. The results show the limitations of current text-based age-setting approaches. Considering this aspect as a fixed variable ignores the variety of ways in which a user is unique in natural language. A disadvantage of the system is the tendency to underestimate the age of older people: Twitter users are perceived to be younger than they are.

The article [9] is devoted to determining the age of the text author using the Support Vector Machine (SVM). Texts from social networks were used in the study. The authors divided user posts into different groups. Short texts from 12 tokens were used in the experiments. For the class "16+ years", the accuracy was 71.3%, for the class "18+ years" - 80.8%, and for the class "25+ years" - 88.2%. n-gram words were the best signs for age classification in this study.

The authors of [10] offer several classification models, using methods of deep learning and classical machine learning. These methods determine the age of the Twitter user based on the text content of tweets. The age classifier was trained on a data set of 11160 users, which includes combinations of words and symbolic n-grams as attributes derived from Latent semantic analysis (LSA). When solving the problem of author's profiling in the framework of PAN-2018 (international competition on tasks of digital text forensics and stylometry), this model reached the best accuracy - 82.21% for the English language.

Article [11] is devoted to the comparison of SVM and Bayesian neural network (NN). The texts for the experiments were obtained from online diaries and divided into 4 age categories: up to 18, 20-27, 30-37, 40 and more years. The accuracy obtained by the SVM model was 50.1%, and the Bayesian NN — 48.2%.

Many of the previously mentioned methods for determining the age of the author of a text have low accuracy. Another drawback is their strict length dependency. They are oriented either to short texts (posts in social networks and blogs, comments, etc.) or to long texts (articles, posts, reviews, etc.). This approach is wrong, as it does not allow to estimate the universality of the model when solving real problems. The slang that users use in their posts is often not structured and contains noise. Intentional and unintentional illiteracy also makes classification difficult. The main part of the described shortcomings can be eliminated at the stage of text preprocessing, but bringing the data into an appropriate form does not guarantee high accuracy. Many factors affect the accuracy of a model. The most significant of these is the original data. If they are unreliable, noisy, or unbalanced, training the model will not be effective enough. For example, real users of social networks may deliberately indicate the wrong age in their profile. This requires filtering the data. The solution to this problem can be the use of computer vision (CV) models intended for the related problem of determining the age of a user from a photograph. The overwhelming majority of approaches to its solution are based

on convolutional neural network (CNN) architectures and their more complex modifications. Let's consider in more detail the most effective of them.

The work [12] is based on VGG-16 architecture and the Deep EXpectacion (DEX) method. The essence of the method consists in detecting a face in a photo, for which an ensemble of 20 different architectures is used to predict age. As experimental data we used celebrity photos from IMDB and Wikipedia. DEX method showed the best result in the ChaLearn Looking at People contest in 2015, and ε-error was only 0.264975.

In [13] SVM approach was used. The approach is inspired by the success of deep NN [14, 15], 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 improves the adaptation of neurons to input data. The accuracy of the face images collected by the authors reaches 70%.

A new CNN model, Soft Stagewise Regression Network (SSR-Net), is presented in [16]. Age determination using a multi-class classifier is performed in several stages. Each stage of age determination is responsible for refining the results of the previous one, which allows for a more accurate assessment. To solve the task of quantization of age grouping by classes, SSSR-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 obtained model is its compactness - it takes only 0.32 MB. Despite this, SSR-Net performance is close to modern methods, the size of the models which are often 1500 times larger. The MAE of this model was 2.52.

The study [17] is based on the hyperplane ranking algorithm. In the proposed approach, age labels are used to predict rank. Age rank is 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 best results of the system are shown in images with neutral facial expression, emotions influence the results of age assessment to the detriment. The MAE value for this approach was 3.82.

The article [18] shows the VGG-Face model developed on the basis of the well-known VGG-Very-Deep-16 architecture. The performance of the model was estimated on marked faces in the wild [19] and YouTube Faces [20]. The obtained accuracy was 98%. In this paper it is proposed to refer to the experience of foreign researchers and to use the advantages of NLP and CV models to solve the problem of determining the age of the author of the Russian-language text.

