22nd International Symposium INFOTEH-JAHORINA, 15–17 March 2023

Automated Authorship Attribution using CNG Distance on Blog Posts in the Serbian Language

Vlado Kešelj
Faculty of Computer Science
Dalhousie University
6050 University Ave, NS, Canada
vlado@cs.dal.ca

Abstract—**The automated authorship attribution problem is a task of identifying the author of a given text using an objective algorithmic method based on previous texts written by the candidate authors. We are particularly interested in methods that do not rely on any language-specific knowledge or preprocessing, and that are based on a low-level text representation such as a sequence of letters and other characters. The previous work has shown that author profiles consisting of the most frequent character n-grams are effective in the authorship attribution in a number of languages, but not many results are reported on languages with sparse resources, such as the Serbian and related languages. We show that a character n-gram based method has also a very good performance in the Serbian language. Another contribution of this work is a new dataset prepared as a good benchmark for the authorship attribution task, comparable to the previously published similar datasets for English, Greek, and some other languages. This dataset for authorship attribution prepared in this work consists of blog posts published as commentary columns on a news and commentary portal, and as such is a grammatical and well-written language corpus, and a good representative of current normative language. The CNG distance method, which was shown to work well in a number of languages before, shows high accuracy of 94% over 5 authors, and 83% over 10 authors in the authorship attribution for this dataset as well. As expected from the results for other European languages, the highest accuracy is obtained around n-grams of size $n = 6, 7$, or a wider range of $n = 3, \ldots, 8$, with L parameter from 500 to 9000, although even for the parameters $n = 2$ and $L = 500$ some relatively high accuracies are achieved.**

Keywords*—*Natural Language Processing (NLP); Text Classification; Automated Authorship Attribution; CNG distance;

I. Introduction

Authorship attribution is a study area of determining the author of a text based on the text content and other text samples of potential authors to be considered. This is an important problem with a long history [1] in the areas such as literature study, history and political social science. For example, the text could be a historical text, where the author is not known, the authorship may be disputed, or there may be other reasons that the actual author of a text is a person not previously associated with a text. One approach to the problem is to rely on linguistic experts and historians, who would typically analyze the text and present arguments based on some chosen features of the text, about why we should believe that a certain person was or was not an author of a text. It was recognized that this may be a subjective process, and that one could use mathematical and statistical methods to establish more objective arguments about authorship, based on assumption of stationary probability of authors using certain words, type of words, or other linguistic features.

Automated authorship attribution (AATT[1]) is an approach to the authorship attribution problem which does not rely on a subjective human factor, but the process of establishing authorship is automated using a computational method. An automated approach to authorship attribution can be based on statistical analysis [2], but it can also use many other forms of algorithms to establish authorship. AATT can be considered to be a special case of the general text classification problem [3]. The availability of massive amounts of text over the Internet made the AATT problem even more important since it can be used in different social media contexts such as detecting restaurant reviews written by the same person and plagiarism detection. AATT is based on style detection and the same methods are frequently applied to other style-based text classification problems, such as genre classification [4], author's age, gender and other demographic based classifications, dementia detection from speech [5], and health disorders detection from written comments.

Language-independent AATT is particularly interesting because such methods do not use language-specific features, they are very generic in processing text as a sequence of tokens, and as such can be generalized to many other domains of sequence classification. For example, the AATT methods have been applied to music classification [6], genomic classification[7], and malicious code detection [8].

AATT became particularly active area in the last two decades (since year 2000 or so), after a number of publications in this area appeared [9], [1], where more datasets were prepared, and a number of new approaches and challenges were explored. For example, the PAN workshop series on authorship attribution and other forms of authorship analysis were ran for many years (2011–2023) and are still going on [10].[2] However, while a number of major world languages are evaluated with different authorship evaluation methods, we are far from really understanding how generalizable these methods are to other languages. Contribution of this work is exploring the effectiveness of the CNG method [3] and character n-gram analysis on a particular corpus in the current standard Serbian language in the ijekavian dialect, but also further understanding how to determine the best n-gram profile sizes in general, and how minimal text size and number of

[1] AAAT is sometimes used as the abbreviation for Automated Authorship Attribution, or one could use AAA, but we find AATT to be easier to use, remember, and associate with the field.

[2] PAN workshops on authorship analysis: https://pan.webis.de/shared-tasks.html

978-1-6654-3778-3/22/$31.00 ©2023 IEEE

authors affect accuracy using the CNG method.

In the rest of the paper, we will proceed with discussion about background and related work in Section II, authorship attribution methodology that we used is discussed in Section III, the approach in preparing the dataset Serbian-Frontal-10 was discussed in Section IV, the obtained results are shown and discussed in Section V, and finally the conclusion and future work is presented in Section VI.

II. Background and Related Work

AATT has a relatively long history with many contributions and we will here mention some main publications leading to this work.

A famous example of disputed authorship was the case of the *Federalist Papers*, a list of 85 articles published mostly during 1787-8 in New York City's newspapers under a pseudonym, and later revealed to be written by Alexander Hamilton and James Madison. The authorship of some of the papers was disputed and some of the first work in AATT was the book by Mosteller and Wallace in 1964 [2], which used statistical methods, such as Bayesian inference, to make conclusions about the authorship of the disputed papers.

Holmes and Forsyth (1995) [11] revisited the case of the Federalist papers using a multivariate approach on vocabulary richness and most frequent words, and also applied a genetic algorithm. Since then, in parallel with machine learning and text mining advances, there were a number of publications examining different machine learning algorithms on a number of languages in the area of authorship attribution, authorship analysis, and similar style analysis tasks [12], [13], [10] from 2000 to 2023.

