Author Identification with Chicken Swarm Optimization Algorithm and Adaboost Approaches

Merve Güllü
Computer Engineering.
Gazi University, Faculty of Technology
Ankara, Turkey
mervegullu@gazi.edu.tr

Hüseyin Polat
Computer Engineering.
Gazi University, Faculty of Technology
Ankara, Turkey
polath@gazi.edu.tr

Aydın Çetin
Computer Engineering.
Gazi University, Faculty of Technology
Ankara, Turkey
acetin@gazi.edu.tr

***Abstract*—Author identification takes an important place in Natural Language Processing (NLP). Each written document carries the trail of its author. In this study, we aim to realize the author identification via the traces belonging to author be retrieved from the text. A raw dataset was created with 25 columnists and randomly selected 2024 texts from different newspapers in the Turkish language. A dataset with character and lexical features with natural language processing methods were prepared over the raw dataset. The feature selection process was realized with the combination of the Chicken swarm optimization and the ensemble learning algorithms on the prepared dataset. The results were evaluated before and after the feature selection method was applied. The highest success rate with 93.99% was achieved when Adaboost with J48 algorithm was applied after the feature selection process carried out.**

Keywords—authorship identification, chicken swarm optimization, AdaBoost, feature selection, boosting

I. INTRODUCTION

Author identification can be defined as the process of determining the author of a text. In author identification processes, the language of the author and the subject of the text is also thought to affect the success of the models. Writing modeling style independent of the subject of the text is a difficult problem [4]. To identify the author of the text, a proper classification algorithm along with the properly selected attributes should be used. For this purpose, the feature sets created over the text; Using n-gram [1-3] use of punctuation marks [1], character-based attributes [1,2,4,5], misuse [1,2,5], lexical features [2,6] word-based features like stop words [2], the frequency of use of characters and spokes varies as [2,3,6,7].

Studies have been carried out for different languages, using different feature selection methods in different languages. The literature has been studied in many languages such as Arabic [5], German [6], Hebrew [7], Chinese [8], Greek [9], Danish [10], Italian [11]. There are a limited number of studies [14 -16] reported in the literature related to author identification in texts written in Turkish Language. In [14], 20 authors and 35 style markers were used to identify 20 authors in their texts. 22 style markers were selected using the CFS Subset Evaluator with the sequence search method. With this selection process, the highest success rate of the model trained with classical machine learning algorithms has been increased from 70.75% to 80%. In [15], 17 authors in total and 850 texts belonging to them were used. The study was carried out on two separate data sets. Dataset 1 contains 10 authors, Dataset 2 includes 7 authors and their own text. The success rate was 99.7% with the use of Chi-square feature selection method as machine learning algorithms and feature selection algorithm. The other study [16] on Turkish, a total of 18 authors and 630 texts were used. In the study, 3 different data sets were used. The number of authors to be classified in the datasets they created is given as 18, 9, and 9 respectively. The average classification success of datasets was 72.4%, 80.0% and 82.9%, respectively. In the study, which also used feature selection algorithms, with dataset 3 that 9 authors and containing their texts was achieved 96.9% as the highest success rate. In their results, they also emphasized that the number of authors was directly influenced by the classification success.

Classification success in author identification is directly related to the attributes. A high number of attributes when combined with large datasets increases the cost of classification algorithms. The ultimate aim of these techniques is to increase the classification success while reducing the classification cost. To overcome this drawback, the application of feature selection approaches is commonly preferred. Feature selection is the process of creating a smaller and more informative subset over the original dataset features [13]. Feature selection aims to find the smallest subset of features that are required and sufficient to estimate the class tag. However, feature selection methods alone may not be sufficient enough since it may result in several global solutions. To find the best global solution for feature selection methods, swarm optimization algorithms have been used [16,17] recently.

In this study, author identification with Chicken Swarm Optimization (CSO) Algorithm and Adaboost approaches is proposed. In the study, the data set consisting of 25 authors and their 2024 texts were created. The study analyzes general lexical and character traits, regardless of the subject content of the texts, and can be used in this way for text written on any subject.

The rest of the study is edited as follows: In the second part, the dataset, feature selection method, and ensemble method are described. The results of the study are in Section 3 and the conclusion is presented in Section 4.

II. MATERIAL AND METHOD

The feature selection process used in this study is based on three main functions, namely creation, evaluation, and selection of subsets as shown in Fig. 1

978-1-7281-7565-2/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on April 22,2022 at 21:14:15 UTC from IEEE Xplore. Restrictions apply.