Comparison of the results presented in works devoted to this topic is shown in Table 1.

Table 1

Comparison of methods proposed in works

Method	Accuracy±Std(%)
MLP [7]	67.18
SVM [9]	71.3, 80.8, 88.2
CNN [10]	82.21
SVM, Bayesian NN [11]	50.1, 48.2
SVM [13]	70
DeepFace[21]	97.35±0.25
VGG-Very-Deep-16 [18]	98
light-CNN [27]	98.8
DeepID2 [22]	99.15±0.13
DeepID3 [23]	99.53±0.10
FaceNet [24]	99.63±0.09
Baidu [25]	99.77
VGGface [26]	98.95
light-CNN [27]	98.8
Center Loss [28]	99.28
L-softmax [29]	98.71
Range Loss [30]	99.52
L2-softmax [31]	99.78
Normface [32]	99.19

CoCo loss [33]	99.86
vMF loss [34]	99.58
Marginal Loss [35]	99.48
SphereFace [36]	99.42
CCL [37]	99.12
AMS loss [38]	99.12
Cosface [39]	99.33
Arcface [40]	99.83
Ring loss [41]	99.50

4. Methodology

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

1. Clearing text from the excessive and disturbing data is performed during pre-processing. Noises make it difficult to perform classification. Spam messages are often found in social networks, which is why texts are cleared of such spam words as "asset", "subscription", "like", "hacking", "mutual subscription", as well as from duplicates. Another characteristic feature of the online platform texts is the frequent use of emoticons, in particular by the audience under 18 years of age. All emoticons are replaced with the "@emoji". Short comments that consist mainly of emoticons and include less than 5 Russian words should also be deleted, as they are not informative enough.

2. For additional data verification, a filtering step is required. Often, users of social networks indicate the wrong age in their profile. This may be due to various reasons; however, this is mostly done for the purpose of accessing 18 or more content or registering directly on an online platform. As a tool for effective filtering of inaccurate data, it was decided to use photos from users ' pages in the social network and a CV model.

According to the results obtained by other researchers, the VGG-Face 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 method that includes several steps: the user's age is determined by the photo, then 2 years are added/subtracted to it. If the age specified by the user falls within the interval, then the counter increases, differently 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.

3. Using the data filtered in the previous stage is necessary for performing the deep neural network training stage. The task of dividing texts into age groups is not trivial, so a separate preliminary study was devoted to it [42], aimed at identifying the most effective architecture of neural networks. It was found that FastText and CRNN [43] - a hybrid of CNN and recurrent neural network (RNN) are the most effective architectures for determining the author's age from the text.

Skip-gram with negative sampling is used for vector representation of words. Negative sampling provides negative examples — to connect words that are not pairs in context. For one word, from 3 to 20 negative words are selected. Skip-gram ignores word structure, so a model was added that breaks words into n-grams. Usually, the value of n can be from 3 to 6. The full word is also written to the chain of n-grams. This approach allows to work with words that the model has never met before. Hashing was used to process a large dimension of features obtained by splitting into n-grams.

Recurrent layers respond to temporary changes and correct contextual dependencies. The CNN architecture is an effective combination for analyzing text sequences; the convolutions of various dimensions available in it make it possible to select the most significant features for classification, regardless of their size.