A. Character N-gram based AAAT and CNG

Our particular interest is in the methods based on character sequence analysis since these methods are very language independent and can be applied to any sequence of tokens (characters). For example, some languages do not use spaces to separate words (e.g., Chinese, Japanese, Thai), so even word segmentation is a non-trivial language dependent task. Character n-grams are all sequences of the fixed size n of consecutive characters obtained from a text in a process which can be visualized as a sliding window moving over the text. If $n = 1$ the n-grams are simply individual characters and their frequencies are character frequencies; if $n = 2$ they are called bi-grams, if $n = 3$ tri-grams, and so on.

Bennett in 1976 was first to report, as far as we know, that character n-gram frequencies, letter bi-grams more precisely, can be successfully used in authorship attribution [14] (sec. 4.10, pages 127–128). In his textbook about problem-solving using computers, he used programs in the BASIC programming language to obtain matrices of letter-pair correlations; i.e., frequencies of consecutive letter pairs, or letter bi-grams in other words. Using a similarity measure, he reported 100% accuracy of the method on "statistically significant samples of works" of 8 authors Hemingway, Poe, Baldwin, Joyce, Shakespeare, Cummings, Washington and Lincoln. The data was not released and it remains unclear about how much data would be needed for an authorship sample to be called a statistically significant sample. Bennett was also limited by the BASIC language and computer capacity at that time, so even though he mentioned using letter tri-grams; i.e., third-order correlation matrix as he referred to it, he did not use tri-grams or longer n-grams.

In 1999, Stamatatos *et al.* [9] created a Greek language corpus of 10 authors from articles of a weekly newspaper and using a linear regression approach with 22 style markers have achieved an accuracy of 65% in authorship attribution. This work was followed by the work of Peng, Kešelj *et al.* [15], [3], where they showed improved performance of authorship attribution by using character n-gram language models [15] and new character n-gram based distance between text profiles CNG [3]. The accuracy of 85% and 97% was obtained on two Greek datasets. The method CNG with Weighted Voting achieved the 1st rank result in one of the challenges of the AAAC competition in 2003 [16].

Since then, there were many uses of the CNG distance and other forms of classifications using character n-grams for authorship attribution and related tasks and it is sometimes used as a baseline approach [10], [17], [18]. However, to further establish how universally it can be applied to different natural languages the benchmark datasets and experiments need to be significantly expanded. Another question to be addressed is establishing how to choose the optimal n-gram sizes and profile lengths based on the size of training and testing data, as well as individual texts.

B. AATT for Serbian and Related Languages

Serbian language is one from a group of very related and mutually intelligible languages: Serbian, Croatian, Bosnian, and Montenegrin (SCBM).[3] It was formerly known as one language until break-up of Yugoslavia in 1990s, after which the four normative languages were established. An estimate is that these languages have about 19 million speakers (Wikipedia), who are mostly situated in the Balkans region of the Southern Eastern Europe. Some work has been published on authorship attribution in these languages but not much, and more research is needed, and in particular more datasets should be prepared.

Reicher *et al.* (2010) [19] reported results on authorship attribution experiments in Croatian. As claimed in the paper, this seems to be the first publication in authorship attribution in this group of languages (SCBM). Three relatively large datasets were prepared of news, blogs, and book chapters (25, 22, and 20 authors; 4571, 3662, and 1149 texts). The approach presented is a machine learning approach with up to 1241 features and the SVM classification method. The maximal achieved accuracies were very high: 91%, 93%, and 95%. This is a quite high accuraccy considering the number of authors in each dataset. Used methodology was very language dependent, and it uses information on POS tags, morphological categories, function words, and so on, and as such requires additional tools or manual annotation. The SVM and similar machine learning methods represent the contrastive classification approach where classifiers learn to detect authorship in contrast with other given authors, while the CNG approach uses self-contained author and text profiles and aims at finding objective distance measure to determine if a text is of the same author or not. The machine learning approach requires a lot of instances of texts, not typically available for one author, and this was addressed by treating chapters of a book as separate texts of each author, and breaking up other longer texts in a similar

[3] https://en.wikipedia.org/wiki/Serbo-Croatian

way. It seems that this may have introduced some bias if we have situation that several texts are part of the same subject matter, and some end up in the training set and others in the test set.

Zečević (2011) [20] prepared a dataset of articles of 3 authors in Serbian and obtained accuracy of 96% using the CNG distance. Zečević and Utvić (2012) [21] prepared a dataset of articles of 5 authors and reported results on n-gram based and syllable based profiles, where CNG distance with character n-grams have shown highest accuracy of 96%. Both datasets prepared by Zečević *et al.* were in the Serbian ekavian dialect.

Jamak *et al.* (2012) [22] applied principal component analysis (PCA) to analyze authorship style of classical writers in the SCBM languages, using books written at the time of Serbo-Croatian as the official language by the authors: Ivo Andrić, Meša Selimović, and Derviš Sušić. Samples of thousands of paragraphs from 6 books of these authors were analyzed. The analysis has shown visible difference in styles, but the analysis was qualitative so no accuracy could be measured.

Brodić *et al.* (2015) [23] successfully used a method for characterization and distinction between Serbian and Croatian languages using a feature representation and the GA-ICDA algorithm. This is a promising method that could be applied to the task of authorship attribution as well.

