Expert Systems With Applications 203 (2022) 117280

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A text as unique as a fingerprint: Text analysis and authorship recognition in a Virtual Learning Environment of the Unified Health System in Brazil

Marcella Andrade da Rocha [a,*], Philippi Sedir Grilo de Morais [a], Daniele Montenegro da Silva Barros [a], João Paulo Queiroz dos Santos [b], Sara Dias-Trindade [c], Ricardo Alexsandro de Medeiros Valentim [a]

[a] Laboratory of Technological Innovation in Health - LAIS/UFRN. Natal, Rio Grande do Norte, Brazil
[b] Federal Institute of Education, Science and Technology of Rio Grande do Norte - IFRN. Natal, Rio Grande do Norte, Brazil
[c] Univ Coimbra, Centre for Interdisciplinary Studies, Faculty of Arts and Humanities, Coimbra, Portugal

ARTICLE INFO

Keywords:
Stylometric features extraction
AVASUS
Text analysis
Authorship attribution
Author recognition

ABSTRACT

Authorship Attribution, the act of deducing the author of a given text based on its writing characteristics, is an issue with an extensive history. It refers to the task of properly recognizing the text's author within a specific group of candidates, based on relevant features extracted from the text (Stylometry). Hence, stylometry identifies relevant attributes that define a space in which authors can be distinguished. Because writers use language in different ways to express their ideas, linguistic variations make it possible to recognize authorship. The definition of the author's text is discussed, in this article, as an auxiliary tool in the distance education platform of the Ministry of Health, AVASUS. Therefore, the stylometric features were extracted from the collected data set, and different classification algorithms were trained. The objective was to predict the authorship of texts with more than 30 characters. As a result, an acknowledgment text for the Virtual Learning Environment of the Unified Health System in Brazil was obtained. The precision achieved in the classification process was 92% in some classifiers. This aspect suggests that techniques for extracting stylometric features may be used to recognize the author of a given text.

1. Introduction

In the early nineteenth century, it was considered challenging to determine the authorship of a document with less than 1000 words. However, at the beginning of the twenty-first century, it was considered possible to do so for a 250-word document. The need for this limit to be smaller and smaller is exemplified by the tendency of many shorter communication tools such as Twitter, Facebook, Short Message Services (SMS), and forums (Bhatnagar & Srinivasa, 2013).

Authorship Attribution exerts a substantial role in many applications, including authorship recognition and forensic investigation. Approaches to this problem attempt to identify the author of a document by analyzing the individual's writing style and/or the subjects/topics on which the author typically writes. The problem has been extensively studied, and a wide range of resources has been explored in the literature by Hürlimann, Weck, van den Berg, Suster, and Nissim (2015), Schwartz, Tsur, Rappoport, and Koppel (2013), Seroussi, Zukerman, and Bohnert (2014), Stamatatos (2013), and others.

However, the analysis of stylometric features in data sets of distance learning platforms or using a series of classifiers has been scarce in the literature so it can be better studied and analyzed to deepen the theme. Consequently, it is difficult to predict which features will be more effective for a given data set in authorship recognition.

Authorship Attribution makes a comparison of the author's writing style with the categorization of texts. When the differences between documents written by distinct authors are noticeable, the resources applied to text categorization or related to the content may be more effective. However, it is more plausible that resources based on the author's style are more efficient for more uniform sets of texts (Rocha, 2019).

Conventionally, the Authorship Attribution task is performed within two scenarios. The first refers to literary or historical research, in which attribution is requested for a text of revealed or unrevealed origin. In addition, such a process may identify potential authors in a set of candidates. The second scenario is forensic linguistics. In this way, it

* Corresponding author.
E-mail addresses: marcella.rocha@lais.huol.ufrn.br (M.A.d. Rocha), sedir.morais@lais.huol.ufrn.br (P.S.G.d. Morais), daniele.barros@lais.huol.ufrn.br (D.M.d.S. Barros), joao.queiroz@navi.ifrn.edu.br (J.P.Q.d. Santos), ricardo.valentim@lais.huol.ufrn.br (R.A.d.M. Valentim).

https://doi.org/10.1016/j.eswa.2022.117280
Received 6 November 2020; Received in revised form 23 October 2021; Accepted 19 April 2022
Available online 5 May 2022

must be determined whether or not an author (Halteren, 2007) has written a specific text, probably an incriminating one. Hence, the task is to verify the authorship by confirming or denying it by a single known author Halteren (2007). In this article, authorship analysis was performed in both described scenarios:verification and recognition of authorship.

