International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2022, 6 (1), 45-50
DOI: DOI: 10.36287/ijmsit.6.1.45
Received: May 20, 2022; Accepted: June 20, 2022

Author Identification with Machine Learning Algorithms

İbrahim Yülüce[1*,2], Feriştah Dalkılıç[3]

[1*]*The Graduate School of Natural and Applied Sciences, Dokuz Eylul University, İzmir, Turkey (ibrahim.yuluce@ceng.deu.edu.tr) (ORCID: 0000-0002-3652-7184)*
[2]*Department of Computer Engineering, Ege University, İzmir, Turkey (ibrahim.yuluce@ege.edu.tr) (ORCID: 0000-0002-3652-7184)*
[3]*Department of Computer Engineering, Dokuz Eylul University, İzmir, Turkey (feristah.orucu@deu.edu.tr) (ORCID: 0000-0001-7528-5109)*

Abstract – Author identification is one of the application areas of text mining. It deals with the automatic prediction of the potential author of an electronic text among predefined author candidates by using author specific writing styles. In this study, we conducted an experiment for the identification of the author of a Turkish language text by using classical machine learning methods including Support Vector Machines (SVM), Gaussian Naive Bayes (GaussianNB), Multi Layer Perceptron (MLP), Logistic Regression (LR), Stochastic Gradient Descent (SGD) and ensemble learning methods including Extremely Randomized Trees (ExtraTrees), and eXtreme Gradient Boosting (XGBoost). The proposed method was applied on three different sizes of author groups including 10, 15 and 20 authors obtained from a new dataset of newspaper articles. Term frequency-inverse document frequency (TF-IDF) vectors were created by using 1-gram and 2-gram word tokens. Our results show that the most successful method is the SGD with a classification performance accuracy of 0.976% by using word unigrams and most successful method is the LR with a classification performance accuracy of 0.935% by using word bigrams.

Keywords *– Author identification, Natural Language Processing, Tf-Idf, Text Mining, Machine Learning*

Citation: Yülüce, İ., Dalkılıç, F. (2022). Author Identification with Machine Learning Algorithms. International Journal of Multidisciplinary Studies and Innovative Technologies, 6(1): 45-50.

I. INTRODUCTION

Author profiling, authorship verification, and author identification are the applications of text mining. Author profiling is the examination of authors' texts to determine their class, including gender, age group, etc. Authorship verification is the task of determining whether two or more texts were written by the same author by analyzing linguistic patterns. Author identification deals with estimating the author of an anonymous text from a predefined set of candidate authors.

In this study, we deal with the authorship identification task that is used in many areas including literary studies, history and forensic linguistics. The need to identify the content creator on the internet, detect plagiarism and prevent copyright infringement has increased the interest in authorship identification. In the identification process, stylometric features expose the patterns that appear in the texts belonging to a specific author. Various number of features have been presented including vocabulary richness measures, syntactical features, function words frequencies, character n-gram frequencies, latent semantic analysis (LSA), and Bag-of-Words (BOW) [1, 2, 3] in many previous studies. In addition, deep learning and machine learning based methods have been employed for the feature extraction and author identification tasks in recent studies [4].

In the scope of this study, we collected the newspaper articles of the top three online news portals of Turkey from 2005 to the present. We created a novel dataset to use in the task of identifying the author of a Turkish language text. By using this dataset, we created sub-datasets of different sizes, including 10, 15 and 20 authors with the highest number of articles. Author identification is a multi-class classification problem that deals with labeling an anonymous text with one of the potential authors. We tested some classical machine learning methods including Support Vector Machines (SVM), Gaussian Naive Bayes (GaussianNB), Multi-Layer Perceptron (MLP), Logistic Regression (LR), Stochastic Gradient Descent (SGD) and ensemble learning methods including Extremely Randomized Trees (ExtraTrees), and eXtreme Gradient Boosting (XGBoost) by python implementations using scikit-learn library. Additionally, we used the JPype python module to provide full access to Zemberek Java Library and the matplotlib and seaborn libraries for data visualization. The contributions of our work are comparison of classical machine learning methods with the new generation ensemble learning methods and reveal the effect of working with different sizes of the author candidate pool and n-grams.

