MCIT
Modeling, control & information technologies

# *A Comprehensive Methodology for Authorship Attribution of Literary Textual Works*

https://doi.org/10.31713/MCIT.2025.043

Rustam Azimov
Laboratory of recognition, identification
and methods of optimal solutions
Institute of Control Systems
Baku, Azerbaijan
rustemazimov1999@gmail.com

*Abstract*—**This study presents a comprehensive framework for authorship recognition, specifically tailored for literary works of fiction. There are two peculiarity of the considered problem. First, the problem involves authorship attribution of both large and small literary works. Second, the texts which are available for identification and testing of a recognition model are quite limited, since many writers write only a few literary works throughout their lives. In the study approaches used for addressing the aforementioned issues. Computer experiments were carried out on the Azerbaijani writers example.**

**Keywords—authorship attribution, stylometry, machine learning, feature engineering, Azerbaijani language**

## I. Introduction

The analysis of textual authorship has evolved from a niche literary pursuit into a critical discipline with applications in cybersecurity, forensic linguistics, legal disputes, etc. The core challenge lies in identifying quantifiable stylistic markers that are consistent for an author yet discriminative across different authors. This problem, known as authorship attribution or authorship identification, operates on the premise that every author possesses a unique, measurable "stylistic fingerprint" and each text of that author is considered to constitute a stylistic event as a variation of "stylistic fingerprint" in some sense [1, 2].

Most of the scientific studies regarding authorship attribution relies on the following types of texts:

- journalistic texts in news, magazines,
- software source codes,
- fiction literature.

The first attempts to use computational textual stylistics – stylometry for analysis of a text was carried out the fiction texts which were about the historical disputes, such as the authorship of the Federalist Papers or the works of Shakespeare. The advent of the internet and the proliferation of digital text archives have exponentially increased the availability of texts, consequently, research has expanded to include the authorship analysis of online content, such as news articles, blog posts, and social media messages, alongside traditional literary works [3–5].

The paper is devoted to analyze the effectiveness of the used methodology for authorship attribution of literary fiction textual works. The methodology allows to build machine learning based models for recognition of authors of small as well as large literary works (e.g. novels, short stories).

## II. Dataset

### *A. Data Acquisition*

The foundation of any robust authorship attribution system is a high-quality, representative dataset. For this study, we compiled a collection of literary works from eleven renowned Azerbaijani writers. The initial dataset, designated as Dataset-0, consisted of 151 texts: 28 large works (e.g., novels, novellas) and 123 small works (e.g., short stories). The distribution of works per author was uneven, reflecting their actual literary output, with the number of large works ranging from 1 to 5 and small works from 0 to 46 per author.

### *B. Data Augmentation*

A common challenge in machine learning, particularly with literary texts, is the scarcity of data for training complex models. To mitigate this, we used a dataset augmentation strategy. Each of the 28 large works in Dataset-0 was divided into 10 separate, non-overlapping parts of approximately equal length. This process transformed Dataset-0 into Dataset-1, which contained 403 individual text samples (123 original small works and 280 text segments from large works). This approach significantly increased the number of observations, providing more data for the parametric identification of models and for feature selection procedures. Dataset-1 was subsequently split into a training set (approximately 80% of the texts, or 325 samples) and a test set (approximately 20%, or 78 samples), ensuring a representative distribution of authors in both sets.

The effectiveness of the trivial data augmentation strategy was analyzed using the two empirical analyses approaches.

The first empirical analysis checks the similarity of the segments to each other using a known statistical metric – variance to ensure the text segments of a literary work shows similar characteristics in terms of stylometry. This analysis was done on an example of a writer's large novel, which was segmented into 10 parts, then number representations of these 10 parts was compared using the variance value over each text feature.

The second analysis measures the relation among the following on a given feature:

- mean difference among the 10 segments of a large work,
- mean difference among 10 short stories (of the same author),
- pairwise mean difference among short stories and segments of the large text.

If the latter is somewhere in the middle of the other two over a feature, it means this feature is representative and effective in terms of data augmentation for text classification problems (here authorship attribution).