In this context, this paper unveils the development of computational methods based on natural language processes applied to identify student authorship in online courses. The developed algorithms were used in the Brazilian "Virtual Learning Environment of the Unified Health System" (AVASUS) (Araújo et al., 0000; Rocha, Nóbrega, de Medeiros Valentim, & Alves, 2020; Valentim et al., 2021), as it is an online teaching platform with a large volume of data due to its thousands of students.

Virtual learning environments like AVASUS, with millions of users and enrollments, entail challenges on several dimensions. Hence, in identifying student authorship concerning their work and content produced online, this research provides results that contribute to the solution of such challenges in an automated way, with an excellent degree of accuracy. Besides, this study is carried out in a Portuguese-language database, while most of the research carried out in the field is based on English-written texts.

2. Background

2.1. Virtual learning environment of national health system — SUS in Brazil

The Brazilian Unified Health System (SUS) is one of the largest and most complex public health systems in the world. The assistance provided ranges from primary care, which includes common services as assessment of blood pressure, to the highly complex care, such as organ transplantation. Significantly, SUS provides full support and free access to the entire Brazilian population.

About five million professionals are part of SUS (DATASUS, 2020), such as involving community health agents; agents who combat endemic diseases; social workers; biologists; biomedical professionals; professors; nurses; students; pharmacists; physical therapists; speech therapists; managers; doctors; military; nutritionists; dentists; physical education professionals; psychologists or psychoanalysts; dentistry technicians (assistants and oral health technicians); technicians in clinical pathology; radiology technicians; occupational safety technicians; nursing technicians and assistants; occupational therapists; and orthopedists. The need to bring the training of these professionals closer to the actual needs of users and the system has been a challenge (SUS, 2020).

The use of technology-mediated courses in online environments aims at the continuing education of these professionals. Furthermore, it presents itself as an excellent solution for mass training. For this reason, the Virtual Learning Environment of the Unified Health System (AVASUS) was created from the integration of educational modules and training trails for health professionals. It is a project developed by the Laboratory of Technological Innovation in Health (LAIS) and by the Secretary of Distance Education (SEDIS), in cooperation with the Secretary of Labor Management and Health Education (SGTES) of the Ministry of Health, Brazil (Nóbrega, Souza, Barbosa, Coutinho, & Valentim, 2016).

In March 2020, when the last data collection was carried out, the platform had 217 active courses; 494,134 total registered users; 1,056,635 enrollments in courses; and 560,658 users entitled to certification were available for enrollment. The distribution of professionals or users enrolled in AVASUS courses by federation unit can be observed in Fig. 1. These data show the robustness of the platform, combining far-reaching and quality education on a large scale with mass training of professionals (AVASUS, 2020).

Fig. 1. Distribution of users enrolled in AVASUS courses by Brazilian State.

Currently, the version 2.0 of the AVASUS platform can be accessed through the website <https://avasus.ufrn.br/>. This recent version has an intuitive design and easy access, allowing users to navigate and reach their goal quickly through the "one click navigation" feature. Still, one can access menus, summaries, and other information without having to enroll. Enrolled students have access to all its content and can to make their educational itinerary by choosing modules related to their education. To obtain a certificate of completion, one must complete 100% of the module. In addition, version 2.0 was developed with responsive technology to improve the users' experience. Thus, the user can access the website from a smartphone, tablet or computer without impairing navigation (AVASUS, 2020).

In AVASUS courses – subdivided into specialization, extension modules or educational modules, web lectures and accessibility (educational modules with audio description or subtitling) –, clinical, work process and organizational topics are addressed based on SUS demands (Vieira et al., 2017). Core modules are: Emerging Respiratory Viruses, including COVID-19, and Introductory Course for Community Health Agent.

The first module had 64,273 students enrolled on April 6, 2020. It was developed to train health professionals and the general public with a focus on the COVID-19 pandemic, and to bring knowledge about emerging respiratory viruses and how to respond effectively to an epidemiological outbreak. The second one has 52,839 students enrolled, thus being designed to reinforce the significance of actions developed by community health agents and their insertion in improving the quality of the health services provided to the population. AVASUS courses are available in Portuguese; although subtitles/translation in English, Spanish, and Swahili are available for some courses. Such an aspect facilitates the worldwide reach of the platform and the sharing of knowledge with people from locations where more considerable difficulties in training professionals exist (AVASUS, 2020; da Silva et al., 2019).