The first function involves creating subsets that differ from the original dataset. The chicken swarm optimization (CSO) algorithm was used in the process of creating a subset. Then subsets that are formed are subjected to the evaluation process. In the subset evaluation process, it is based on the process of assigning the suitability value of the subsets and including them in the solution class according to the suitability value. Ensemble learning algorithms (ELA) were used to calculate the suitability value. Subsets formed are given as input to ELA. Resulted accuracy values obtained in classification were assigned as the suitability value of each subset. The assigned values were analyzed and additions to the solution class were performed.

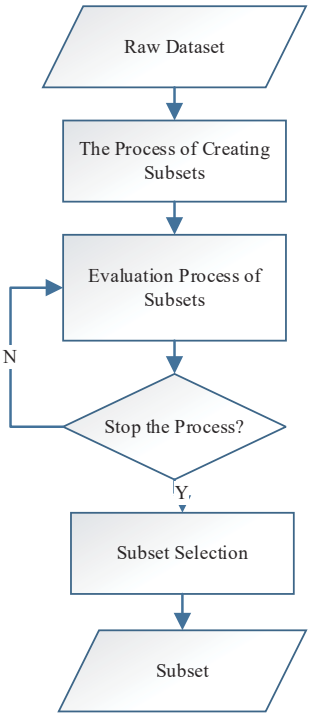


Fig. 1. The flowchart of the main process

The stop criterion is checked after subset evaluation. These two main functions are repeated until the stopping criteria are met. When no criteria are met, subsetting and evaluation processes are repeated The stopping criterion may differ depending on the design as reaching the pre-determined suitability value or the number of cycles. When the stop criterion is met, the subset selection process, the last main function, takes place. The highest suitability value of all subsets formed in the function is considered as the solution. In this study, the stop criterion was determined as the number of generations.

The last main function, the subset selection, covers the subset selection process with the highest suitability value in the solution class. The output is the subset of the final solution and its suitability value.

A. DataSet

In this study, 2024 column articles are belonging to 25 authors taken from different websites as raw data. The features of natural language processing techniques have been extracted on the columns. Table I. shows the content of the dataset used in this study.

In the Dataset distribution, the sample numbers are close to each other. At least 48 texts are belonging to a writer. In general, the dataset contains 80 texts for each author.

There is no proven feature set for author identification. In studies conducted in this field, a certain set of qualities was not used and generally lexical, syntactic or structural features are examined on the texts. It is thought that the features used in this study will yield successful results in the author's identification.

TABLE I. CONTENT OF DATASET

Name	*Feature explanation*	*Type*	*Min-Max*
tw	Total No of words in the text	int	10-1187
V1	No of words that appear only once in the text	Int	1-947
V2	No of words appearing only twice in the text	Int	0-134
ratioV2 toV1	The ratio of words specified once in the text to the No of words passed twice.	Float	0-35.0
vt	No of vowels used in the text	Int	10-4774
qmt	No of question marks used in the text	Int	0-45
emt	No of exclamation marks used in the text	Int	0-29
pst	The parenthesis sign used in the text '(' and ')'	Int	0-18
dt	No of dashes used in the text	Int	0-28
st	No of semicolons used in the text	Int	0-26
dst	No of dots in the text	Int	0-573
ct	No of comma in the text	Int	0-159
sqt	No of single quotes in the text	Int	0-4
dqt	No of double quotes in the text	Int	0-80
nst	No of Sentence in the text	Int	1-160
lt	No of letters in the text	Int	25-11181
nct	No of characters in the text	Int	36-13430
alw	Avg. of the no of letters used in the word	Float	6,01-10,4
aws	Avg. no of words in a sentence	Float	1,5-51,6
ant	Avg. word distance of nouns in the text	Float	0-29
avt	Avg. word distance of verbs in text	Float	1-12.4
aat	Avg. word distance of adjectives in the text	Float	0-23,5
apt	Avg. word distance of pronouns in the text	Float	0-25
abt	Avg. word distance of abbreviations in text	Float	0-250
aut	Avg. word distance of words of unknown type in the text	Float	1-33.2
acnt	Avg. word distance of certain nouns in the text	Float	0-33,2
an	Avg. use of no in the text	Float	0-88

B. Subset Creation Process

Different size datasets have been prepared to increase success on the dataset. The size change was made only on attributes. CSO algorithm is used for subset creation.
The CSO algorithm is an optimization method based on the hierarchical order within the chicken swarm and the food-seeking behavior that simulates the chicken swarm behavior. CSO has been used in many different research areas because of its convergence speed and accuracy that include robot path planning[17], solving maximum sidelobe suppression

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on April 22,2022 at 21:14:15 UTC from IEEE Xplore. Restrictions apply.

optimization problems of antenna arrays [18], unlocking migration problem for virtual machine consolidation [19], parameter estimation of nonlinear systems in the industry [20], and increasing energy efficiency in wireless sensor networks [21].