III. Methodology

It the methodology part we will first describe the CNG method, and then we will discuss details of data preprocessing.

A. CNG Method for AATT

The inception of the CNG method comes from the work of Bennett [14], where it was suggested that two texts can be compared for the same authorship by evaluating the following similarity measure:

$$\sum_{I,J}[M(I,J)-E(I,J)]\cdot[N(I,J)-E(I,J)], \qquad (1)$$

where $I, J \in \{1, 2, \dots, 26\}$ are indices of all letters of the English alphabet, the matrix values $M(I, J)$ are normalized frequencies of pairs of letters c_I and c_J of the first author, $N(I, J)$ analogue frequencies of the second author, and $E(I, J)$ analogue frequencies of the "standard English." *Normalized frequencies* were calculated by counting consecutive letter pairs (c_I, c_J) in the text and then dividing these counts by the total of all letter pairs, which means that all numbers in a matrix add up to 1. Larger value of the above expression should mean more similarity since the higher value is obtained when the same letter pairs have higher or lower frequency than the standard English in both matrices M and N.

The frequencies of the standard English may be difficult to determine and it is a language-depended parameter, so another somewhat similar measure is offered by Bennett as follows:

$$\sum_{I,J}[M(I,J)-N(I,J)]^2 \qquad (2)$$

which always gives a non-negative value, and where a smaller value means more similarity between texts. This measure is obviously the square of the Euclidean distance if we treat matrices as two vectors in the 26×26 space.

Our later experiments have shown that this Euclidean distance over character bi-grams does not give always a high performance, so we first extended it to use longer character n-grams ($n > 2$). Number of distinct n-grams obtained from a text rapidly increases as n grows, and they become more sparse in the sense of many possible n-grams not appearing at all in text. To address this problem we use only the first L most frequent n-grams for a given profile length L. Thus, for any text t, and given two parameters: n-gram size n and profile length L, we define the profile of the text to be the set of L pairs of the L most frequent n-grams and their frequencies:

$$f_t = \{(x_1, f_1), (x_2, f_2), \dots, (x_L, f_L)\} \qquad (3)$$

Any ties between frequencies can be resolved by lexicographic order of n-grams. In this way, a profile is defined as a function f_t which maps most frequent n-grams x to their frequencies $f_t(x)$. For any n-gram x which is not among the L most frequent n-grams; i.e., for which $f_t(x)$ is not defined, we will extend $f_t(x)$ to be zero:

$$f_t(x) = \begin{cases} f_t(x_i) & \text{if } x = x_i \text{ for some } i \in \{1, 2, \dots, L\} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Using Euclidean distance between profiles, as suggested by Bennett in formula (2), does not work that well as n grows because the frequencies of n-grams in a profile tend to decrease exponentially towards zero, similarly to the Zipf's law, and thus lower frequency n-grams do not affect significantly the distance between profiles and hence they are practically ignored in authorship attribution. This goes against the well-known observations in the authorship recognition that relatively rare features are still important in recognizing an author's style. For example, difference in frequencies between 0.03 and 0.01 between more frequent n-grams has much more effect on the Euclidean distance than the difference between 0.0003 and 0.0001 between less frequent n-grams, even though these differences are the same relatively to their absolute values and in our experience they should have about the same effect in the authorship attribution classification. For this reason, the following relative difference between n-gram frequencies is used:

$$\frac{f_1(x) - f_2(x)}{\frac{f_1(x)+f_2(x)}{2}} = \frac{2 \cdot (f_1(x) - f_2(x))}{f_1(x) + f_2(x)} \qquad (5)$$

For example, using this formula both differences between frequencies 0.03 and 0.01, and between 0.0003 and 0.0001 are the same (1.0). This is the motivation for introducing the CNG distance [3] between two profiles f_1 and f_2, which uses the sum of the squares of the relative differences:

$$\mathrm{cng}(f_1, f_2) = \Sigma_{x \in D(f_1) \cup D(f_2)} \frac{2 \cdot (f_1(x) - f_2(x))}{f_1(x) + f_2(x)} \qquad (6)$$

where $D(f_1) \cup D(f_2)$ is the union of all n-grams in the profiles f_1 and f_2; i.e., union of domains of the functions f_1 and f_2. As defined by the extension of f in eq. (4), if $x \notin D(f)$, we assume $f(x) = 0$. Finally, Algorithm 1 gives the algorithm for determining authorship of a given text t using the CNG algorithm. In summary, we concatenate all training texts written by the same author into one long text, and the profile of this text is the author's profile. Given a text t we choose the author whose profile is closest in terms of the CNG distance to the profile of the text t.

Algorithm 1 AATT_cng(t)

```
Require: $t$ is given text, $A = \{a_1, a_2, \ldots, a_k\}$ set of authors,
    $T$ is set of training texts with known authorship, profile
    parameters $n$ and $L$
Ensure: returns best matching author $a^*$ of text $t$
for all $i \in \{1, 2, \ldots, k\}$ do
    concatenate all texts $t' \in T$ written by $A_i$ into text $t_i$
    create profile $f_i$ of text $t_i$ (author profile)
create profile $f$ of the given text $t$
$i^* = \arg\min_{i=1..k} \mathrm{cng}(f, f_i)$
$a^* = a_{i^*}$
Return $a^*$
```