2.1.1. Automation of Distance Learning Systems

Automation of Distance Learning Systems is starting and soon there will be no need for a tutor for the activities, but purely artificial intelligence for this task. Even though the Distance Learning System is incorporated on the internet, it is not completely independent of human

intermediation for teaching correction and assisting the students. The development of intelligent systems is dominating the environment and automating several systems, and this also involves distance learning, as AVASUS. Nonetheless, it is an expensive activity that requires research in several areas and that is carried out gradually. With the automation of the systems, machines will perform the tasks through artificial intelligence. Such a factor will lead to a reduction in costs and an improvement in quality, bringing satisfactory results and safety.

In the AVASUS platform, several courses do not have tutoring, thus improvements are being applied (Nóbrega et al., 2016; da Rocha, da Costa, de Medeiros Valentim, & de Pinho Dias, 2019; Souza, Vieira, Coutinho, & Valentim, 2018; Vieira et al., 2017) to improve the platform, making it more automated and reducing human intermediation. Authorship Recognition of texts typed by users in the course activities will further enrich the system and increase the automation of AVASUS, ensuring the integrity of the user who correctly accesses the platform in an automated way. Therefore, it may represent the beginning of a distance learning system mediated by technology and artificial intelligence.

2.2. Text analysis automation

Text Analysis Automation is a necessary task that is increasingly demonstrating its relevance because of the significant increase in the volume of texts generated every day on the Internet. As of yet, some approaches have been proposed (Stamatatos et al., 2014). Text analysis automation techniques are divided into three parts:

1. Lexical Characteristics: words, n-grams, slang, functional words, dialects, punctuation marks, n-gram characters, frequent suffixes, etc;
2. Text characteristics: phrases, whitespace, line length, nonalphanumeric characters, etc;
3. Syntactic characteristics: n-gram syntactic function, sentence types, morphological complexity, etc.

The item (2) is generally used for analysis of non-literary texts, such as source code; in contrast, the item (3) typically requires peculiar linguistic knowledge. Then, the item (1) is widely utilized for authorship recognition. There are several studies that employ lexical elements, such as characters (Markov, Gomez Adorno, Sidorov, & Gelbukh, 2016; Stamatatos, 2013), phonemes (Khomytska & Teslyuk, 2018) and morphemes (Rao, Raju, Latha, & Varma, 2017).

Moreover, there are the n-gram lexical elements, the subtraction of n elements of a sequence and the syntactic role a word makes in sentences. The representative of these automation techniques is the Bag of Words (BoW). It portrays the text as a generalized set of words, without considering their grammatical structure and order, but preserving their diversity and associating the frequency in which they occur (Markov, Stamatatos, & Sidorov, 2017).

Other studies have obtained results when utilizing lexical characteristics, as Altheneyan and Menai (2014) and Shojaee, Murad, Azman, Sharef, and Nadali (2013a). The latter used Hápax Legómenon, a word that appears registered only once, and Hápax Dyslegómenon, a word that appears registered twice in a text. To be specific, n-grams characters are the most popular for its noise tolerance and its effectiveness in unstructured documents such as emails. Although n-gram resources have proven to be effective, classification based on that is complex. What is more, data processing takes time (Brocardo, Traore, & Woungang, 2015).

2.2.1. Patterns in texts

Natural Language Processing Research (NLPR) is dedicated to developing techniques for extracting patterns in texts (Huang & Haralick, 2009). Patterns can be identified as:

- Syntactic patterns: parts of noun phrases, which are defined by grammatical rules;
- Semantic patterns: polysemic words can be identified by the contexts of the word.

3. Related works

3.1. Authorship attribution

Analyzing whether an established author composed a document, or determining of whom the authorship of a particular text is, incites relevant questions and arouses researchers' curiosity. To this end, models of texts by various authors are gathered and saved in a database, from where the style characteristics of each author are extracted.