The main rules of the algorithm;

(1) There may be more than one subgroup in the swarm, and each subgroup contains one rooster and consists of chickens and chicks, the number of which can be determined parametrically.

(2) In the swarm, individuals with the best condition are appointed as roosters and act as leaders in the assigned group. Individuals with the worst condition value are appointed as chicks, the rest of the individuals are designated as chickens. The group where the chickens will live is chosen randomly. The relationship between mother hens and chicks is also randomly established.

(3) The hierarchical order is updated after the number of generations determined parametrically.

(4) Chickens in each group follow the rooster, their group leader, to look for food. At times, stealing between chickens may occur. It is assumed that the chicks can steal the food found by other individuals randomly and that each chick follows his mother to look for food.

The general structure mentioned above has been used and it has been made with the Adaboost method, which is the condition value calculation part is ELA.

C. Subset Evaluation Process

Subsets created with CSO are given as input to the Boosting method, which is an ELA. Boosting method is a developed community learning method to minimize errors that occur in classical machine learning algorithms [22]. In this study, Adaptive Boosting (Adaboost) algorithm [23], which has been transformed into an effective boosting method thanks to the improvements made in recent years, has been used. Adaboost[24], Adaboost aims to organize the erroneous classification in the model it creates with its training data, with other models. It continues until there is no possibility to improve in the model[25]. The main purpose of the method is to produce models with different weights from the same training data and to perform the test results according to the weighted majority vote [26].

The working principle of Adaboost algorithm is given in Fig. 2. Initially, it creates a random size training set from the training data set. It creates a model with the selected machine learning technique and training set. For the correct prediction of the final model to be created, it makes the selection of the training set at each step. During the selection process, more weight is assigned to previously misclassified sample data. The operations continue until the number n determined in this way or until all of the training data are classified correctly.

Subsets created with CSO were evaluated by Adaboost algorithm. In this process, J48 algorithm was chosen as the classifier within the Adaboost algorithm. While selecting the algorithm, the classification success made before the feature reduction is taken into consideration (Table II).

The evaluation was based on Accuracy *Acc*, Eq. (1), which is frequently used in the literature in evaluating machine learning algorithms.

$$Acc = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

where *TP* is true positive, *TN* is true negative, *FP* is false positive and *FN* is false negative.

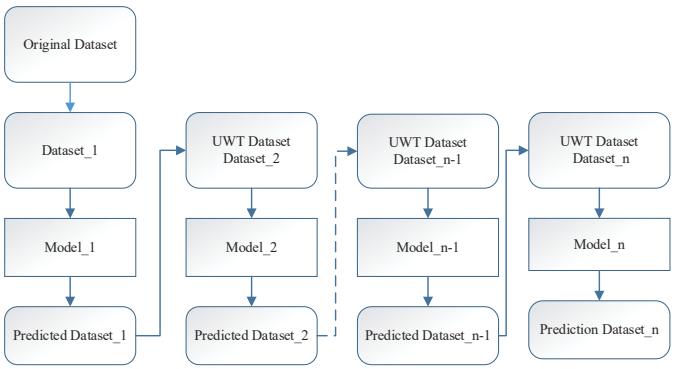


Fig. 2. General working principle representation of Adaboost algorithm (UWT: Updated Weighted Training)

D. Subset Selection Process

In this study, the feature selection process was realized by the combination of CSO and Adaboost method. In the CSO algorithm, subsetting and evaluation operations have been realized as much as G value determined parametrically. G value was determined as 32 in the study.

According to the termination criteria of the optimization algorithm, the subset with the Acc value closest to 100% classification success was chosen as the result dataset.

In addition to the Adaboost(J48) algorithm results, the selected dataset was also tested in the Adaboost algorithm with Naive Bayes, SVM, KNN, Decision Table algorithms.

III. Result of Study

In the evaluation of the result of this study, the metrics generally used in the literature to evaluate the results of machine learning algorithms are used.

