# Language-Independent Features for Authorship Attribution on Ukrainian Texts

Yuliia Hlavcheva[a], Maksym Glavchev[a], Victoria Bobicev[b], Olga Kanishcheva[a]

[a] *National Technical University "KhPI", 2 Kyrpychova str., Kharkiv, 61002, Ukraine*
[b] *Technical University of Moldova, Bd. Ştefan cel Mare, 168, Chişinău, MD-2004, Republic of Moldova*

**Abstract**

Authorship attribution is the natural language processing task of the author identification of an input text. The main goal of this task is to define the salient characteristics of documents that capture the author's writing style. In this paper, we analyze language-independent features for authorship attribution. All experiments were realized on the corpus of Ukrainian scientific papers. For the experiments we used Bayes Based Algorithms (Naive Bayes Multinomial), Support Vector Machine (SMO) and Decision Trees (LMT, J48) methods. The experimental results of the scientific text classification demonstrated that Decision Trees method in most cases outperforms other machine learning methods, and the proposed in the paper language-independent features are appropriate for the Ukrainian scientific documents authorship attribution.

**Keywords**

Writing Style, Language-Independent Features, Authorship Attribution, Text Classification, Machine Learning Methods

## 1. Introduction

The task of authorship identification is not new. The results of authorship detection studies are actively used in various spheres of human life. Authorship research can be divided into three main areas: *authorship identification* (author identification by analyzing the writing styles of other works of this author), *authorship characterization* (determination of the author's characteristics (gender, education, culture, language skills, etc.) and generation of the author's profile), *similarity detection* (comparison of several texts and definitions were created by one author, actually identifying the identity of the author) [1, 2, 3, 4]. Similarity detection is most often used to identify potential academic plagiarism. The advantage of this method is small value of input textual data.

The relevance of the research topic is confirmed by the dynamics of publications and citations in scientific databases. Data from the Web of Science Core Collection on the publication activity of scientists confirms this. We selected documents by the keyword "Writing Style" and received 6,679 documents for the previous 10 years (2010-2019). During this time, the number of publications has almost tripled, but the publication citation on the topic has significantly increased (Figure 1).

Methods and approaches to the author identification and the definition of writing style differ in the studies of various authors [4, 5, 6, 7, 8, 9,10].

In the paper [7] the authors proposed to detect differences between writings on the same topic provided by a set of users and tested whether these differences are enough to use for an authentication system. They observed 74% accuracy in detecting the actual authors and concluded that with additional features the accuracy can be pushed to above 90%. Also they analyzed the impact of some data cleaning systems like removing stop words and punctuation marks, and how they affected the final results.

*IT&I-2020 Information Technology and Interactions, December 02–03, 2020, KNU Taras Shevchenko, Kyiv, Ukraine*
EMAIL: yuliia.hlavcheva@khpi.edu.ua (A. 1); maksym.glavchev@khpi.edu.ua (A. 2); victoria.bobicev@gmail.com (A. 3); kanichshevaolga@gmail.com (A. 4)
ORCID: 0000-0001-7991-5411 (A. 1); 0000-0001-9670-9118 (A. 2); 0000-0003-4450-3964 (A. 3); 0000-0002-9035-1765 (A. 4)

In the paper [6] was proposed a new methodology for authorship attribution based on a profile of indices related to the generalized coupon collector problem, called coupon-collector-type indices.

The authors in the work [5] proposed an alternative AV approach that considers only topic-agnostic features in its classification decision. In addition, they presented a post-hoc interpretation method that allows understanding, which particular features have contributed to the prediction of the proposed AV method.

The authors [8, 10] for the Authorship verification task also allocate a subtask Author obfuscation in a situation where the author deliberately changes the writing style.

Author obfuscation is the adversarial task of preventing a successful verification by altering a text's style so that it does not resemble that of its original author anymore. The paper [8] introduces new algorithms for both tasks. They proposed an approach that models writing style difference as the Jensen-Shannon distance between the character n-gram distributions of texts, and manipulates an author's writing style in a sophisticated manner using heuristic search.

The length of text is influence on the authorship identification. In the work [9] the authors tried to identify authors of tweet messages, which are limited to 280 characters.

This paper focuses on the third direction of similarity identification. Therefore, by identification of the author, we mean the definition of a potential author of a text from a certain number of applicants. The decision is based on the group of properties that reflect the author's style measurement and comparison. By the author's style, we mean the author's own writing style, which he uses unconsciously when writing texts. Text properties that reflect the author's style are called stylometric. Stylometric properties by which the author's style can be identified make up a list of style markers.