The rest of this paper is structured as follows: In section 2, we give some important studies performed in this topic. Section 3 gives an overview about the dataset, preprocessing tasks and classification algorithms used in our experiments. In section 4, we present the experimental results obtained by the proposed method under different cases. Finally, the last section concludes the paper and presents the relevant future work.

II. Literature Review

The task of identifying authors has been studied in different languages for different purposes since 2000. An important part of the literature consists of studies on English language [4, 5, 6, 7, 8]. There are also many studies done in many different languages including Japanese [9], Mongolian [10], Persian [11], Albanian [12], Indian [13, 14], Brazilian [15], Russian [16, 17], German [18], and Arabic [19]. When the existing studies were examined, it was seen that different types of data sets were used for author identification tasks. Some studies have been carried out on newspaper articles [4, 15, 18, 19], while others were carried out on poems [13], novels [11, 12, 16], email content [20], song lyrics [21], source codes [22], or tweets, blog posts, and forums [8, 9, 23]. In some cases, different types of data sources were combined or compared [17, 25]

Early studies in author identification focused on different stylometric techniques. These techniques are based on identification of style markers including lexical and character features or syntactic and semantic features that quantify writing style [9, 26]. The style markers can be exemplified as sentence length, function word and character n-gram frequencies, the number of verbs and punctuation marks in the sentences, vocabulary richness measures etc. With the development and widespread use of machine learning models over the last decade, machine learning-based author identification has become a promising solution for author identification. Mohsen et al. [4] applied a deep learning method with name Stacked Denoising Auto-Encoder for extracting document features and then used the SVM classifier. They used a subset of RCV1 dataset, which contains 100 documents from each of the top 50 authors and reached classification accuracy up to 95.12% under different settings. In [5] authorship attribution experiments were carried out using a Feedforward Neural Network model (FNN) and LR and 95.93% of accuracy was achieved on one of the four widely used datasets. In another recent study [6], pre-trained language models were applied in the field of author identification. They demonstrated that BERT and ELMo pre-trained models achieve the best results (as 92.86%) on a cross-domain dataset. Ramezani [7] employed seven well-known classifiers by using the TF_IDF scheme on two English and Persian datasets and obtained 0.902 and 0.931 accuracy, respectively. Fourkioti et al. [8] combined the three language models based on characters, words, and POS trigrams and achieved the best generalization accuracy of 96% on movie reviews.

A few research on author identification have been carried out in Turkish language. Some studies are based on NLP techniques, while others are based on machine learning techniques. One of the first studies in this field was [26]. Diri and Amasyalı extracted 22 style markers for the 18 different authors to determine the author of an anonymous text. They obtained a success rate of %84 on average. Örücü and Dalkılıç [27] proposed two methods for determining the corresponding author of an anonymous text. Author-specific N-gram Method and Support Vector Machine (SVM) were applied to newspaper columns of 16 authors. The first method reached a success ratio of 87% with 1-grams while SVM had a success ratio of 77% with 2-grams.

Atar et al. [25] collected the columns from the electronic archives of two different newspapers and created a dataset containing 100 training and 20 test articles for each of the 237 authors. They trained the Word2vFisher and Doc2Vec models using a large corpus in Turkish and used the SVM classifier to classify the columns. They stated that the Skip-Gram approach is more successful when compared to the CBOW approach. Kuyumcu et al. [28] used the same dataset in [25] and applied the Tf-Idf weighting method for the vector space that was a combination of word 1-3-n grams and character 2-6-ngrams. They used Ridge Regression as a classifier and achieved an accuracy of 89.6%.