In addition to accuracy calculation, sensitivity *Sen*, Eq. (2) and specificity *Spe*, Eq. (3) metrics were used.

$$Sen = TP / (TP+FN) \quad (2)$$

$$Spe = TN / (FP+TN) \quad (3)$$

The main dataset prepared in the study (Table I) was subjected to training and testing with the Adaboost algorithm and 5 different classifier methods. The results are given in Table II. The evaluation process was carried out with the Adaboost with J48 algorithm since it has achieved the highest classification success as seen in Table II.

TABLE II. Success Metrics and Values Before Feature Selection Operation

Adaboost model algorithms	*Acc (avg.)*	*Sen (avg.)*	*Spe (avg.)*
w/Naïve Bayes	72.818%	73.9%	72.8%
w/SVM	79.79%	82.4%	79.8%
w/KNN	82.41%	82.6%	82.4%
w/Decision Table	76.16%	88.2%	76.2%
w/J48	89.82%	90.5%	89.8%

Once the new dataset obtained after the feature selection process, the same algorithms were trained and tested. The

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on April 22,2022 at 21:14:15 UTC from IEEE Xplore. Restrictions apply.

results are given in Table III. With the use of CSO in this study, the aim is to exceed the highest success value made on the main dataset. When the results are evaluated, the effect of the feature selection process on success is observed. Although subset evaluation is performed only by one algorithm Adaboost with J48, in general, an increase in success evaluation criteria is observed in all algorithms except Adaboost with Decision Table. The highest level of improvement rates was obtained in Adaboost with Naïve Bayes with 8.4% increase in accuracy Acc with CSO optimization. However, it was less than the Acc value of Adaboost with J48 even before feature selection was applied.

TABLE III. SUCCESS METRICS AND VALUES AFTER FEATURE SELECTION OPERATION

Adaboost model algorithms	*Acc (avg.)*	*Sen (avg.)*	*Spe (avg.)*
w/Naïve Bayes	81.22%	81.22%	81.2%
w/SVM	80.13%	79.9%	80.1%
w/KNN	84.68%	84.3%	84.7%
w/Decision Table	73.56%	74.9%	73.6%
w/ J48	93.99%	93.9%	93.7%

The results showed that the dataset created using natural language processing techniques and the use of CSO and Adaboost algorithm can be used in author identification studies.

IV. CONCLUSION

In this study, a dataset was created using natural language processing techniques on Turkish columns compiled from 25 different authors. The feature selection process has been realized with the method created by the combination of the dataset CSO and Adaboost algorithms. Success values before and after the feature selection process were compared. It was observed that the feature selection method used in the study increased success in Adaboost algorithms with Naïve Bayes, SVM, KNN, and J48.

Increased success with the combination of the chicken swarm optimization algorithm and community learning algorithm has been found to have a strong potential in predicting the identity of an author along with the feature selection process.

We think that the diversity of authors as well as the diversity of the languages are also important in the measurement of the success rate of the classification algorithms. The authors can be identified in different languages with the same techniques to monitor whether the language affects the models. Future studies will be a focus on larger datasets with an increased number of authors and added languages.

REFERENCES

[1] P. Canbay, H. Sever, and E. Akcapinar Sezer, "Determining of discriminative blog size for authorship attribution on the Turkish texts," *6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding*, vol. 2018-Janua, no. May, pp. 1–5, 2018, doi: 10.1109/ISDFS.2018.8355373.

[2] O. Tsur and A. Rappoport, "Using classifier features for studying the effect of native language on the choice of written second language words," no. June, pp. 9–16, 2007, doi: 10.3115/1642025.1642027.

[3] M. Koppel, J. Schler, and K. Zigdon, "Determining an author's native language by mining a text for errors," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. May 2014, pp. 624–628, 2005, doi: 10.1145/1081870.1081947.

[4] S. Bischoff *et al.*, "The Importance of Suppressing Domain Style in Authorship Analysis," no. 1, 2020, [Online]. Available: http://arxiv.org/abs/2005.14714.

[5] H. Sayoud, "Author discrimination between the holy Quran and Prophet's statements," *Lit. Linguist. Comput.*, vol. 27, no. 4, pp. 427–444, 2012, doi: 10.1093/llc/fqs014.

[6] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Appl. Intell.*, vol. 19, no. 1–2, pp. 109–123, 2003, doi: 10.1023/A:1023824908771.

[7] M. Koppel, N. Akiva, and D. Mughaz, "New methods for attribution of rabbinic literature," *Hebr. Linguist. A J. Hebr. Descr. Comput. Appl. Linguist.*, vol. 57, no. Sebastiani 2002, pp. 5–18, 2006.

[8] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, Feb. 2006, doi: 10.1002/asi.20316.