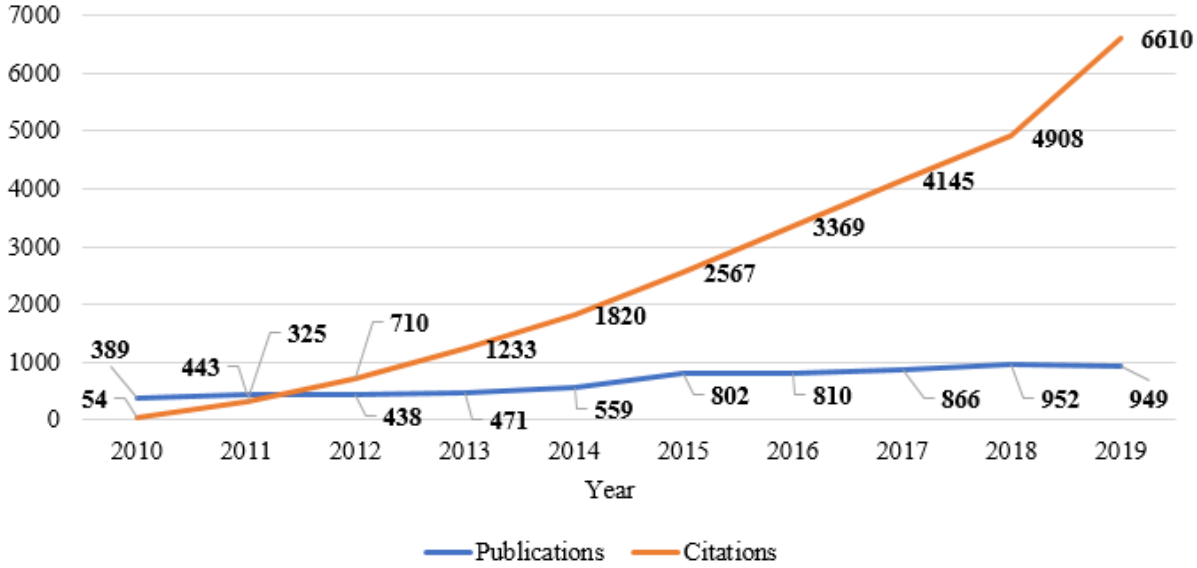


**Figure 1**: The number of publications of "Writing Style" topic for 2010-2019 years

The main stylometric properties include the following features: lexical, symbolic, syntactic, semantic, and application-specific [1, 2]. The author [1] highlights the following application-specific properties that can reflect a written style: Structural, Content-specific and Language-specific.

The search for symbolic properties is based on the analysis of text as a sequence of symbols. This type of information is easily identifiable for any natural language. We can easy analyze total number of characters, number of alphabetic characters, number of upper and lower case characters, number of punctuation marks, etc. On practical experiments, it was proved that the use of symbolic features, in combination with other markers, is useful for determining the style of writing [11].

All investigated structural units of the text depend on the language, general content, and have a probabilistic nature. Each language has certain lexical, syntactic, semantic, stylistic features. Therefore, the same approaches to attribution for different languages give results of different quality (accuracy). This is confirmed by previous studies of the stylometric properties of scientific texts in Ukrainian and Russian, carried out by the authors [12, 13].

Also, all style markers are equally effective when used for different languages [11]. Based on this, we distribute the stylometric characters into two groups: language-dependent features and language-independent features.

The paper [11] analyzes in detail and investigates language-independent and language-dependent features of the stylometric properties: average sentence length in the text, percentage of capital letters in relation to number of lowercase letters, percentage of lowercase letters in relation to total number of characters in the text, percentage of punctuation signs in relation to total number of spaces in the text, percentage of numeric characters in relation to total number of letter characters in the text, average word length in a sentence in the text, frequency of the most frequent stop word in the text, the most frequent starting word in a sentence in the text, the frequency of the most frequent starting letter of the starting word, the frequency of the most frequent starting letter of the stop word, the number of the words occurring just once in the text, the number of the words that occurred twice in the text, the number of words for the given word length in the text.

Some of them were used in experiments by the authors of this publication. This paper focuses on the research of the statistical characteristics of scientific texts in the Ukrainian language, which can be attributed to language-independent stylometric properties.

## 2. Data Description

For the experiments we used our own preprocessed text corpus. The data source is the repository of the National Technical University "Kharkiv Polytechnic Institute" (http://repository.kpi.kharkov.ua) and the portal of scientific publications of the National Technical University "Lviv Polytechnic" (http://science.lpnu.ua/uk). The text corpus consists of individual scientific publications in Ukrainian. For stylometric properties we used only the paper main text, which best reflects the author's written style. For each author, a collection of paper fragments is formed. Thus, the text corpus represents a set of data based on identical fragments (instances), all documents of which are considered individually. In this paper, we have implemented an instance-based approach to authoring style research. Two existing approaches [1, 14, 15] were used so far: profiles (profile-based approaches), instances (instance-based approaches). We concentrated on the instance-based methods, because the profile approaches defines the overall writing style of the author and does not account for style changes in individual documents. The instance-based approaches determine the writing style for each instance of the text and thus accommodate the changes in the author's writing style. In our case, stylometric properties are determined for each fragment of the paper separately. Text data statistics are shown in Table 1. However, the distribution of text volumes among the authors and the files is highly unbalanced. The smallest files contained only 150-160 words. Therefore, we cut files and are working with shortest scientific texts in our collection are around 150 words. We created two subsets for our experiments, one of them consist of 8 classes (authors), other – 32 classes (authors). These classes were selected randomly but the sets are balance.