Karaman et al. [29] used a total of 1295 news articles of 10 different authors to predict unknown authors of an articles by using TF-IDF technique and Random Forest, Decision Tree, Naive Bayes and SVM algorithms. They achieved success rates 80%, 69%, 94% and 97% of F-measure, respectively.

It can be noticed from the above examples that the author identification studies in Turkish language are open to development. In this paper, we introduced an author identification method using some classical machine learning and ensemble learning techniques. We also investigated the classification performance of the selected techniques on three different sized author groups and two distinct n-gram profiles.

III. Materials and Methods

A. Dataset

In this study, the dataset was gathered by us from plenty of Turkish News Websites. The collected articles include independent topics written by the authors and it includes articles written from 2005 to the present. Dataset contains 86,852 articles, 49 authors and an average of 1772 texts per author. While the maximum number of articles written by an author is 3,495, the minimum number of articles is 106. The information about the top 20 authors is given in Table 2. One of our authors has a total of 2,391 articles and the average number of words she/he used in these articles is 849.54. Additionally, we also see the minimum average word count per article is 335.11. While the maximum total word count for an author is 2,031,274, the average word count per article is approximately 445. No pre-processing or filtering was applied to the original corpus. Some of the texts have large spaces, url links, special characters and full capital words.

Since there is a great difference in the number of articles collected between the authors, 10-15-20 authors with the most articles were studied. The statistics about the generated three datasets are given in Table 1. The largest dataset with 20 authors, contains 55,108 articles and 24,218,293 total words.

Table 1. Article, word and average word

Dataset	Total Article	Total Word Count
Dataset-10	30,372	12,770,000
Dataset-15	43,451	17,861,840
Dataset-20	55,108	24,218,293

B. Pre-Processing

The original data cannot be sent to a machine learning model without pre-processing. Because real-world data is often noisy, inconsistent and incomplete. If we sent the data to the machine learning algorithm without pre-processing, we may encounter undesirable low scores. So, the below mentioned steps were applied to clean up the texts:

- Removing url links
- Normalizing text to lowercase
- Filtering stop-words

Table 2. Article, word and average word count per article of the top 20 authors

Author	Article Count	Total Word Count	Average Word Count per Article
Author 1	3,494	1,349,874	386,34
Author 2	3,386	1,161,662	343,07
Author 3	3,369	1,093,380	324,54
Author 4	3,084	1,562,320	506,58
Author 5	2,966	1,101,166	371,26
Author 6	2,933	1,274,944	434,68
Author 7	2,864	1,733,088	605,12
Author 8	2,777	1,273,953	458,75
Author 9	2,754	951,911	345,64
Author 10	2,745	1,267,702	461,82
Author 11	2,742	953,534	347,75
Author 12	2,670	1,017,759	381,18
Author 13	2,663	960,657	360,74
Author 14	2,553	1,338,523	524,29
Author 15	2,451	821,367	335,11
Author 16	2,391	2,031,274	849,54
Author 17	2,377	1,086,461	457,07
Author 18	2,362	989,805	419,05
Author 19	2,281	974,750	427,33
Author 20	2,246	1,274,163	567,30

- Filtering special characters
- Removing digits
- White space formatting

After completing these steps, we used the Zemberek library [24]. Zemberek is a natural language processing library that can be used for open-source Turkish languages, developed by using the Java programming language. By using this library, we did lemmatization for words, corrected spelling mistakes, and checked whether a word is Turkish. We compared the Zemberek and NLTK TurkishStemmer, and prefered using the Zemberek library.

C. Tf-Idf Vectorizer

In this paper, The TF-IDF vectorizer has been used to explore similarity between text documents. This is a very common algorithm for converting text to a meaningful representation of numbers, also known as vector representation to feed into machine learning algorithms. It is easy to implement and works fast. TF-IDF was used in the early 1970s to solve an information retrieval problem and then has been successfully used in document classification, topic modelling etc. TF represents the number of times each word occurs in text, article or any kind of datasets. For example, if the word “save” occurs 20 times in a text and the entire text has 1000 words, the TF value is 0,02 (20/1000). IDF shows the importance of the word “save” for a text. It is obtained by dividing the total number of texts by the number of texts containing the term. A score closer to zero indicates that the word is used more often.