4. The final stage includes the validation of the selected NN models. The procedure for validation was cross-checking for 10 blocks. It is used as a way to get a reliable assessment and improve the learning process. The most accurate model based on validation results is chosen as a decision-making tool.
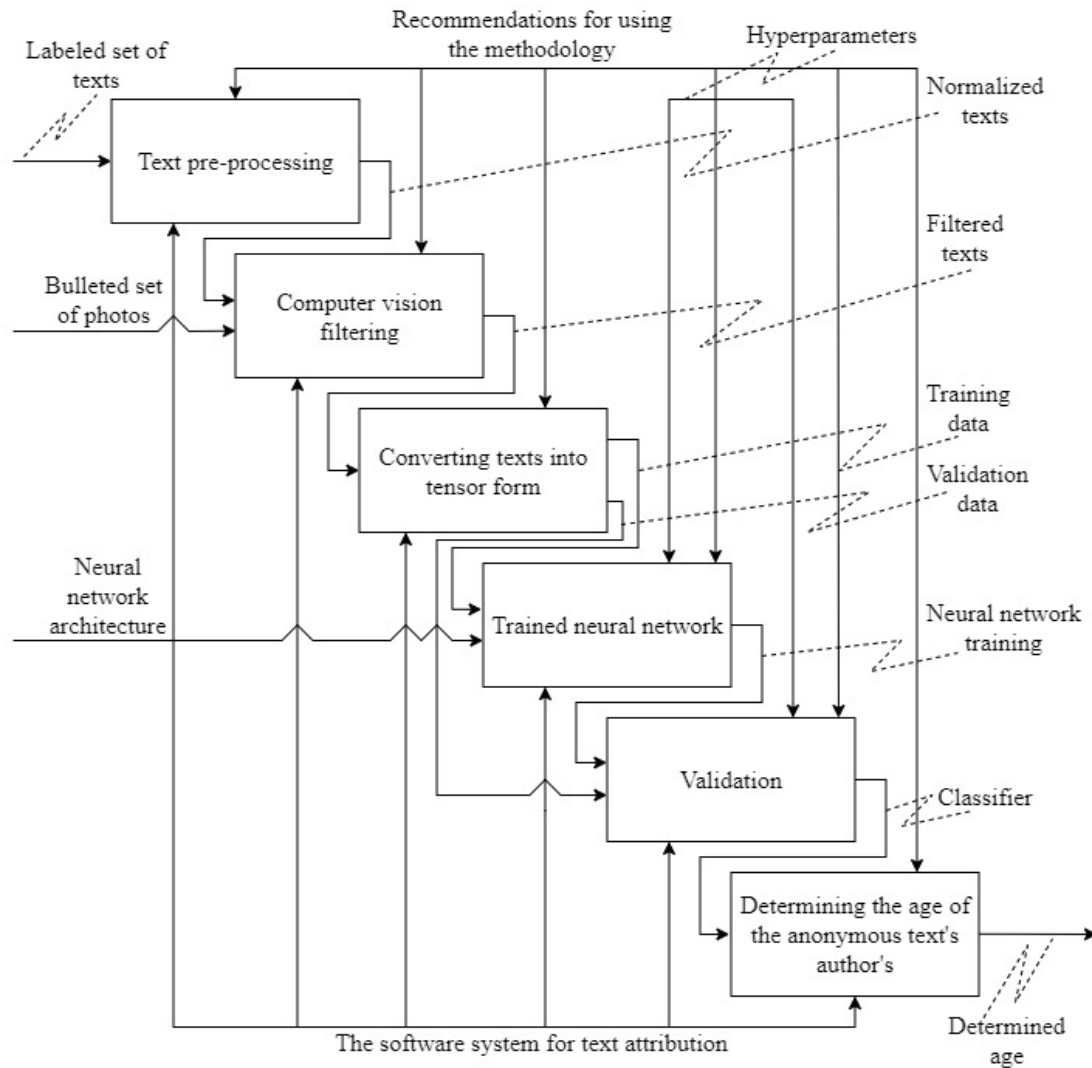
Figure 1: The technique for determining the age of the text's author

5. Results

To carry out an effective classification of texts, a large set of representative data is required. There is no set of marked-up texts for solving the problem of determining the age of the author of a Russian-language text. Therefore, such a data set was formed and marked up on its own.

Since countering pedophilia in social networks is one of the key areas of application of solutions to this problem, we used real data from users of The Vkontakte online platform. On average, the most popular social network in Russia and the CIS, Vkontakte, the number of daily messages reaches 550 thousand messages left by more than 30 thousand users. Data was collected from community pages using the API. Thus, more than 50 thousand links and 70 thousand photos were received. The data collection process was automated - a special script extracted the last 100 entries left on the community pages and their accompanying comments. The total amount of data received was more than 2 million records. Each of these entries includes: a short comment, 5 photos from the author's account, and a tag containing information about the age of the comment's author.