## III. Textual Feature Extraction and Selection

The core of authorship attribution lies in defining and extracting features that capture an author's unique stylistic signature. We investigated five primary classes of textual features, calculating their values for each text in our dataset.

- **Sentence Length Frequency:** This feature captures an author's propensity for using short or long sentences. We calculated the normalized frequency distribution of sentences based on their word count (e.g., the proportion of sentences with 5 words, 10 words, etc.) for each text. This reveals syntactic complexity and narrative pacing habits.
- **Word Length Frequency:** This measures the normalized distribution of word lengths (number of characters per word) within a text. It reflects an author's vocabulary complexity and preferences, indicating whether they favor concise or more elaborate diction.
- **Character N-gram Frequency:** This powerful feature type involves calculating the normalized frequencies of all possible contiguous sequences of 'n' characters (e.g., for n=1: "a", "b", "n"; for n=2: "ab", "an", "in"). Character n-grams are language-agnostic and can capture sub-word stylistic patterns, such as common prefixes, suffixes, and other morphological constructs inherent to the Azerbaijani language. We focused on unigrams (n=1) and bigrams (n=2).
- **Character N-gram Frequency Variance:** We proposed a novel feature based on the stability of character n-grams within a text. Each text was divided into several contiguous parts. For each character n-gram, we calculated its frequency within each part and then computed the variance of these frequencies across all parts. A low variance indicates an n-gram that is used consistently throughout the text, potentially offering a more robust and stable stylistic marker than raw frequency alone.
- **Word Frequency (Bag-of-Words):** This classic approach involves calculating the normalized frequencies of specific words in a text. We explored different strategies for selecting which words to use as features. Beyond simple frequency, we also experimented with using meta-features: the frequency of a word in the unified text of each candidate author (from the training set) and its frequency in the entire training corpus. This adds a relative, author-specific context to the raw word count.

Given the potentially vast number of features (especially for character n-grams and words), we implemented and compared several feature selection procedures. For words, one procedure selected the most frequently used words for *each author individually*, then merged these lists. Another procedure selected the most frequent words across the *entire training set*, irrespective of author. We also experimented with manual curation, removing non-discriminative words like common function words or character names. For character n-grams, a similar procedure selected the most frequent n-grams across the training set. Additionally another procedure was used to select character n-grams which differs in characterizing an author from authors, here characterizing degree of an author on a n-gram is defined based on the mean difference among the texts of an author.

These features were then grouped into different feature sets. Some sets contained features from a single type (e.g., only character bigrams), while others were mixed, combining features from different types (e.g., sentence length and word length frequencies together) to investigate potential synergistic effects.

## IV. Experiment Results

### *A. Machine Learning methods used in the study*

To perform the actual authorship attribution, we employed several established machine learning models, known for their effectiveness in text classification tasks [6]:

- **Support Vector Machine (SVM):** We utilized the Radial Basis Function (RBF) kernel.
- **Random Forest (RF):** We used 100 trees with a maximum depth of 5.
- **Artificial Neural Network (ANN):** We used multilayer feedforward networks with two hidden layers. The number of neurons was varied based on the input feature vector size. The models were trained using the Adam optimization algorithm with an early stopping criterion.

The experiments were carried out using the program libraries Scikit-Learn and Keras [7, 8].

### *B. Results on analyzing the effectiveness data augmentation approach to discriminate texts of different authors*

Let's look through the variance values among the text segments of a literary work on the following bigram examples

ər, di, də, ən, lə, la, ar, in, da, an

are the following

0.0983, 0.0012, 0.0014, 0.0014, 0.0014, 0.0015, 0.0015, 0.0017, 0.0018, 0.002.

Since the provided variance values are quite small it means that using these bigrams ensures that dividing a literary work into non-overlapping text segments does not distort the author stylistics significantly.