TF-IDF Vectorizer in scikit-learn library takes ngram_range (1,1) as default parameter. The parameter (1,1) means only unigrams, (2,2) means only bigrams are used to create TF-IDF vectors. The larger n value indicates a larger probability pool. As the number of n_grams increase, we will see the decrease in the accuracy score in the experimental studies section.

D. Classification Models

In this study, machine learning and ensemble learning algorithms that have been applied to different fields and have shown successful performances were preferred and performance comparisons were made by applying them on three different data sets. The techniques used are briefly described below.

1. SVM (Support Vector Machine)

SVM is one of the supervised learning methods generally used in classification problems. Basically, it tries to separate two classes with a line or plane. It also makes this separation according to the elements at the boundary.

2. Gaussian NB (Naïve Bayes)

Naive Bayes is a probabilistic machine learning model used for classification problems. It can do good work with small data. Naive Bayes assumes that each class follows a Gaussian distribution and NB is a generative model.

3. MLP (Multi Layer Perceptron)

MLP has emerged as a result of the studies done to solve the XOR Problem. It works effectively especially in classification problems.

4. LR (Linear Regression)

LR is a popular, uncomplicated and a supervised learning algorithm. It is the simplest form of regression that is also used to examine the mathematical relationship between variables.

5. SGD (Stochastic Gradient Descent)

SGD is a linear classifier such as SVM or Linear Regression and has been successfully applied to large scale machine learning problems frequently encountered in text classification and natural language processing. It is easy to implement and has many possibilities in code tuning

6. Extra Trees

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is also easy to use and set various hyperparameters. It gives better

Table 3. Author Identification accuracies of the classifiers

N-gram	Classifier	Dataset-10	Dataset-15	Dataset-20
Unigrams	XgBoost	0.968	0.956	0.946
	SGD	**0.975**	**0.976**	**0.971**
	LR	0.962	0.956	0.954
	SVC-1	0.873	0.840	0.815
	SVC-2	0.945	0.932	0.926
	MLP	0.915	0.846	0.710
	Gaussian NB	0.832	0.775	0.746
	Extra Trees	0.936	0.904	0.871
Bigrams	XgBoost	0.831	0.855	0.822
	SGD	0.932	0.918	**0.900**
	LR	**0.935**	**0.920**	**0.900**
	SVC-1	0.812	0.707	0.633
	SVC-2	0.913	0.886	0.863
	MLP	0.896	0.765	0.607
	Gaussian NB	0.793	0.712	0.668
	Extra Trees	0.875	0.820	0.768
Average		0.900	0.861	0.819

performance than the Random Forest (RF) algorithm.

7. XgBoost (eXtreme Gradient Boosting)

XgBoost is a high-performance version of the Gradient Boosting algorithm optimized with various arrangements. The most important features of the algorithm are that it can achieve high predictive power, prevent overfitting, manage empty data and do them quickly. It is one of the popular algorithms preferred recently.

IV. Experimental Studies

We tested several combinations of hyperparameters' values by using Exhaustive Grid Search technique supplied by scikit-learn library and selected combinations with maximal classification accuracy. Final classification experiments were performed using selected hyperparameters for each classifier as given in Table 4. We used XgBoost and GaussianNB algorithms with default hyperparameters. We got the best results at default parameters even though we tried a wide variety of parameters. Two different settings of the SVC classifier are employed to see performance comparison of different kernel functions.