TABLE I. Serbian-Frontal-10 Dataset: labels, authors, minimal and maximal test blog size, and total data size (bytes)

		test blog size		data size (bytes)		
L.	Author	min	max	train	test	total
A0	Barašin S.	5258	12,092	54,266	77,025	131,291
A1	Grmuša M.	4554	15,942	115,126	82,183	197,309
A2	Trbojević R.	4138	9,935	82,762	75,311	158,073
A3	Grujić M.	3459	4,164	50,219	39,003	89,222
A4	Jokić A.	2980	9,174	57,631	52,236	109,867
A5	Vuković M.	2722	8,461	41,643	45,532	87,175
A6	Puhalo S.	2516	4,539	35,332	35,062	70,394
A7	Šehovac D.	2429	10,473	42,339	42,787	85,126
A8	Knežević M.	1448	2,825	19,566	22,995	42,561
A9	Mojović N.	1439	6,329	48,565	43,815	92,380
10	10 authors			547,449	515,949	1,063,398

B. Data Preprocessing

The raw data was downloaded from the web in HTML format and then processed using Perl scripts, which will be made publicly available. The title and content of each blog is extracted and saved. The blogs are visually checked to make sure they contain clean text, they do not contain additional notes that can reveal an author, and similar. According to these inspections, the following final preprocessing steps were developed:

1) **Alphabet normalization:** Most blogs use Latin and some Cyrillic alphabet, so they are all converted to the Latin alphabet so that n-grams could be properly matched, and to make conversion simpler since some blogs contained English words occasionally;
2) **Tag removal:** All HTML tags are removed, and paragraph tags replaced with empty lines
3) **New-line normalization:** It was made sure that each paragraph is in a single long line, and paragraphs separated by one empty line
4) **URL masking:** Any URLs are reduced to string `https://...` to remove details but still leaving minimal information since the use of URLs is a characteristic of an author's style
5) **HTML entity normalization:** A few HTML entities (e.g., `&scaron;`) were appropriately replaced
6) **Boilerplate text removal:** The end of the post sometimes included disclaimers, or even author names, and they were removed

The text included Serbian-specific letters (e.g., č, ć, etc.) and they were encoded in UTF-8, as in the original site. We used character n-grams of type 'byte', and all characters were included, including punctuation and new-line characters. Since paragraphs were single lines, we avoided accidental bias if some authors texts used shorter physical lines than others, and in this way the new-line character frequency correlated with paragraph length, which is a true style feature of an author.

N-gram extraction from texts and profile creation was implemented using the Perl module `Text::Ngrams.pm` [24].

IV. Dataset Serbian-Frontal-10

One of the contributions of this work is creation of a new dataset for evaluation of authorship attribution. The dataset and the code used in this study is available publicly at the web site https://vlado.ca/serbian-frontal-10 and on the GitHub as https://github.com/vkeselj/serbian-frontal-10.

For further and robust evaluation of a language-independent method like CNG, we need representative datasets from different languages with comparable characteristics. Following a similar approach of Stamatatos *et al.* [9], we looked into creation of corpus of about 10 authors, with well-written published articles on general topics in a contemporary Serbian language, with a focus on the dialect used in Bosnia and Herzegovina, or more precisely in its entity Republika Srpska.[4] Opinion columns are a good source of sampling a standardized language, as used by Stamatatos *et al.* [9], so we found a relatively popular blog portal Frontal[5] as a good source of a large number of blogs in the normative language that we want to sample. The blogs at this site are opinion columns in the areas of political, economical, and literary commentary. We would like to note that the content of blogs was not analyzed other than checking it over for necessary preprocessing, and verifying that it seemed to be genuinely written by the author. The goal was to create a reasonable sample for authorship attribution and opinions expressed in blogs were ignored in the process of selection.

Dataset preparation methodology is used as follows: the authors who published lately, starting from January 2023, were chosen first. A condition was that the author needed to have 20 blogs, which were collected in a reversed chronological order. The older 10 blogs of each author were used for training data and 10 more recent blogs were used as testing data. This corresponds to one typical application of authorship attribution where we would use older written data to detect authorship of newer articles. This creates a very balanced dataset in terms of files with 10 train blogs and 10 test blogs for each author. The total number of test blogs is 100, which also conveniently makes accuracy calculation rounded to the number of percentages equal to the number of correctly classified test blogs.

Test blogs that are very short might be a particular challenge for authorship attribution, so we decided to sort the authors according to the byte length of their shortest blog used in testing in a decreasing order. The list of authors with basic data size statistics shown in Table I. The minimal test blog length goes from about 5.3kB down to 1.4kB. The total training and testing size is a big over 0.5MB, so the total dataset size is a bit more than 1MB.

Table II shows minimal and maximal test blog size in words, the train, test, and total data size in words, and also average test blog size in words and bytes.

[4] Bosnia and Herzegovina has three official languages: Bosnian, Croatian, and Serbian.