A given author is recognized among all authors who are in the experiment as a whole. One can mention the cases of the questioned text of the Federalist Papers, used in Cerra, Datcu, and Reinartz (2014), investigating digital crimes (Schmid, Iqbal, & Fung, 2015) and identifying terrorist messages via the web. Many of them consider data sets collected from online sources, such as websites, blogs, and social networks. Still, many follow the same general approach of preprocessing or filtering text to remove unwanted characters, digits, punctuation marks, etc. Next, the extraction of characteristics is carried out and, if necessary, reduction techniques are applied. In summary, standard classifiers such as Naive Bayes (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM).

In the scientific field, Authorship Attribution has been considerably developed over the past decade, taking advantage of advances in computing techniques, as machine learning, information retrieval, and natural language processing. The diversity of digital texts available – email messages, blogs, online forums, source codes, etc. – points out that the existing technology may handle noisy texts by several candidate authors because of a wide variety of applications (Stamatatos, 2011).

The term Authorship Attribution refers to writing style and to the task of recognizing the author of a text in a group of candidate authors based on relevant features extracted from the text. Furthermore, stylometry occurs from identification of relevant attributes that define a space in which authors can be distinguished from each other. Hence, variations in language usage make Authorship Assignment possible because writers express their ideas in unique ways.

In recent studies (Akimushkin, Amancio, & Oliveira, 2018; Al-Ayyoub, Jararweh, Rabab'ah, & Aldwairi, 2017; Albadarneh et al., 2015), researchers mention that most of the Authorship Attribution algorithms are based on a simplified representation model used in natural language processing and information retrieval, known as Bag of Words (BoW). In BoW's approach, texts are represented by the frequency of words without considering the grammatical relationship. The algorithms made using BoW conjecture that an author's style is basically described by the probability distribution of certain words, phrases or any other relevant structure (Neme, Pulido, Muñoz, Hernández, & Dey, 2015). Several studies have obtained results using lexical characteristics (Altheneyan & Menai, 2014; Shojaee, Murad, Azman, Sharef, & Nadali, 2013b). The latter research (Shojaee et al., 2013b) used Hápax legómenon (word registered only once) and Hápax dilegómenon (word registered twice) in one language.

Some papers using syntactic n-grams (sn-gram) (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014) and binary n-gram (Peng, Choo, & Ashman, 2016), make variations to the n-gram method (sequence close to n items of a certain text sample) to achieve better results. In particular, character n-grams are the most popular for their noise tolerance and effectiveness in unstructured documents such as e-mails. Although the n-gram resources have proven to be effective, classification based on them is complex (Brocardo et al., 2015). Nonetheless, characters n-grams are not consistently better at sorting accuracy (Cerra et al., 2014). In brief texts, the word n-gram approach becomes sparse, as the combination of words is not found. Such an aspect makes it difficult to classify by algorithms.

Authorship Attribution involves the following tasks, the first being the focus of this paper's proposal:

- Authorship Verification/Recognition (i.e., to decide if a certain text was written by a certain author, or what is the authorship of the text?) (Brocardo et al., 2015);
- Plagiarism detection (i.e., comparing two texts and finding similarities in writing) (Franco-Salvador, Rosso, & Montes-y Gómez, 2016);
- Author's profile or characterization (i.e., extraction of information about the age, education, sex, ideology, etc. of the author of a given text) (Ashraf, Iqbal, & Nawab, 2016);
- Detection of stylistic inconsistencies (as can happen in collaborative writing) (Tschuggnall & Specht, 2013).

This paper concentrates on the problem of authorship verification and authorship recognition, exposing the criteria of the methods of Authorship Attribution. What may distinguish this paper from others is that authorship recognition is presented as a resource for a Virtual Learning Environment (VLE) and the use of the Portuguese language. In addition, the use of short texts (between 70 and 300 characters) characterizes the originality of this research.

4. Methodology

4.1. Authorship recognition system

Fig. 2 shows all the steps of the complete author-ship recognition system model. Its detailed description will be discussed in the following sessions. Initially, the texts are acquired and form a database which then goes through the preprocessing stage. Thus, the preprocessed data are divided into validation and testing, setting them for training and validation. Subsequently, the step of feature extraction is applied to the data. Afterwards, it goes through the training step, constituting the model for authorship recognition. For training, a set of 80% of the data was used, and, for validation, 20%. The data used both for training and validation are composed of several stylometry vectors.