Table 4. HyperParameters of Classifier Methods

Classifier	HyperParameters
XgBoost	Default
SGD	Alpha=1e-05, max_iter=50, penalty=elasticnet
LR	Max_iter=1000, solver=lbfgs
SVC-1	Kernel=linear, gamma=scale, c=0.025
SVC-2	Kernel=rbf,gamma=2,c=1
MLP	Alpha=1,max_iter=1000
Gaussian NB	Default
Extra Trees	n_estimators=100,max_depth=1000, min_samples_split=2

Author identification performance of the classifiers are shown in Table 3. Accuracy was used as the evaluation metric to measure authorship identification performance. When the performances of the classifiers are evaluated, it is seen that the most successful method for unigrams is SGD. This is followed by the XgBoost and LR methods with almost similar performance. SVC-2 (with rbf kernel) and Extra Trees classifier were also successful by showing over 90% performance. Considering the bigrams, LR and SGD performed better than other classifiers.

The accuracy score is going down with the increasing author count in experiments. Because it is getting harder to predict the right author in a larger pool of authors. Increase in the number of authors also increases the number of tags to be predicted and causes a performance decrease of approximately 4%.

We worked with 1-grams and 2-grams separately for comparison purposes. As the results show us, 1-grams produce better accuracy scores than 2-grams considering all the classifiers. There is a significant difference in accuracy scores of XgBoost and SVC-1 (with linear kernel) classifiers (approximately %12) for different n-gram settings. On average, classifiers performed 5%, 8%, 10% better in unigrams compared to bigrams for Dataset-10, Dataset-15, and Dataset-20, respectively.

