Analysis of texts features for the author recognition system

Kamil Aida-zade[1], Rustam Azimov[2]

[1,2]Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan, Baku, Azerbaijan

[1]*kamil_aydazade@rambler.ru*, [2]*rustemazimov1999@gmail.com*

***Abstract* — In the study a comparative analysis of the recognition effectiveness of the use of text features used in the author recognition computer system in combination with different methods and models of machine learning to recognize the authors of literary works in the Azerbaijani language was carried out.**

Keywords* — *authorship recognition of texts; authorship attribution; authorship identification; machine learning

I. INTRODUCTION

Text authorship recognition has actual applications, such as the authorship of unattributed fictional works, as well as authorship identification of e-mails, texts on online forum pages, and blog posts.

Recognition of the authorship of texts is understood as selection of the author of a given text among certain author candidates based on the texts of those author candidates. For this, text samples known to be written by those author candidates are used [1]. Various methods and models of machine learning were used to establish the relationship between the features of these text samples and their authors.

After T.C. Mendenhall showed that the number of words of certain lengths (e.g. number of 1-letter words, number of 2-letter words, etc.) remained invariant within the text of C.J.H. Dickens' Oliver Twist, he compared the features of the texts of Shakespeare's and Bacon's works [2], [3]. A.A. Markov used the frequency of occurrence of vowel-vowel, consonant-vowel, and so on in A.S. Pushkin's "Eugene Onegin" in his work devoted to an application of statistics [4]. Mosteller and Wallace used frequencies of selected words to identify the authors of twelve unknown authors among the Federalist papers [5].

In another study, recognition of the authorship of texts was carried out on 634 fictions by 55 authors [6]. However, as a result of the increase in Internet resources, in recent years there have been more works devoted to authorship identification of short, non-fiction texts such as articles in newspaper columns [7]-[9].

Statistical methods, machine learning methods and models were used to recognize the authorship of texts in the Azerbaijani language [10]-[13].

In the study based on the results of computer experiments, a comparative analysis of the effectiveness of a number of text features and several machine learning methods and models that can be used to recognize the authorship of texts in the Azerbaijani language on an example of artistic works was conducted.

II. Problem Statement

In this study, the problem authorship recognition was considered on an example of the works of several famous Azerbaijani writers. A comparative analysis of the recognition effectiveness of the use of different feature sets with different methods and models of machine learning for use in a computer system that recognizes the authorship of texts in the Azerbaijani language was carried out.

A. *Used Dataset*

The dataset (dataset-0) consisting of 23 large-volume, 128 small-volume, total 151 works of 11 Azerbaijani writers was used in this work. Each of the large-volume works was divided into 10 parts (with almost equal number of characters), which were considered as separate texts, thus another dataset (dataset-1) consisting of 23x10+128=403 texts was prepared. The observations in this dataset-1 are split into training and test sets on 80-20 ratio (325 and 78 observations).

The recognition models were trained with the training set obtained from dataset-1, and their accuracy was evaluated on both the test set obtained from dataset-1 and dataset-0 itself. At the same time, accuracies were also evaluated on the test base obtained from dataset-1 with a certain confidence (recognition is rejected if a certain degree of confidence is not obtained, even if the recognition result is correct for that observation, it is considered wrong).

B. *Used Text Features*

Here, as text features, the frequencies of letter n-grams (a letter n-gram is a combination of n letters), the frequencies of words of a certain length (for example, the frequency of 5-letter words, the frequency of 6-letter words, etc.), the frequencies of sentences of a certain length (for example, the frequency of 5-word sentences, the frequency of 6-word sentences, etc.), letter n-gram variance features, which express how letter n-grams are distributed within a text (a given text is divided into parts, n-gram frequencies are calculated in each part, variance of them are found) were used. The feature sets consisting of different numbers of features selected from these features are described in Section III.

C. *Used Machine Learning Methods and Models*

Artificial Neural Network (ANN), Support Vector Method (SVM) and Random Forest (RF) were used in the study. Radial Basis Function was used as kernel in SVM. The characteristics of ANN architectures were defined according to the number of features. One set of letter 2-grams has also been used with a Convolutional Neural Network (CNN) in the form of a two-dimensional matrix.

III. Description of Used Feature Sets and Informative Features Selection Algorithm

A. *Used Feature Sets*

From the features of the frequencies of sentences of a certain length and the frequencies of words of a certain length (let's call them sentence frequencies and word frequencies for short), two feature sets were made each. In one of the sentence frequency set, there are frequencies of 5-word, 6-word, ..., 14-word sentences, and in the other, there are frequencies of 5-word, 6-word, ..., 29-word sentences. One of the word frequency feature sets contains the frequencies of 3-letter, 4-letter, ..., 7-letter words, and the other has the frequencies of 3-letter, 4-letter, ..., 12-letter words. The frequencies of 5-word, 6-word, ..., 14-word sentences and the frequencies of 3-letter, 4-letter, ..., 12-letter words were also used.