On the second analysis we consider that if the mean value of pairwise differences among small work and large work part pairs is in the middle of mean value of small works and mean value of the large work parts, the large text segmentation approach is legitimate in terms of discriminating different authors' texts. For example on the following bigrams

cl, şh, mp, ıı, aə, bk, je, ğs, pt, mx

the differences from the difference indicator among short works and large work parts and the mean of the other mean values are the following:

1.76438E-07, 1.35125E-07, 1.05477E-07, 9.55283E-08, 9.46659E-08, 9.4265E-08, 5.83959E-08, 5.30525E-08, 4.16172E-08, 6.44E-10.

These differences are quite small that it means "being in the middle of the other" assumption is not incorrect.

The results on the other most bigrams were analogical to the results provided before.

### *C. Results of recognition effectiveness*

The maximum recognition accuracies on different machine learning models and features selection procedures were given in Figure 1. As it seems from the figure 1 using the character n-gram frequencies and bag of words features were quite successful in the dataset-0 – fiction literature works. As it is clear from the results on the Test set with 75% reliability criterion frequencies of character n-grams were better than the frequencies of words – bag of words features.

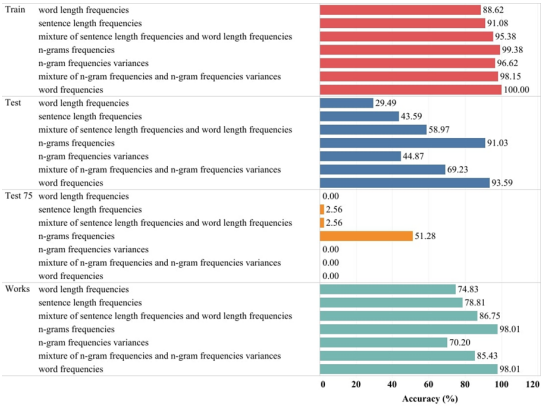


Figure 1. Recognition effectiveness evaluation metric results of text feature classes

Maximum recognition accuracies on different machine learning methods were given according to the results on different text sets in Figure 2. As it is clear from the figure multilayer feedforward artificial neural networks were not quite effective in recognition. Nevertheless the recognition accuracies when using SVM or Random Forest is about 98%. The results show that SVM was more reliable (see the results on the Test 75 row.

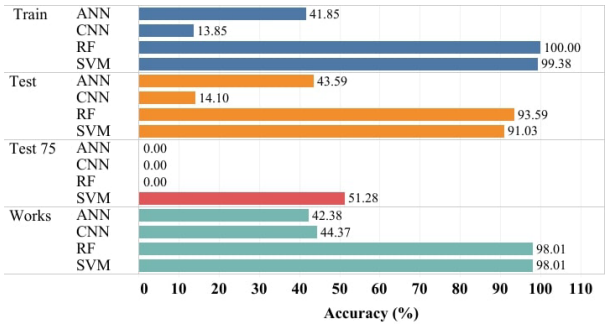


Figure 2. Recognition effectiveness evaluation metric results of machine learning methods

## V. CONCLUSION

In the scientific study challenges of authorship attribution of fiction literature works were addressed by a framework – methodology. This methodology includes diving a large fiction work to parts, analyzing the effectiveness of this on discriminating authors' texts, examining different text representations and feature selection procedures for using machine learning methods on authorship recognition.

## REFERENCES

[1] He, X.; Lashkari, A.H.; Vombatkere, N.; Sharma, D.P. Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey. Information 2024, 15, 131. https://doi.org/10.3390/info15030131.

[2] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

[3] Anisimov, A.V.; Porkhun, E.V.; Taranukha, V.Y. Algorithm for construction of parametric vectors for solution of classification problems by a feed-forward neural network. Cybern. Syst. Anal. 2007, 43, 161–170.

[4] Aida-zade, K.R., Azimov, R.B.. Analysis of the use of text features in the authorship identification of literary works in the Azerbaijani language // Informatics and Control Problems, – Baku, Azerbaijan: – 2024, v. 44, no. 1, – pp. 51-58.

[5] Azimov, R.B. (2024). Comparative analysis of using different text features, models, and methods in text author recognition. Cybernetics and Systems Analysis.

[6] Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

[7] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8] Chollet, F. (2015). Keras. GitHub repository. https://github.com/fchollet/keras.