[5] https://www.frontal.ba/blogovi

TABLE III. ACCURACY OF CNG ON SERBIAN-FRONTAL-10 DATASET, 10 AUTHORS

	N-gram size														
L	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
30	0.48	0.41	0.36	0.40	0.41	0.43	0.33	0.38	0.34	0.26	0.22	0.16	0.15	0.20	0.19
60	0.49	0.37	0.52	0.45	0.49	0.45	0.49	0.37	0.41	0.40	0.36	0.25	0.19	0.23	0.24
100	0.54	0.51	0.52	0.54	0.46	0.54	0.54	0.51	0.49	0.50	0.41	0.28	0.30	0.28	0.28
200	0.54	0.61	0.60	0.61	0.58	0.57	0.57	0.52	0.57	0.49	0.43	0.37	0.33	0.29	0.33
300	0.54	0.65	0.69	0.67	0.66	0.65	0.69	0.61	0.55	0.48	0.46	0.44	0.44	0.37	0.38
400	0.54	0.72	0.66	0.70	0.72	0.70	0.71	0.66	0.58	0.51	0.50	0.45	0.42	0.41	0.35
500	0.54	0.76	0.72	0.76	0.71	0.72	0.72	0.71	0.63	0.60	0.60	0.54	0.48	0.42	0.37
600	0.54	0.76	0.61	0.73	0.72	*0.79*	0.72	0.75	0.58	0.62	0.59	0.55	0.49	0.47	0.37
700	0.54	0.70	0.71	0.73	*0.81*	0.78	0.78	0.75	0.63	0.63	0.58	0.53	0.55	0.45	0.39
800	0.54	0.69	0.73	0.71	0.75	0.74	0.77	0.72	0.68	0.69	0.59	0.56	0.52	0.45	0.40
900	0.54	0.70	0.76	0.73	0.77	0.76	0.78	0.76	0.70	0.72	0.59	0.58	0.49	0.47	0.42
1000	0.54	0.63	0.74	0.77	0.77	0.73	0.77	0.74	0.72	0.70	0.62	0.59	0.50	0.47	0.40
1100	0.54	0.61	0.76	0.73	**0.83**	0.73	*0.79*	0.73	0.74	0.73	0.62	0.61	0.49	0.45	0.39
1200	0.54	0.61	0.75	0.76	*0.82*	0.74	0.75	0.71	0.74	0.68	0.63	0.61	0.50	0.45	0.42
1300	0.54	0.61	0.73	0.77	*0.81*	0.77	0.75	0.73	0.72	0.69	0.63	0.61	0.52	0.45	0.43
1400	0.54	0.61	0.76	0.73	*0.80*	0.75	0.74	0.76	0.73	0.69	0.62	0.61	0.52	0.46	0.41
1500	0.54	0.61	0.77	0.75	*0.80*	0.76	0.78	0.74	0.74	0.69	0.63	0.60	0.52	0.45	0.41
2000	0.54	0.61	0.77	0.77	0.72	0.76	0.74	0.77	0.74	0.72	0.64	0.54	0.50	0.47	0.41
3000	0.54	0.61	0.77	0.71	0.73	0.72	*0.79*	0.77	0.73	0.68	0.61	0.55	0.48	0.47	0.43
4000	0.54	0.61	0.73	0.74	0.74	0.71	*0.79*	0.76	0.73	0.68	0.58	0.52	0.52	0.47	0.45
5000	0.54	0.61	0.77	0.74	0.74	0.75	0.78	0.76	0.73	0.67	0.61	0.53	0.52	0.45	0.44
6000	0.54	0.61	0.74	0.74	0.74	0.73	0.74	0.78	0.72	0.67	0.61	0.55	0.51	0.45	0.43
7000	0.54	0.61	0.70	0.74	0.73	0.74	0.73	0.77	0.73	0.66	0.62	0.54	0.56	0.47	0.43
8000	0.54	0.61	0.70	0.73	0.75	0.72	0.74	0.77	0.71	0.65	0.63	0.57	0.56	0.49	0.45
9000	0.54	0.61	0.70	0.73	0.72	0.69	0.73	0.74	0.71	0.65	0.63	0.58	0.60	0.51	0.48
10000	0.54	0.61	0.70	0.74	0.69	0.69	0.72	0.73	0.72	0.68	0.64	0.58	0.62	0.52	0.49

TABLE II. SERBIAN-FRONTAL-10 DATASET SIZE IN WORDS; MINIMAL AND MAXIMAL TEST BLOG SIZE IN WORDS AND BYTES

	test blog size(w)		data size (words)			avg.test blog size	
	min	max	train	test	total	words	bytes
A0	826	1827	8024	11690	19714	1169	7703
A1	739	2305	17209	12282	29491	1228	8218
A2	675	1627	13627	12320	25947	1232	7531
A3	507	615	7181	5719	12900	572	3900
A4	473	1401	8760	8021	16781	802	5224
A5	390	1213	6103	6676	12779	668	4553
A6	395	693	5401	5384	10785	538	3506
A7	355	1387	6149	6124	12273	612	4279
A8	210	445	2947	3554	6501	355	2300
A9	224	1003	7626	6852	14478	685	4382
			83027	78622	161649		

TABLE IV. ACCURACIES PER NUMBER OF AUTHORS IN A TASK

# au.	2	3	4	5	6	7	8	9	10
acc.	0.90	0.93	0.93	0.94	0.93	0.90	0.86	0.84	0.83
corr.	18	28	37	47	56	63	69	76	83
test	20	30	40	50	60	70	80	90	100
err.	2	2	3	3	4	7	11	14	17

V. RESULTS

Table III shows accuracy over all 10 authors of the CNG distance approach, with the maximal accuracy of 0.83 achieved for $(n, L) = (5, 1100)$. The results use the standard train-and-test evaluation approach and show accuracies for a grid of combination of parameters n and L. To better observe the grid accuracies, the maximal accuracy is in bold and underlined font, accuracies within 5% of the best are in italic and underlined font, and accuracies within 10% of the best accuracy are in the underlined font.

This is a relatively good accuracy on fairly short blogs, where test blogs are as short as about 200 words, and most for most authors they range from about 400 to 1400 words. Similarly to other European languages, most high values in accuracy are achieved for n-grams of size 3–8, and with values within 5% of top in n-grams from 5–7.