**Table 1**
Statistics of the corpus

| Characteristics | 8 classes | 32 classes |
|---|---|---|
| Number of authors | 8 | 32 |
| Number of documents | 67 | 271 |
| Average number of documents per author | 8 | 8 |
| Number of fragments | 1019 | 2633 |
| Average number of fragments per author | 127 | 82 |
| Total size (tokens) | 154643 | 415565 |
| Average number of tokens per author | 19330 | 12986 |
| Total number of sentences | 10385 | 24743 |
| Average number of sentences per author | 1298 | 773 |

Statistical characteristics were determined using our own software "Determination of statistical characteristics of text fragments". This program helps to obtain different kind of features from text and is adapted to the processing of a wide range of character sets.

This program implements its processing of text data according to its own algorithm in order to form a wide range of text statistical parameters. The general interface consists of two areas: working with data and calculation results (Fig. 2). Calculations are carried out on the tabs for following elements: sentences, words, symbols, set (user-selectable).

The program implements the functions of calculating and saving the obtained experimental data for all elements.

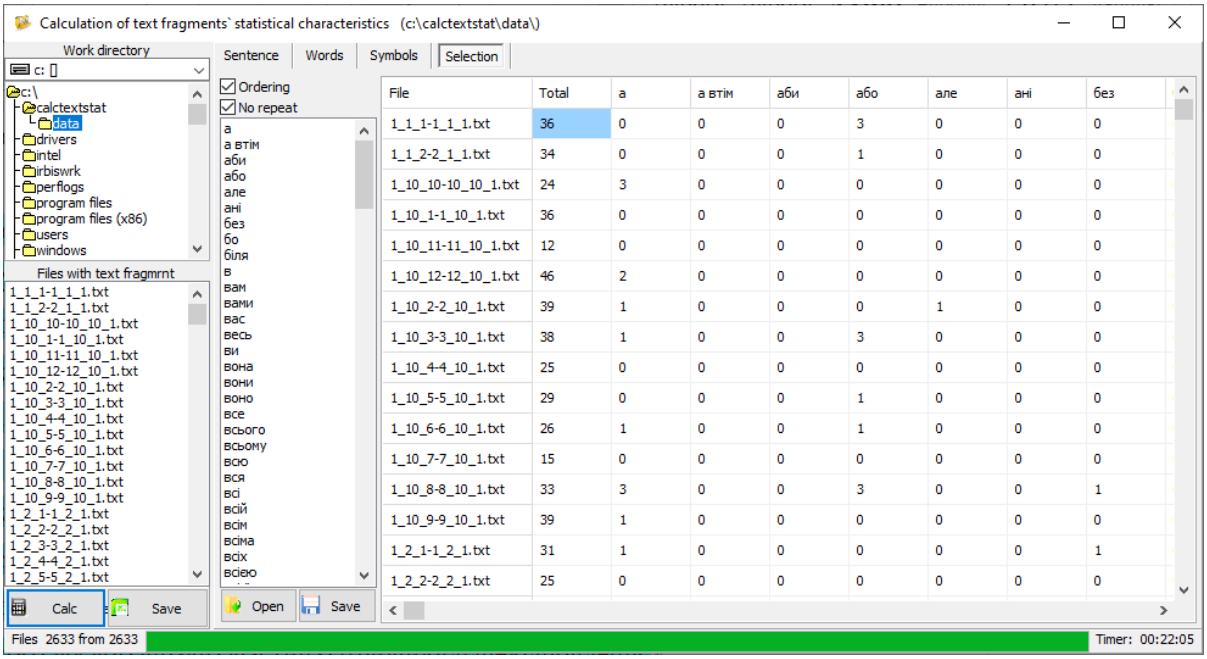


**Figure 2**: The interface of the "Determination of statistical characteristics of text fragments" software

## 3. Authorship Attribution the Base on Language-Independent Features

In this Section, we describe our language-independent features for text classification in Ukrainian and show the main results of our experiments of authorship attribution.