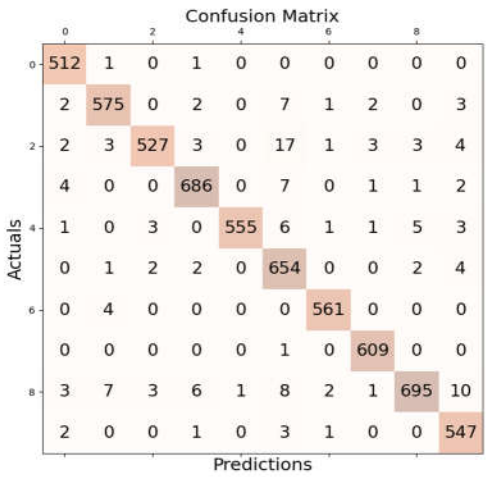

Fig. 1. Confusion Matrix of the SGD classifier by using unigrams for Dataset-10

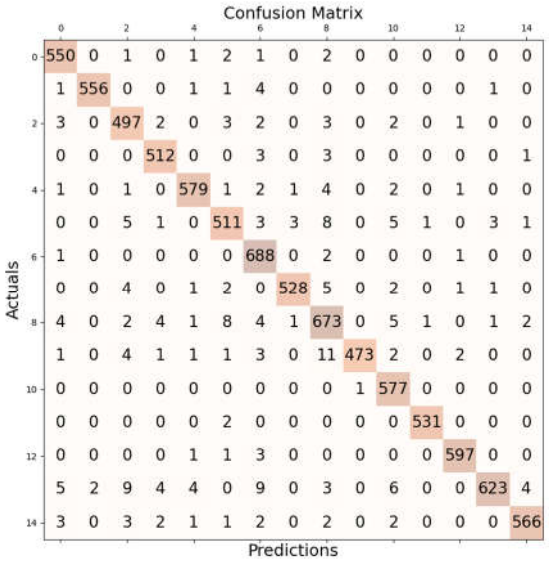

Fig. 2. Confusion Matrix of the SGD classifier by using unigrams for Dataset-15

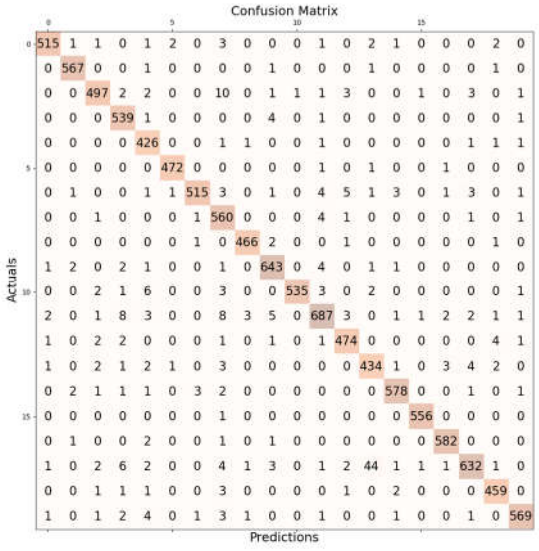

Fig. 3. Confusion Matrix of the SGD classifier by using unigrams for Dataset-20

As shown as in Figure 1, Figure 2 and Figure 3. SGD classifier has a state-of-the-art performance by using unigrams.

V. Conclusion

In this paper, we experimented with the author identification task for Turkish articles. A large amount of news articles has been collected and various text cleaning and pre-processing operations have been applied on it. Three different sizes of the author datasets have been created and unigram and bigram features have been investigated on these datasets. We also used TF-IDF to expose the author's specific ngram features. We set up the maximum features parameter of TFidfVectorizer as 3000 to build a vocabulary that only considers the top 3000 features ordered by term frequency across the particular datasets. Some important machine learning and ensemble learning algorithms that have been applied to different fields and have shown successful performances were trained on the author datasets. The GridSearchCV exhaustive search technique has been employed in the hyperparameter selection process. For all combination of values in the specified range, the network is trained and selected with the best hyperparameter for best accuracy rate.

It has been found that the most successful method for unigrams is SGD according to performance evaluation metrics. It is followed by the XgBoost and LR methods with almost similar performance. LR and SGD performed better than other classifiers in terms of bigrams. Another fact observed is that the prediction accuracy of classifiers is going down approximately 4% with the increasing author count of successive datasets. In addition to that, unigrams produce better accuracy scores than bigrams considering all the classifiers.

As a future work, a number of studies are planned with hybrid models based on BERT and deep neural networks to achieve more efficient models. Additionally, we are going to set up a different model instead of the TF-IDF model, such as Word2Vec word embedding method.

References

[1] Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology* 60.3 (2009): 538-556.

[2] Alhuqail, Noura Khalid, Author Identification Based on NLP (April 6, 2021). European Journal of Computer Science and Information Technology, Vol.9, No.1, pp.1-26, 2021, Available at SSRN: https://ssrn.com/abstract=3820262

[3] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. "BertAA : BERT fine-tuning for Authorship Attribution." *In Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

[4] A. M. Mohsen, N. M. El-Makky and N. Ghanem, "Author Identification Using Deep Learning," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 898-903, doi: 10.1109/ICMLA.2016.0161.

[5] Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. *In Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[6] Barlas, G., Stamatatos, E. (2020). Cross-Domain Authorship Attribution Using Pre-trained Language Models. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, vol 583. Springer, Cham. https://doi.org/10.1007/978-3-030-49161-1_22

[7] Ramezani, Reza. "A language-independent authorship attribution approach for author identification of text documents." *Expert Systems with Applications* 180 (2021): 115139.

[8] Olga Fourkioti, Symeon Symeonidis, Avi Arampatzis, Language models and fusion for authorship attribution, Information Processing & Management, Volume 56, Issue 6, 2019, 102061, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.102061.

[9] S. Okuno, H. Asai and H. Yamana, "A challenge of authorship identification for ten-thousand-scale microblog users," 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 52-54, doi: 10.1109/BigData.2014.7004491.

[10] Z. Damiran and K. Altangerel, "Author Identification-An Experiment Based on Mongolian Literature Using Decision Trees." *2014 7th International Conference on Ubi-Media Computing and Workshops*. IEEE, 2014. pp. 186-189.

[11] Ramezani, Reza, Navid Sheydaei, and Mohsen Kahani. "Evaluating the effects of textual features on authorship attribution accuracy." *ICCKE 2013*. IEEE, 2013.