To examine effect of the number of authors to the accuracy, we ran experiments for 2 authors (A0, A1), 3 authors (A0, A1, A2), 4 authors (A0, A1, A2, A3), and so on to 8, 9, and 10 authors. The obtained accuracies are shown in Table IV. Based on this table we can see that there is large variation in difficulty of separating styles of different authors. The first two authors A0 and A1 are relatively hard to separate with 2 posts misclassified and 90% accuracy. The third author A2 and fifth author A4 are easier to recognize since they do not introduce any additional misclassified posts. The author A7 seems to be relatively hard to recognize since this class introduces 4 more incorrect classifications.

The classification of 5 authors has highest accuracy of 94% and its full grid is shown in Table V. The highest accuracy of 94% is achieved for $(n, L) = (7, 3000)$. The area of parameters (n, L) where accuracy is highest is similar as in Table III, being a kind-of round area in the ranges $3 \leq n \leq 8$, and $500 \leq L \leq 9000$.

Table VI shows the confusion matrix for all 10 authors in the best classification case with $n = 5$, $L = 1100$, and accuracy 0.83. We can see that the blogs by the authors A2, A3, A5, and A9 had their authorship perfectly recognized. Two authors hardest to recognize were A6 (5/10 recognized), and A8 (6/10), and it is interesting that for those two authors the amount of training data was smallest, and the average test blog size was shortest, which seem like a reasonable explanation for this reduced performance.

If we examine different classification experiments by using

TABLE V. ACCURACY OF CNG ON SERBIAN-FRONTAL-10 DATASET, 5 AUTHORS

	N-gram size														
L	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
30	0.56	0.54	0.64	0.62	0.72	0.64	0.42	0.54	0.48	0.42	0.34	0.34	0.26	0.34	0.32
60	0.54	0.56	0.66	0.62	0.74	0.68	0.68	0.60	0.54	0.58	0.48	0.34	0.32	0.36	0.38
100	0.74	0.58	0.70	0.78	0.70	0.66	0.72	0.66	0.62	0.62	0.50	0.40	0.42	0.40	0.38
200	0.74	0.78	0.72	0.78	0.74	0.76	0.72	0.78	0.68	0.68	0.52	0.46	0.46	0.38	0.34
300	0.74	0.86	0.84	0.74	0.86	0.86	0.78	0.82	0.68	0.60	0.58	0.58	0.54	0.48	0.42
400	0.74	0.88	0.84	0.78	0.88	0.88	0.82	0.82	0.68	0.66	0.64	0.58	0.58	0.54	0.40
500	0.74	0.88	0.78	0.86	0.86	*0.90*	0.86	0.80	0.76	0.74	0.68	0.64	0.66	0.54	0.44
600	0.74	0.86	0.80	0.84	0.88	0.88	0.86	0.82	0.72	0.72	0.68	0.64	0.62	0.58	0.48
700	0.74	0.82	0.84	0.88	0.88	*0.92*	0.84	0.80	0.76	0.74	0.66	0.62	0.70	0.60	0.50
800	0.74	0.78	0.86	0.88	0.88	*0.90*	0.84	0.82	0.76	0.76	0.68	0.66	0.64	0.62	0.52
900	0.74	0.80	*0.90*	0.86	0.88	0.88	0.86	*0.90*	0.76	0.78	0.68	0.70	0.64	0.60	0.56
1000	0.74	0.70	0.88	*0.90*	0.88	0.88	0.84	0.84	0.80	0.82	0.70	0.70	0.64	0.60	0.56
1100	0.74	0.68	*0.90*	0.88	*0.92*	0.86	0.86	0.88	0.80	0.86	0.72	0.70	0.68	0.62	0.56
1200	0.74	0.68	0.88	*0.90*	*0.92*	0.86	0.86	0.82	0.78	0.80	0.72	0.72	0.70	0.60	0.54
1300	0.74	0.66	0.88	*0.90*	*0.90*	0.88	0.84	0.82	0.86	0.82	0.74	0.72	0.72	0.60	0.56
1400	0.74	0.66	*0.90*	0.84	*0.90*	0.88	0.88	0.84	0.86	0.80	0.74	0.72	0.70	0.62	0.54
1500	0.74	0.66	0.88	0.86	*0.90*	0.88	*0.92*	0.82	0.88	0.82	0.74	0.72	0.70	0.62	0.54
2000	0.74	0.66	0.84	*0.90*	0.88	0.88	0.84	0.88	0.86	0.84	0.80	0.74	0.68	0.60	0.54
3000	0.74	0.66	0.86	0.86	0.88	0.88	**0.94**	0.86	0.88	0.82	0.78	0.66	0.64	0.60	0.56
4000	0.74	0.66	0.80	0.88	*0.92*	*0.90*	0.88	0.86	0.86	0.82	0.72	0.68	0.62	0.58	0.58
5000	0.74	0.66	0.86	0.86	*0.90*	*0.90*	*0.92*	0.88	0.86	0.78	0.72	0.68	0.62	0.56	0.58
6000	0.74	0.66	0.82	0.86	0.86	0.88	0.88	0.88	0.82	0.78	0.72	0.66	0.62	0.60	0.56
7000	0.74	0.66	0.76	0.88	0.86	*0.92*	0.86	0.86	0.82	0.76	0.70	0.66	0.62	0.62	0.56
8000	0.74	0.66	0.76	0.86	*0.92*	*0.90*	0.86	0.86	0.82	0.76	0.70	0.68	0.62	0.64	0.58
9000	0.74	0.66	0.76	0.86	*0.90*	0.86	0.84	0.86	0.82	0.76	0.70	0.68	0.66	0.66	0.60
10000	0.74	0.66	0.76	0.82	0.86	0.86	0.84	0.84	0.82	0.76	0.72	0.68	0.66	0.64	0.60