The data obtained by the described method was pre-processed according to the author's method: cleared of spam, uninformative messages, duplicates, etc. The preprocessed data was then filtered by the VGG-Face model. As a result of applying the data filtering method based on the G-Face model, 5.5 thousand texts were selected out of 75 thousand texts where the user's age was reliable.

In this experiment, distributions of VK authors by age before and after filtering (Figure 2) was obtained. The graph shows a drop in the share of users under the age of 18 and an increase in the share of users over the age of 18. Based on the obtained distributions, we can conclude that some users deliberately underestimate their age in the profile.
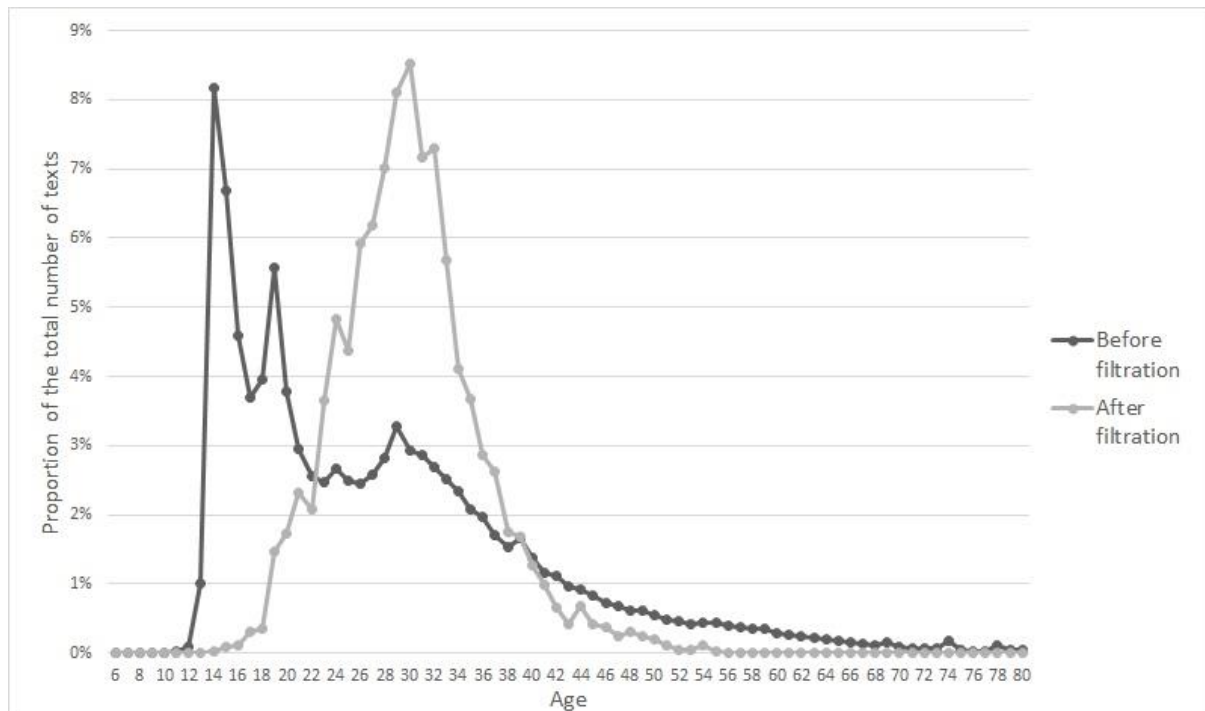

Figure 2: Distribution of users by age before and after filtering

Two data sets were created for model training. The first set included texts, age marks and images that had not been filtered using the VG-Face method, while the second set included filtered data respectively. It was decided to implement both binary and multiclass classification. For this purpose, the data contained in the set were divided into categories. In the first case, the sets included two categories: up to 18 years old and over 21 years old, in the second set - up to 18 years old, from 21 to 27 years old and over 30 years old. The choice of categories is related to the need to distinguish between juvenile and adult users. It should be noted that the texts written by authors aged 18 to 21 years were not taken into account when teaching the models, because the separation ability of the models in this interval is unsatisfactory. This has a negative impact on the end result. The situation is similar to the 27-30 years interval. The main distinguishing features of the selected groups are the level of education (school, university), writing style (official business, conversational, slang), and vocabulary.