### 3.1. Identification of Language-Independent Features

Groups of properties, which refer to the text statistical parameters and allow to determine the author's style with high accuracy are described in [2, 11]. The authors of this paper created their own list of text statistical properties of Ukrainian, which are divided into 5 groups:

*Group 1:* average number of words in a sentence, average word length, average word frequency, punctuation (5 indicators).

*Group 2*: the number of words with length from 1 to 20 characters, the number of words with a word frequency from 1 to 8 times (28 indicators).

The range of up to 20 letters was selected due to the use of similar words: multifunctional (19), competitiveness (22), etc. The statistics on the words with a length of 1 to 20 characters and the number of words with a word frequency of 1 to 8 times are shown in Fig. 3 and Fig. 4.

*Group 3*: frequency of using letters of the Ukrainian alphabet (33 indicators).

Less commonly used letter "г"/"g" and most often used letters "о"/"o", "н"/"n" and "а"/"a". The full information about these letters are presented in Table 2.

**Table 2**

Frequency of less and most commonly used letters in our corpus

| Letters | Frequency of 8 classes | Frequency of 32 classes |
|---|---|---|
| "r"/"g" | 157 | 431 |
| "o"/"o" | 99876 | 276108 |
| "н"/"n" | 93057 | 257314 |
| "a"/"a" | 83034 | 222646 |

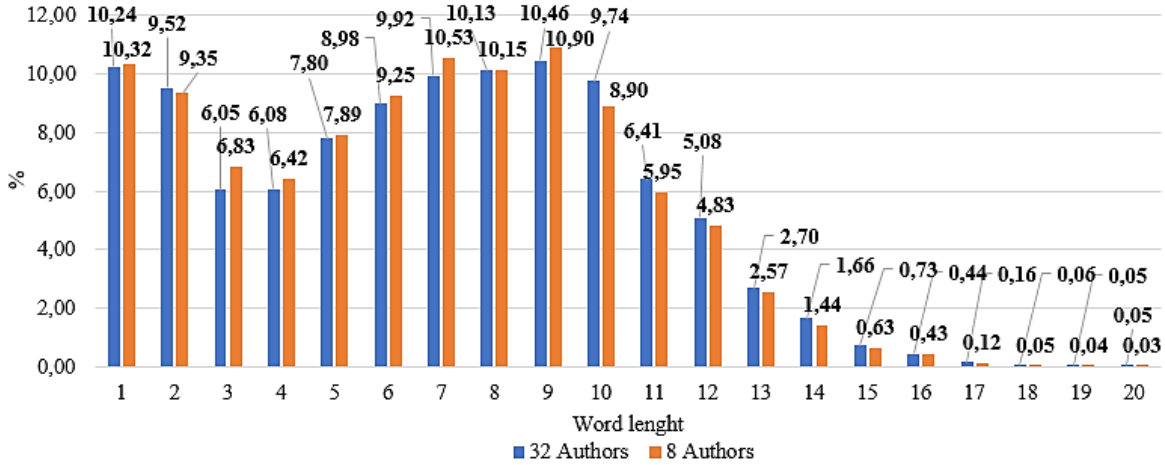


**Figure 3**: Percentage of words used in the corpus depending on their lengths

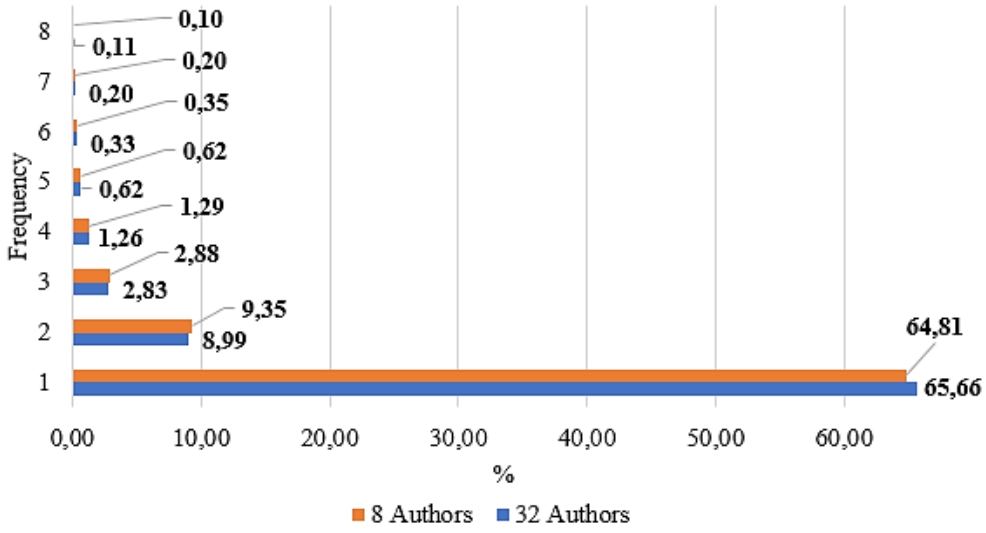


**Figure 4**: The number of words with a certain frequency

*Group 4*: frequency of using stopwords and pronouns (65 indicators). The dictionary of prepositions, pronouns, conjunctions was formed for the calculation of group 4 indicators. We used words with a frequency of 100 or more for the classification process. The stop-words and pronouns are shown in Table 3 and Table 4.