TABLE VI. CONFUSION MATRIX FOR 10 AUTHORS, $n = 5$, $L = 1100$

	Predicted									
Actual	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9
A0	7	2	0	0	0	0	0	0	0	1
A1	0	9	1	0	0	0	0	0	0	0
A2	0	0	10	0	0	0	0	0	0	0
A3	0	0	0	10	0	0	0	0	0	0
A4	1	0	0	0	8	0	0	0	0	1
A5	0	0	0	0	0	10	0	0	0	0
A6	1	0	2	0	1	0	5	1	0	0
A7	0	0	0	0	0	0	2	8	0	0
A8	2	1	0	0	0	0	0	0	6	1
A9	0	0	0	0	0	0	0	0	0	10

2, 3, …, 10 authors, we can observe that parameters $n = 5$ and $L = 1100$ show consistently optimal or almost optimal performance. This could be a generally optimal parameter settings for this language, dialect, encoding, and general text length. When none of these parameters is known, an option is to use the CNG-wv (CNG with Weighted Voting) algorithm [16], which uses results in the CNG distance for an array of parameters, such as $n \in \{3, \ldots, 8\}$ and $L \in \{1000, 2000, 3000, 4000, 5000\}$, and an weighted voting method is used. The weighted vote used is the score $r = 1 - a/b$, where a is the CNG distance to the best label and b is the second closest distance. This approach leads to a 79% accuracy for 10 authors without dependence on parameters n or L.

Table VII shows for which pairs of parameters n and L the higest or nearly highest accuracy is achieved for classifications of 2 to 10 authors. We can see that certain areas of best parameters are relatively stable with consistently highest accuracy achieved for $n = 5$ and $L = 1100$.

In summary, these findings are relatively aligned with the previous findings for other European languages, such as Greek and English, where n-grams of size 6 or 7 have shown the best performance. We find that for this dataset, the best performance is generally achieved for $n = 5$ and $L = 1100$, although the n-gram sizes of 6 and 7 show also very high performance across a range of L values. The achieved performance is 94% for five authors, and 83% for 10 authors.

VI. CONCLUSION

We have presented a new dataset for authorship attribution for the Serbian language, focusing on a particular ijekavian dialect, based on blog post commentary articles. The dataset consists of 10 authors and for each author includes 10 posts for training and 10 posts for testing. The CNG method for authorship attribution shows highest accuracy of 94% for 5 authors, and 83% for 10 authors. For example, the 83% accuracy is achieved for the combination of parameters $n = 5$, $L = 1100$.

A. Limitations and Future Work

Since this dataset contains some very short blogs, an important question for future work is to find effect of these short blogs on accuracy, and determining the minimal size of training data, and minimal size of individual testing posts to achieve higher accuracy in authorship attribution.

The created dataset should be a valuable additional benchmark to evaluation other AATT algorithms, including those that require language-specific tools if and once they are available.

More classical machine learning algorithms that work in contrastive way should be applied on the dataset. It is expected that they may achieve higher accuracy in this standard setting

TABLE VII. Higest accuracies achieved for 2–10 author classification (⋆ denotes maximal accuracy, + is > 95% maximal, . > 90% maximal, and _ is < 90% maximal)

	N-gram size								
L	2	3	4	5	6	7	8	9	10
200	⋆..______	_________	__.______	__.______	_________	_________	_________	_________	_________
300	.++..____	_.._._.__	_.+______	.+⋆....__	__.._..__	_________	_________	_________	_________
400	.+⋆..++__	_________	__.______	⋆+⋆.._.__	..+....__	__.__..__	_________	_________	_________
500	.+⋆.+++_.	_..______	_.+...+..	.+⋆.._.__	.+⋆+...._	......+__	_________	_________	_________
600	_.+..++_.	__.______	_.+_._.__	..+....__	..+.+.+.+	__.....__	_____....	_________	_________
700	__.__..__	_.+___.__	_.+.._.__	.+⋆.+.+.+	⋆+⋆++.+..	____..+..	_____....	_________	_________
800	______.__	_.+...+__	.+⋆.._.__	_.+.+.+..	..+++++._	____.....	_________	_________	_________
900	______.__	⋆⋆⋆++.+..	_.+.._.__	_.+.+.+..	_...+++_.	__....+..	.++++⋆⋆+.	_________	_____..__
1000	_________	_.+...+__	.+⋆+++⋆..	__..++...	__.....__	____.....	_____..__	_____..__	_____..__
1100	_________	.+⋆++++_.	_.+...+._	.+⋆+⋆+⋆⋆⋆	__.....__	__......+	_..._..__	_____..__	__.....__
1200	_________	_.+...+_.	.+⋆++.+..	.+⋆+⋆+⋆++	___.._.__	__.......	_________	_____..__	_____..__
1300	_________	_.+...+__	.+⋆+..+..	_.+++++++	_.....+..	__.___._.	_____..__	___..++._	_____..__
1400	_________	.+⋆+++⋆..	_.+_._.__	_.+++.+.+	__..+.+..	.+⋆....._	__._.++..	__...++._	_____..__
1500	_________	.+⋆.++⋆..	_.+...._.	_.++++⋆.+	..+.+.+..	.+⋆++++..	____...__	_.+.+++._	_____.___
2000	_________	__._..+_.	.+⋆++.+..	_.+.._.__	__..+.+..	____...__	_...++⋆..	__....+__	____...__
3000	_________	_.+..++_.	_.+._____	__..._.__	__..._.__	.+⋆⋆++⋆.+	__...++..	__....+__	_________
4000	_________	__.__..__	_.+.._.__	_.++..+__	_.++._.__	__...+⋆.+	___._.+..	___._..__	_________
5000	_________	.+⋆..++_.	__.._..__	_.++._.__	__.+..+_.	_.++++⋆..	__....+..	___._..__	_________
6000	_________	__.__..__	__....+__	___.__.__	__..._.__	__..._.__	__....+..	______.__	_________
7000	_________	_________	_.+.._.__	___.._.__	_.++._.__	__..__.__	___._.+..	______.__	_________
8000	_________	_________	__..._.__	_.++._._.	__.+._.__	___.__.__	___._.+..	_________	_________
9000	_________	_________	__..._.__	__.+._.__	___._____	______.__	___.__.__	_________	_________
10000	_________	_________	__.___.__	___._____	___._____	_________	_________	_________	_________