[12] H. Paci, E. Kajo, E. Trandafili, I. Tafa and D. Salillari, "Author Identification in Albanian Language," *2011 14th International Conference on Network-Based Information Systems*, pp. 425-430.

[13] Pandian, A., V. V. Ramalingam, and R. V. Preet. "Authorship identification for Tamil classical poem (Mukkoodar Pallu) using C4. 5 algorithm." *Indian Journal of Science and Technology* 9.46 (2016).

[14] Kale Sunil Digamberrao, Rajesh S. Prasad, Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi, Procedia Computer Science, Volume 132, 2018, Pages 1086-1101, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.024.

[15] Oliveira W Jr, Justino E, Oliveira LS. Comparing compression models for authorship attribution. Forensic Sci Int. 2013 May 10;228(1-3):100-4. doi: 10.1016/j.forsciint.2013.02.025. Epub 2013 Mar 24. PMID: 23597746.

[16] Romanov, Aleksandr & Kurtukova, Anna & Shelupanov, Alexander & Fedotova, Anastasia & Goncharov, Valery. (2020). Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks. Future Internet. 13. 3. 10.3390/fi13010003.

[17] Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. *Future Internet* 2022, 14, 4. https://doi.org/10.3390/fi14010004

[18] Sage, M., Cruciata, P., Abdo, R., Cheung, J.C., & Zhao, Y.F. (2020). Investigating the Influence of Selected Linguistic Features on Authorship Attribution using German News Articles. *SwissText/KONVENS.*

[19] Otoom, Ahmed & Abdallah, Emad & Jaafer, Shifaa & Hamdallh, Aseel & Amer, Dana. (2014). Towards author identification of Arabic text articles. 2014 5th International Conference on Information and Communication Systems, ICICS 2014. 1-4. 10.1109/IACS.2014.6841971.

[20] O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec*. 30, 4 (December 2001), 55–64. https://doi.org/10.1145/604264.604272

[21] B. Kırmacı and H. Oğul, "Evaluating text features for lyrics-based songwriter prediction," 2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES), 2015, pp. 405-409, doi: 10.1109/INES.2015.7329743.

[22] Upul Bandara, Gamini Wijayarathna, Source code author identification with unsupervised feature learning, Pattern Recognition Letters, Volume 34, Issue 3, 2013, Pages 330-334, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2012.10.027.

[23] Alonso-Fernandez, Fernando & Belvisi, Nicole & Hernandez-Diaz, Kevin & Muhammad, Naveed & Bigun, Josef. (2021). Writer Identification Using Microblogging Texts for Social Media Forensics. IEEE Transactions on Biometrics, Behavior, and Identity Science. PP. 1-1. 10.1109/TBIOM.2021.3078073.

[24] Akın, Ahmet Afsin, and Mehmet Dündar Akın. "Zemberek, an open source NLP framework for Turkic languages." *Structure* 10.2007 (2007): 1-5.

[25] M. S. Atar, E. Esen and M. A. Arabaci, "Supervised author recognition with aggregated word embeddings," 2018 26th Signal Processing and Communications Applications Conference (SIU), 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404464.

[26] Diri, B., and Amasyalı, M. F. (2003, June). "Automatic author detection for Turkish texts." In Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP) (pp. 138-141).

[27] Örücü F., Dalkılıç G., "Author Identification Using N-grams and SVM", The 1. International Symposium on Computing in Science & Engineering, ISBN:978-605-61394-0-6 P:130, Kuşadası, 3-5 Haziran 2010

[28] B. Kuyumcu, B. Buluz and Y. Kömeçoğlu, "Author Identification in Turkish Documents with Ridge Regression Analysis," 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, pp. 1-4, doi: 10.1109/SIU.2019.8806242.

[29] Burcu İlkay KARAMAN, Feriştah DALKILIÇ, Emine Eda ÇAM EKER, "Author Recognition In Modern Turkish For Forensic Linguistic Cases Using Machine Learning", 1st International, 17th National Forensic Science Congress, 12-15 November 2020, Online.