*Group 5*: coefficients of language diversity (5 indicators).

The author's unique vocabulary consists of specialized terms, functional words, certain linguistic constructions and influences the author's text variety. This variety is determined using the following coefficients [13, 16]:

- coefficient of text lexical diversity $K_l = \frac{W}{N}$, where $W$ is the number of unique words, $N$ is the total number of words;

- syntactic complexity coefficient $K_s = 1 - \frac{P}{W}$, where $P$ is the number of sentences, $N$ is the total number of words;
- exclusivity index $I_{wt} = \frac{W_1}{W}$, where $W_1$ is the number of words with a frequency of 1, $W$ is the total number of words;
- text concentration index $I_{kt} = \frac{W_{10}}{W}$, where $W_{10}$ is the number of words with a frequency of 1, $W$ is the total number of words;
- speech connectivity coefficient $K_z = \frac{Z+S}{3P}$, where $Z$ is the number of prepositions, $S$ is the number of conjunctions, $P$ is the number of sentences.

**Table 3**

Stopwords and pronouns with a frequency of 20 or more (for 8 classes)

| Word | Frequency | Word | Frequency | Word | Frequency |
|---|---|---|---|---|---|
| а/and | 710 | на/on | 2357 | той/that | 32 |
| або/or | 672 | над/over | 35 | тому/so | 249 |
| але/but | 128 | наприклад/for example | 207 | тільки/only | 87 |
| без/without | 81 | не тільки/ not only | 24 | у/in | 2155 |
| в/in | 2535 | ними/them | 48 | це/it | 363 |
| вона/she | 83 | них/them | 85 | цей/this | 85 |
| вони/they | 112 | ніж/than | 47 | цих/these | 86 |
| все/all | 35 | однак/however | 49 | цього/this | 153 |
| всі/all | 99 | оскільки/because | 91 | цьому/this | 112 |
| всіх/all | 110 | при/at | 593 | ця/this | 41 |
| від/from | 639 | проте/but | 61 | ці/these | 85 |
| він/he | 111 | під/under | 164 | цієї/this | 83 |
| для/for | 1952 | та/and | 2518 | чи/or | 135 |
| до/to | 1115 | так/so | 203 | що/what | 1365 |
| з/with | 2026 | таким чином/so | 57 | щоб/in order to | 95 |
| за/by | 1034 | також/also | 250 | як/as | 663 |
| й/and | 163 | те/that | 49 | якщо/if | 293 |
| його/his | 335 | ти/you | 60 | і/and | 2398 |
| коли/when | 90 | тим/by that | 59 | із/from | 343 |
| ми/we | 75 | тих/then | 128 | їх/their | 408 |
| мов/as | 55 | то/then | 418 | її/her | 323 |
| між/between | 313 | того/that | 327 | | |

After calculating the text diversity coefficients, we determined their average values for each author. Results were grouped by value. Thus, we found that a significant number of authors belong to each group of indicators for each coefficient of text diversity. Unfortunately, it does not allow independent use of these properties to identify the author. In Table 5 are presented the number of unique mean values for the text diversity coefficients and the maximum number of authors with the same indicators.

The analysis of the text diversity coefficients showed that these coefficients could not be a formal feature and could not be used for the author identification. Therefore, these indicators are used together with additional properties. For example, according to the Decision Trees (LMT) algorithm, with the addition of text diversity coefficients to the set of properties, the F-Measure value increased from 0.599 to 0.614.

**Table 4**
Stopwords and pronouns with a frequency of 100 or more (for 32 classes)

| Word | Frequency | Word | Frequency | Word | Frequency |
|---|---|---|---|---|---|
| а/and | 1852 | на/on | 6815 | той/that | 112 |
| або/or | 1460 | над/over | 109 | тому/so | 686 |
| але/but | 479 | наприклад/example | 480 | тільки/only | 336 |
| без/without | 214 | не тільки/ not only | 135 | у/in | 5755 |
| в/in | 7088 | ними/them | 154 | це/it | 1075 |
| вона/she | 253 | них/them | 240 | цей/this | 195 |
| вони/they | 308 | ніж/than | 150 | цих/these | 268 |
| все/all | 137 | однак/however | 144 | цього/this | 404 |
| всі/all | 255 | оскільки/because | 267 | цьому/this | 438 |
| всіх/all | 293 | при/at | 1481 | ця/this | 120 |
| від/from | 1649 | проте/but | 216 | ці/these | 230 |
| він/he | 270 | під/under | 356 | цієї/this | 224 |
| для/for | 4284 | та/and | 7531 | чи/or | 456 |
| до/to | 3368 | так/so | 593 | що/what | 4157 |
| з/with | 5366 | таким чином/so | 265 | щоб/in order | 205 |
| з/by | 2622 | також/also | 705 | як/as | 2088 |
| й/and | 823 | те/that | 185 | якщо/if | 558 |
| його/his | 1075 | ти/you | 129 | і/and | 6948 |
| коли/when | 240 | тим/by that | 198 | Із/from | 733 |
| ми/we | 180 | тих/then | 128 | їх/their | 1211 |
| мов/as | 105 | то/then | 418 | її/her | 850 |
| між/between | 802 | того/that | 327 | | |