of classification with 10 classes. A priority should be given to the other methods for authorship attribution.

The CNG method should also be compared to the methods that have shown high performance in text classification particularly on Serbian and related languages, such as the method based on the letter position features and the GA-ICDA algorithm reported by Brodić *et al.* (2015) [23].

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments and help in improving the work and suggestions for the future directions.

References

[1] H. Love, *Attributing Authorship: An Introduction*. Cambridge University Press, 2002.

[2] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer-Verlag, 1964.

[3] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003, pp. 255–264.

[4] J. Doyle and V. Keselj, "Automatic categorization of author gender via n-gram analysis," in *The 6th Symposium on Natural Language Processing, SNLP'2005*, Chiang Rai, Thailand, December 2005.

[5] C. Thomas, V. Kešelj, N. Cercone, K. Rockwood, and E. Asp, "Detecting and rating dementia of alzheimer type through lexical analysis of spontaneous speech," Fairfax, VA, USA, pp. 154–171, 2013.

[6] W. Jacek and V. Kešelj, "Evaluation of n-gram-based classification approaches on classical music corpora," in *Proceedings of Mathematics and Computation in Music*, Montreal, Canada, June 2013, pp. 213–225.

[7] A. Tomović, P. Janičić, and V. Kešelj, "n-gram-based classification and unsupervised hierarchical clustering of genome sequences," *Computer Methods and Programs in Biomedicine*, vol. 81, pp. 137–153, February 2006. [Online]. Available: http://www.sciencedirect.com/science/article/B6T5J-4J2TS9W-1/2/5403c6c4fc0a6a4a66f9ca2e46977f61

[8] T. Abou-Assaleh, N. Cercone, V. Kešelj, and R. Sweidan, "N-gram-based detection of new malicious code," in *Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004.*, vol. 2. IEEE, 2004, pp. 41–42.

[9] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic authorship attribution," in *Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*. Bergen, Norway: Association for Computational Linguistics, Jun. 1999, pp. 158–164. [Online]. Available: https://aclanthology.org/E99-1021

[10] PAN, "PAN workshops on authorship attribution," 2011–23 (accessed Jan 2023), https://pan.webis.de/shared-tasks.html.

[11] D. Holmes and R. Forsyth, "The Federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, pp. 111–127, 1995.

[12] P. Juola, "Authorship attribution for electronic documents," pp. 119–130, 2006.

[13] P. Juola *et al.*, "Authorship attribution," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2008.

[14] W. R. Bennett, *Scientific and Engineering Problem-solving with the Computer*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1976.

[15] F. Peng, D. Schuurmans, V. Kešelj, and S. Wang, "Automated authorship attribution with character level language models," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary, April 12–17 2003.

[16] V. Kešelj and N. Cercone, "CNG method with weighted voting," in *Ad-hoc Authorship Attribution Competition (AAAC)*, June 2004.

[17] D. Kosmajac and V. Kešelj, "Language distance using common n-grams approach," in *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2020, pp. 1–5.

[18] 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.

[19] T. Reicher, I. Krišto, I. Belša, and A. Šilić, "Automatic authorship attribution for texts in Croatian language using combinations of features," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 21–30.

[20] A. Zečević, "N-gram based text classification according to authorship," in *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, 2011, pp. 145–149.

[21] A. Zečević and M. Utvić, “An authorship attribution for Serbian,” in *BCI (Local)*. Citeseer, 2012, pp. 109–112.

[22] A. Jamak, A. Savatić, and M. Can, “Principal component analysis for authorship attribution,” *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, vol. 3, no. 2, pp. 49–56, 2012.

[23] D. Brodić, A. Amelio, and Z. N. Milivojević, “Characterization and distinction between closely related South Slavic languages on the example of Serbian and Croatian,” in *Computer Analysis of Images and Patterns (CAIP 2015)*, G. Azzopardi and N. Petkov, Eds. Cham: Springer International Publishing, 2015, pp. 654–666.

[24] V. Kešelj, “Perl package Text::Ngrams,” 2003–2023, https://vlado.ca/srcperl/Ngrams/Ngrams.html, https://metacpan.org/pod/Text::Ngrams, or https://github.com/vkeselj/Text-Ngrams.