Letter 1-grams and selected letter 2-grams (The selection of informative letter 2-grams is described in the next paragraph) feature groups were also used. Seven feature sets were obtained by choosing 5, 10, ..., 25, 50, 100 letter 2-grams from the possible letter 2-grams. Feature sets consisting of the variances of frequencies of letter 1-grams and selected 2-grams in separate parts of the given text were also created. Several mixed feature sets consisting of letter n-grams and their within-text variances were also used (4 feature sets with 10, 20, 30, 40 features using 5, 10, 15, 20 features from each of these feature classes).

B. *Selection of Informative Letter bigrams*

For the selection of informative letter 2-grams, all the texts in the training database were combined into one text, the frequencies of all possible letter 2-grams in this single text were calculated, and the high-frequency 2-grams were selected.

IV. Results of Computer Experiments

The maximum recognition accuracies obtained on dataset-0 when using different models and methods of machine learning with different feature sets are given in Table I.

TABLE I. RECOGNITION ACCURACIES OF FEATURE SETS

Feature classes	**ANN**	**CNN**	**SVM**	**RF**
n-gram	31.13%	44.3%	93.38%	96.69%
variation	31.13%	-	39.74%	68.21%
n-gram + variation	9.93%	-	58.94%	85.43%
sentence frequencies	42.38%	-	65.56%	78.81%
word frequencies	5.30%	-	56.95%	74.83%
sentence-word frequencies	15.89%	-	66.89%	86.75%

Conclusion

In this study based on the results of computer experiments, a comparative analysis of the effectiveness of different methods and models of machine learning to recognize the use of different feature sets, composed of different numbers of features belonging to different classes of features for use in a system that recognizes the authorship of texts on an example of the fictional works of several Azerbaijani authors, is given.

The feature sets consisting of letter n-grams outperformed other features in recognition accuracy. Although Random Forest gives the highest recognition accuracy among the models and methods, more reliable results were obtained with Support Vector Machine.

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, pp. 538-556, 2009, doi: 10.1002/asi.21001.

[2] T. C. Mendenhall, "The Characteristic Curves of Composition," *Science*, no. 214, pp. 237-246, 1887, doi: 10.1126/science.ns-9.214s.237.

[3] Mendenhall, "A Mechanical Solution of a Literary Problem," The Popular Science Monthly, no. 60, pp. 97–105, 1901.

[4] A. A. Markov, "An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains," *Science in Context*, vol. 19, no. 4, pp. 591-600, 2006, doi: 10.1017/s0269889706001074.

[5] F. Mosteller and D. L. Wallace, "Inference in an Authorship Problem," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275-309, 1963, doi: 10.1080/01621459.1963.10500849.

[6] Y. Zhao, J. Zobel, "Searching with style: Authorship attribution in classic literature," *In Proceedings of the thirtieth Australasian conference on Computer science*, vol. 62, pp. 59-68, January 2007.

[7] J. Diederich, J. Kindermann, E. Leopold, G. Paass, "Authorship attribution with support vector machines," *Applied intelligence*, vol. 19, no. 1, pp. 109-123, 2003.

[8] I. Erdoğan, M. Güllü, H. Polat, "Development of a Letter Recognition Application with Machine Learning Algorithms," *El-Jezeri*, vol. 9, no. 4, pp. 1303-1314, 2022.

[9] V. Levent, B. Diri, "Author Recognition in Turkish Documents with Artificial Neural Networks," *Academic Informatics*, vol. 14, pp. 5-7, 2014.

[10] K. R. Ayda-zade, S. G. Talibov, "Analysis of the Methods for the Authorship Identification of the Text in the Azerbaijani Language," *Problems of information technology*, vol. 8, no. 1, pp. 14-23, 2017.

[11] R. B. Azimov, E. M. Mustafayev, "Comparison of SVM and ANN methods for recognition of authorship of texts". *Applied Mathematics and Fundamental Informatics: Proceedings of the XII Intern. youth scientific-practical. conf. with elements of science. schools*, pp. 60-61, May 2022.

[12] K. R. Aida-zade, E. M. Mustafayev, R. B. Azimov, "Features analysis for application in a computer recognition systems of Azerbaijani texts authorship," *Second International Bilateral Workshop on Science Between Dokuz Eylül University and Azerbaijan National Academy of Sciences*, p.11, November 2022.

[13] E. M. Mustafayev, R. B. Azimov, "Comparative analysis of different feature groups for use in a computer system for recognizing authors of texts in the Azerbaijani language," *II Republican scientific conference on "Fundamental problems of mathematics and application of intellectual technologies in education"*, pp. 34-39, December 2022.