**Table 5**
The analysis result of the uniqueness level of the average values of the text diversity indicators by the authors

| | Number of unique values (8/32 classes) | MAX number of authors with the same value of the indicator (8/32 classes) |
|---|---|---|
| Coefficient of text lexical diversity | 15/25 | 11/38 |
| Coefficient of syntactic complexity | 9/13 | 17/97 |
| Exclusivity index | 17/32 | 12/28 |
| Concentration index | 7/11 | 38/178 |
| Coefficient of speech connectivity | 52/127 | 5/12 |

## 3.2. Experiments

For our experiments of text classification, we used our corpus (2 subsets) and five groups of features: for separate groups and their combination. Weka software (https://www.cs.waikato.ac.nz/ml/weka/) was used for classification task. Bayes Based Algorithms (Naive Bayes Multinomial, NBM), Support Vector Machine (SMO), Decision Trees (LMT, J48) were used as classification methods with the cross-validation parameter – 10 folds. According to the preliminary experiments of the authors using other stylometric properties for machine learning methods are shown to demonstrate good results [12].

The Figures 5 shows the classification results of 1, 2 and 3 groups separately and together. For individual groups, quality is not assessed using F-measure. F-measure cannot be calculated or has very little value for them. The result is presented in the percentage of correct classification. In the experiments, fragments of documents from 8 authors were used, that is, a classification was carried out into 8 classes. The results of 32 classes are not as good as for 8 classes. In our opinion, this is due to the fact that with an increase in the number of classes, the influence of statistical characteristics decreases.

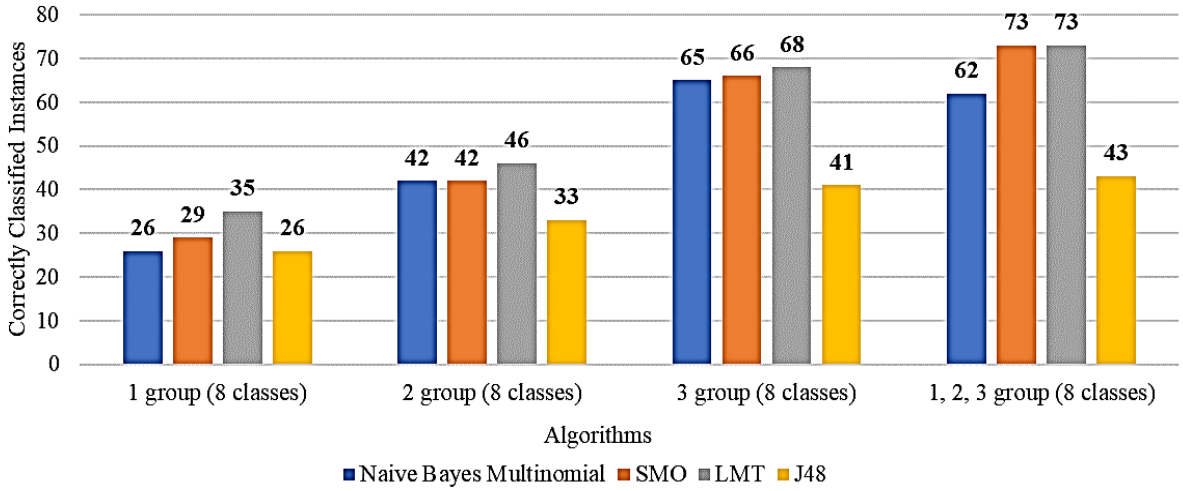


**Figure 5**: Comparison of the classification result for separate groups of properties and their combinations (8 classes)

Experiments have confirmed that the classification quality depends on sets of features and machine learning methods. In addition, scientific publications analyze other problems associated with the preparation of data that affect the quality of the classification [2]:

• Problem Scope – The number of authors in the research, equal to the number of classes in the classification.

• Training Size – The number of documents in the training set.