The results of the experiments are presented in Table. 2. Filtering unreliable data allowed to improve the original accuracy of the models on average by 17.5%. The obtained accuracy values allow to conclude that it is advisable to use CV-models to improve the training set and clean up unreliable data.

Table 2

Results of experiments

Categories	Accuracy for two categories		Accuracy for three categories	
Models	Raw data	Filtered data	Raw data	Filtered data
FastText	66,5%	82,1%	44,5%	62,4%
CRNN	63,8%	81,8%	43%	61,1%

In addition, the dataset was tested on another tool [44] designed to determine gender and age based on quantitative parameters of texts, such as n-grams 3-8 characters. The result is 65% when determining the age.

6. Discussion

The fact that the wrong age is easily indicated in social networks is associated with an improvement in the result, so the classification based on the original data shows the worst result. The guarantee of correct age verification is to determine the age from the photo and compare it with the age specified in the profile. There is a chance that the user will upload photos that are not their own, but the result seems to be improving, which indicates the suitability of the described method.

7. References

The method described in the article is intended for determining the age of the author of a text written in Russian. The approach includes FastText and VGG-Face models. The latter is used to filter user photos.

During the study, it was found. that social media users intentionally indicate an age less than the real one. In particular, this may be done for illegal actions-communication with young users. In turn, the presented approach helps to counter such threats, and, as a result, pedophilia. The best result was 82.1% for the two categories. The result is obtained using the FastText model and filtered VGG-Face data. The obtained accuracy of the model is comparable to the approaches of foreign researchers and is sufficient for use in solving real problems.

References
[1] A.S. Romanov, R.V. Meshcheryakov, Identification of authorship of short texts with machine learning techniques, in: Proceedings of the Conference Dialog, 2010, vol. 9, no. 16, pp. 407–413.
[2] A.S. Romanov, R.V. Meshcheryakov, Gender identification of the author of a short message, in: Proceedings of the Conference Dialog, 2011, vol. 10, no. 17, pp. 620–626.
[3] A. Romanov, A. Kurtukova, A. Fedotova, R. Meshcheryakov, 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), 2019, pp. 22–37.
[4] A.S. Romanov, M.I. Vasilieva, A.V. Kurtukova, R.V. Meshcheryakov, Sentiment Analysis of Text Using Machine Learning Techniques, in: Proceedings 2nd International Conference «R. PIOTROWSKI'S READINGS LE & AL'2017», 2018, pp. 86–95.
[5] A.S. Romanov, A.A. Shelupanov, S.S. Bondarchuk, Generalized authorship identification technique. Proceedings of TUSUR University. (2010) 1.21: 108–112.
[6] A.S. Romanov, A.A. Shelupanov, R.V. Meshcheryakov, Development and research of mathematical models, methods and software tools of information processes in the identification of the author of the text, Tomsk: V-Spektr, 2011.
[7] A. Nemati, Gender and Age Prediction Multilingual Author Profiles Based on Comment, in: FIRE-WN, 2018, pp. 232–239.
[8] D-P. Nguyen, R.B. Trieschnigg, A.S. Dogruoz, R. Gravel, M. Theune, T. Meder, F. de Jong, 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, pp. 1950–1961.
[9] C. Peersman, D. Walter, L. Vaerenbergh, Predicting age and gender in online social networks, in: Proceedings of the International Conference on Information and Knowledge Management, 2011, pp. 37–44.
[10] S. Daneshvar, User Modeling in Social Media: Gender and Age Detection, Ph.D. thesis, University of Ottawa, 2019.
[11] K.S. Tumanova, Algorithm for the classification of texts in Russian by age and gender of the author, 2011. URL: https://studylib.ru/doc/2366008/tumanova-kristina---text.
[12] R. Rothe, R. Timofte, L. Van Gool, DEX: Deep EXpectation of apparent age from a single image, in: IEEE International Conference on Computer Vision Workshops, 2015, pp. 252-257.
[13] E. Eidinger, R. Enbar, T. Hassner, Age and gender estimation of unfiltered faces, in: IEEE Transactions on Information Forensics and Security, vol. 10, 2014, pp. 2170–2179.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton. ImageNet classification with deep convolutional neural networks. Neural Inform. Process. Syst. (2012) 1.2: 4.