[9] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-Gram-Based Author Profiles for Authorship Attribution," *Pacific Assoc. Comput. Linguist.*, pp. 255–264, 2003.

[10] P. Juola, "A Controlled-corpus Experiment in Authorship Identification by Cross-entropy," *Lit. Linguist. Comput.*, vol. 20, no. Suppl 1, pp. 59–67, Jan. 2005, doi: 10.1093/llc/fqi024.

[11] J. Savoy, "Comparative evaluation of term selection functions for authorship attribution," *Digit. Scholarsh. Humanit.*, vol. 30, no. 2, pp. 246–261, Jun. 2015, doi: 10.1093/llc/fqt047.

[12] V. Kothari, J. Anuradha, S. Shah, and P. Mittal, "A survey on particle swarm optimization in feature selection," *Commun. Comput. Inf. Sci.*, vol. 270 CCIS, no. PART II, pp. 192–201, 2012, doi: 10.1007/978-3-642-29216-3_22.

[13] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol. Comput.*, vol. 54, no. February, p. 100663, 2020, doi: 10.1016/j.swevo.2020.100663.

[14] T. Taş and A. K. Görür, "Author Identification for Turkish Texts," *Çankaya Üniversitesi Fen-Edebiyat Fakültesi, J. Arts Sci.*, vol. 7, pp. 151–161, 2007, [Online]. Available: http://www.arastirmax.com/system/files/dergiler/64470/makaleler/7/1/arastirmax_11081_pp_151-161.pdf.

[15] Y. Bay and E. Çelebi, "Feature Selection for Enhanced Author Identification of Turkish Text," 2016, pp. 371–379.

[16] F. Türkoğlu, B. Diri, and M. F. Amasyalı, "Author Attribution of Turkish Texts by Feature Mining," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1086–1093.

[17] X. Liang, D. Kou, and L. Wen, "An Improved Chicken Swarm Optimization Algorithm and its Application in Robot Path Planning," *IEEE Access*, vol. 8, pp. 49543–49550, 2020, doi: 10.1109/ACCESS.2020.2974498.

[18] S. Liang, Z. Fang, G. Sun, Y. Liu, G. Qu, and Y. Zhang, "Sidelobe Reductions of Antenna Arrays via an Improved Chicken Swarm Optimization Approach," *IEEE Access*, vol. 8, pp. 37664–37683, 2020, doi: 10.1109/ACCESS.2020.2976127.

[19] F. Tian, R. Zhang, J. Lewandowski, K.-M. Chao, L. Li, and B. Dong, "Deadlock-free migration for virtual machine consolidation using Chicken Swarm Optimization algorithm," *J. Intell. Fuzzy Syst.*, vol. 32, no. 2, pp. 1389–1400, Jan. 2017, doi: 10.3233/JIFS-169136.

[20] G. Ren *et al.*, "A parameter estimation method for fractional-order nonlinear systems based on improved whale optimization algorithm," *Mod. Phys. Lett. B*, vol. 33, no. 7, 2019, doi: 10.1142/S0217984919500751.

[21] W. Osamy, A. A. El-Sawy, and A. Salim, "CSOCA: Chicken Swarm Optimization Based Clustering Algorithm for Wireless Sensor Networks," *IEEE Access*, vol. 8, pp. 60676–60688, 2020, doi: 10.1109/ACCESS.2020.2983483.

[22] Y. Freund, R. E. Schapire, and M. Hill, "Experiments with a New Boosting Algorithm Rooms f 2B-428 , 2A-424 g," 1996.

[23] J. Cui *et al.*, "AdaBoost Ensemble Correction Models for TDDFT Calculated Absorption Energies," *IEEE Access*, vol. 7, pp. 38397–

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on April 22,2022 at 21:14:15 UTC from IEEE Xplore. Restrictions apply.

38406, 2019, doi: 10.1109/ACCESS.2019.2905928.

[24] Y. Freund and R. E. Schapire, “Experiments with a New Boosting Algorithm,” *Proc. 13th Int. Conf. Mach. Learn.*, pp. 148–156, 1996, doi: 10.1.1.133.1040.

[25] F. Khan, J. Ahamed, S. Kadry, and L. K. Ramasamy, “Detecting malicious URLs using binary classification through ada boost algorithm,” *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 997–1005, 2020, doi: 10.11591/ijece.v10i1.pp997-1005.

[26] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, “On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification,” *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 728–738, 2020, doi: 10.1109/TKDE.2019.2891622.

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on April 22,2022 at 21:14:15 UTC from IEEE Xplore. Restrictions apply.