Therefore, we conducted experiments for 32 authors (32 classes) and 8 authors (8 classes) and compared the results. The result of the text classification using different sets of features for a different number of classes is presented in Table 6.

According to the experiments, we obtained an average increase in the value of Correctly Classified Instances: 20%, MIN Correctly Classified Instances: 15%, MAX Correctly Classified Instances: 28%. The dynamics of changes in the classification quality for 8 authors (8 classes) using a different number of indicators is shown in Fig. 6.

## 4. Conclusions

In this paper we described our experiments on text authorship attribution. These experiments were run on our own text corpus of scientific publications. Bayes Based Algorithms (Naive Bayes Multinomial, NBM), Support Vector Machine (SMO) and Decision Trees (LMT, J48) from the WEKA toolkit were tested for this task. Two sets of experiments have been designed, with selections of texts written by 32 and 8 authors respectively. As a novelty, we proposed our own set of author style indicators, organizing them in 5 groups and testing these groups individually and in various combinations.

The result of the experiments demonstrated the usefulness of the proposed language-independent stylometric properties indicators for text authorship attribution. Experiments showed that for 1-3, 1-4 and 1-5 groups of properties, the classification indicators are similar, despite the increase in the number of features.

**Table 6**
Classification quality metrics for property groups

| Group number | Algorithm | Number of indicators | Correctly Classified Instances (32 authors/ classes) | F-Measure (32 authors/ classes) | Correctly Classified Instances (8 authors/ classes) | F-Measure (8 authors/ classes) |
|---|---|---|---|---|---|---|
| 1 | NBM | 5 | 10.74 % | - | 26.49 % | - |
| 1 | SMO | 5 | 11.31 % | - | 29.24 % | - |
| 1 | LMT | 5 | 14.12 % | - | 34.54 % | - |
| 1 | J48 | 5 | 10.74 % | 0,100 | 26.30 % | 0,26 |
| 1, 2 | NBM | 33 | 18.11 % | 0,184 | 35.52 % | 0,37 |
| 1, 2 | SMO | 33 | 20.96 % | - | 44.35 % | 0,37 |
| 1, 2 | LMT | 33 | 26.81 % | 0,249 | 45.63 % | 0,47 |
| 1, 2 | J48 | 33 | 13.71 % | 0,135 | 33.95 % | 0,33 |
| 1, 2, 3 | NBM | 66 | 42.23 % | 0,433 | 62.21 % | 0,63 |
| 1, 2, 3 | SMO | 66 | 49.79 % | 0,484 | 73.79 % | 0,734 |
| 1, 2, 3 | LMT | 66 | 55.33 % | 0,550 | 73.30 % | 0,731 |
| 1, 2, 3 | J48 | 66 | 18.61 % | 0,182 | 43.27 % | 0,433 |
| 1, 2, 3, 4 | NBM | 131 | 49.03 % | 0,500 | 70.75 % | 0,720 |
| 1, 2, 3, 4 | SMO | 131 | 58.71 % | 0,581 | 79.78 % | 0,79 |
| 1, 2, 3, 4 | LMT | 131 | 60.27 % | 0,599 | 78.60 % | 0,78 |
| 1, 2, 3, 4 | J48 | 131 | 21.07 % | 0,205 | 49.06 % | 0,48 |
| 1, 2, 3, 4, 5 | NBM | 136 | 48.87 % | 0,499 | 70.65 % | 0,71 |
| 1, 2, 3, 4, 5 | SMO | 136 | 59.286 % | 0,586 | 79.4897 % | 0,794 |
| 1, 2, 3, 4, 5 | LMT | 136 | 61.6407 % | 0,614 | 80.6673 % | 0,806 |
| 1, 2, 3, 4, 5 | J48 | 136 | 21.0786 % | 0,206 | 49.4603 % | 0,492 |