[15] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors. ArXiv preprint arXiv:1207.0580, 2012.

[16] T. Yang, Y. Huang, Y. Lin, P. Hsiu, Y. Chuang, 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), 2018, pp. 1078–1084.

[17] K. Chang, C. Chen, A Learning Framework for Age Rank Estimation Based on Face Images with Scattering Transform, in: IEEE Transactions on Image Processing, 2015, vol. 24, no. 3, pp. 785–798.

[18] O. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, in: Proceedings of the British Machine Vision Conference, 2015, vol. 1, pp. 41.1–41.12.

[19] G. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, University of Massachusetts, Amherst, Technical Report, 2007.

[20] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 529–534.

[21] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.

[22] Y. Sun, Y. Chen, X. Wang, and X. Tang, Deep learning face representation by joint identification-verification, in: Advances in neural information processing systems, 2014, pp. 1988–1996

[23] Y. Sun, D. Liang, X. Wang, and X. Tang, Deepid3: Face recognition with very deep neural networks, arXiv preprint arXiv:1502.00873, 2015

[24] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[25] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, Targeting ultimate accuracy: Face recognition via deep embedding, arXiv preprint arXiv:1506.07310, 2015.

[26] O. M. Parkhi, A. Vedaldi, A. Zisserman et al., Deep face recognition, in: BMVC, 2015, vol. 1, no. 3, p. 6.

[27] X. Wu, R. He, Z. Sun, and T. Tan, A light CNN for deep face representation with noisy labels, in: IEEE Transactions on Information Forensics and Security, 2018, vol. 13, no. 11, pp. 2884–2896.

[28] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on computer vision. Springer, 2016, pp. 499–515.

[29] W. Liu, Y. Wen, Z. Yu, and M. Yang, Large-margin softmax loss for convolutional neural networks, in: ICML, 2016, vol. 2, no. 3, p. 7

[30] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5409– 5418.

[31] R. Ranjan, C. D. Castillo, R. Chellappa, L2-constrained softmax loss for discriminative face verification, arXiv preprint arXiv:1703.09507, 2017.

[32] F. Wang, X. Xiang, J. Cheng, A. L. Yuille, Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1041–1049.

[33] Y. Liu, H. Li, X. Wang, Rethinking feature discrimination and polymerization for large-scale recognition, arXiv preprint arXiv:1710.00870, 2017.

[34] M. Hasnat, J. Bohne, J. Milgram, S. Gentric, L. Chen, et al., Von mises-fisher mixture model-based deep learning: Application to face verification, arXiv preprint arXiv:1706.04264, 2017.

[35] J. Deng, Y. Zhou, S. Zafeiriou, Marginal loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 60−68.

[36] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.

[37] X. Qi, L. Zhang, Face recognition via centralized coordinate learning, arXiv preprint arXiv:1801.05678, 2018.

[38] F. Wang, J. Cheng, W. Liu, and H. Liu, Additive margin softmax for face verification, in: IEEE Signal Processing Letters, 2018, vol. 25, no. 7, pp. 926–930.

[39] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.

[40] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[41] Y. Zheng, D. K. Pal, M. Savvides, Ring loss: Convex feature normalization for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5089–5097.

[42] A.A. Sobolev, A.V. Kurtukova, A.S. Romanov, M.I. Vasilieva, Determination of the age of the author of an anonymous text. Electronic instrumentation and control systems, in: Proceedings of the XV International Scientific and Practical Conference, 2019, vol. 2, no. 12, pp. 128–131.

[43] S. Lai, Xu L., Liu K. Recurrent Convolutional Neural Networks for Text Classification, in: Proceedings of the 29 AAAI Conference on Artificial Intelligence, 2015, pp. 2267–2273.

[44] 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. URL: https://github.com/sag111/author_gender_and_age_profiling_with_style_imitation_detection .