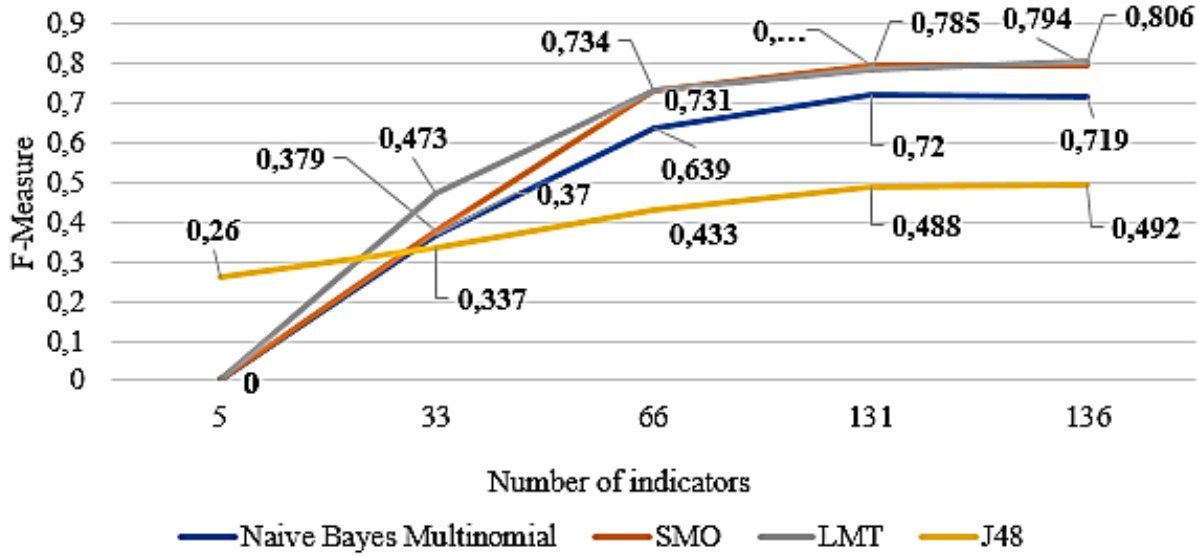


**Figure 6:** Dependence of the F-measure classification quality on the indicator numbers (8 classes)

The best result (F-measure) of 32 classes we received for the SMO method (0.586) and LTM (0.614) for 1-5 groups of properties. The best result (F-measure) of 8 classes we received for the SMO method (0,794) and LTM (0,806) for 1-5 groups of properties.

## 5. References


[1] E. Stamatatos, A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, Vol. 60, no 3, (2009), pp. 538-556. doi:10.1002/asi.21001.

[2] R. Zheng, Li. J. Chen, Z. Huang, A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American society for information science and technology, Vol. 57. №. 3, (2006), pp. 378-393. doi:10.1002/asi.20316.

[3] M. AlSallal, R. Iqbal, V. Palade, S. Amin, V. Chang, An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, Vol. 96., (2019) pp. 700-712. doi:10.1016/j.future.2017.11.023.

[4] B. Alhijawi, S. Hriez, A. Awajan, Text-based authorship identification - A survey. Paper presented at the 5th International Symposium on Innovation in Information and Communication Technology, ISIICT 2018. (2018), pp. 1-7. doi:10.1109/ISIICT.2018.8613287.

[5] O. Halvani, L. Graner, R. Regev, TAVeer: An interpretable topic-agnostic authorship verification method, ACM International Conference Proceeding Series, 2020.

[6] L. Zheng, H. Zheng, Authorship Attribution via Coupon-Collector-Type Indices, Journal of Quantitative Linguistics, Vol. 27, no. 4, (2020), pp. 321-333. Doi:10.1080/09296174.2019.1577939.

[7] N. Sadman, K. Datta Gupta, M. A. Haque, S. Sen, S. Poudyal, Stylometry as a Reliable Method for Fallback Authentication, 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2020, (2020), pp. 660.

[8] J. Bevendorff, T. Wenzel, M. Potthast, M. Hagen, B. Stein, On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification, IT - Information Technology, vol. 62, no. 2, (2020), pp. 99-115. doi:10.1515/itit-2019-0046.

[9] N. M. Sharon Belvisi, N. Muhammad, F. Alonso-Fernandez, Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features, 2020 8th International Workshop on Biometrics and Forensics, IWBF 2020 – Proceedings, 2020.

[10] J. Bevendorff, M. Potthast, M. Hagen, B. Stein, Heuristic authorship obfuscation, ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2020, pp. 1098.

[11] S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, M. Veinovic, Automated language-independent authorship verification (for Indo-European languages). Journal of the Association for Information Science and Technology, Vol. 70.8, (2019), pp. 858-871. doi:10.1002/asi.24163.

[12] V. Bobicev, Y. Hlavcheva, O. Kanishcheva, V. Lazu, Authorship Attribution in Scientific Publications. Proceedings of Corpora-2019 conference. Saint-Petersburg, Russia, (2019), pp. 174-181.

[13] V. Vysotska, O. Kanishcheva, Y. Hlavcheva, Authorship Identification of the Scientific Text in Ukrainian with Using the Lingvometry Methods, 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, (2018), pp. 34-38. doi:10.1109/STC-CSIT.2018.8526735.

[14] A. Bacciu, M. La Morgia, A. Mei, E. N. Nemmi, V. Neri, J. Stefa. Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features. In CLEF, (2019).

[15] R. Shukla, Foundations and Applications of Authorship Attribution Analysis, 2019, 38 p.

[16] V. A. Vysotska, V. V. Pasichnyk, Yu. M. Shcherbyna, T. V. Shestakevych. Matematychna lingvistyka. Knyha 1. Kvantytatyvna linhvistyka, Lviv, Novyi Svit–2000, 2012.