4.2. AVASUS database

The AVASUS textbase covers some course completion articles and forum responses that students post virtually every day on the system. The number of authors, document length and a vast number of text entries prone to growth were the characteristics that influenced the selection of the database used. As there is a daily variation with several authors publishing, the tests were performed with the transcripts of the forums. The texts are followed by user IDs and arranged in a CSV (Comma-Separated Values) extension file named: foruns_do_AVASUS.csv. The AVASUS database has 50,939 unique user IDs who post on the forum. The data is divided into two columns: User ID and Text posted in the forum, a sample of the database is presented in Table 1.

An initial analysis of this database was performed using BoW combined with Term Frequency–Inverse Document Frequency (TFIDF) (da Rocha et al., 2019). The aim of this analysis was to verify the similarity of the writing patterns between the texts of the same author in each of the data settings considered. The motivation for this experiment was that it can be observed that within the datasets there may be clear writing preferences among the authors, which makes the recognition of authorship also influenced by the word classification set used by the authors.

The good efficiency acquired by the authors of the paper indicated a greater (dis) similarity among the authors in the data set. It was observed that the data sets were influenced by the (dis)similarity of word use among the authors. Still, similar patterns of word choice were observed in these tests — i.e., one author has different word usage preferences compared to the others. Therefore, the set of words used can help to evaluate the (dis)similarity between the authors and expose the possibility of authorship recognition.

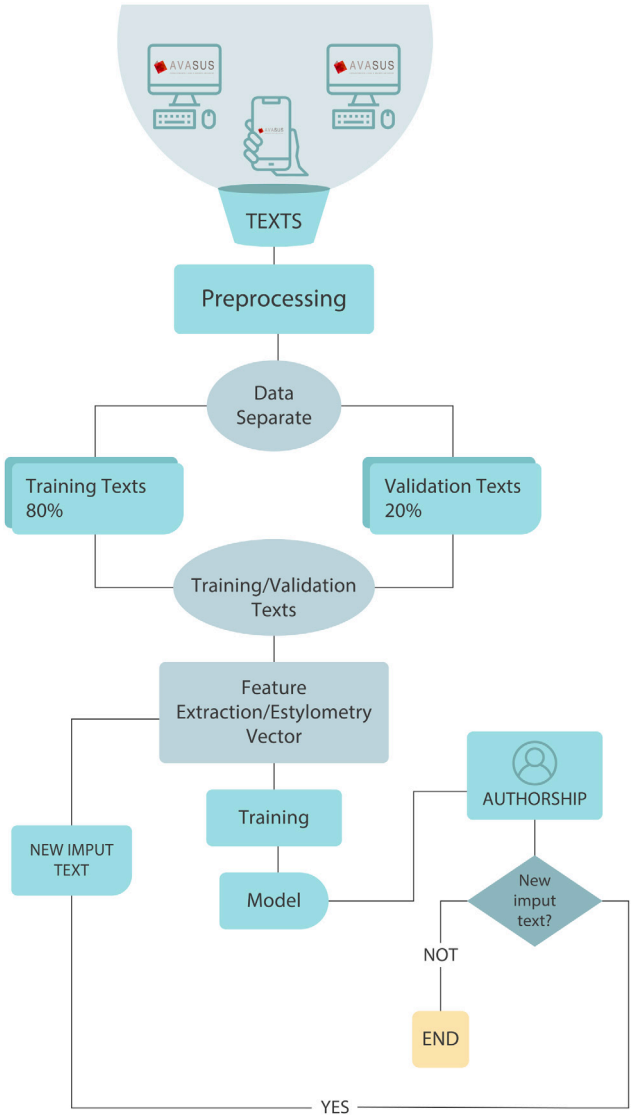

Fig. 2. Diagrammatic representation of the system.

Table 1
Sample of the original AVASUS database.

ID	Text
44765	Boa noite.colegas. a icterícia fisiológica ou ...
5224	É uma etapa muito importante e necessária para...
35443	Acho muito importante o tema da ictericia neon...
24148	PARA A PREVEN«AO DA INFEC«AO POR HPV...
109169	Boa noite na minha comunidade não utilizamos m...
2153	A promoçao e prevençao de saude da equipe na ...
43470	Evitar comportamentos sexuais de risco – ...
5207	»um fato que no Brasil, aproximadamente 200 m...
124517	muito interessante e emocionante...
44765	Por exemplo e' de muita importância alimentação...
2710	Bom dia turma concordo com vocês sobre o acomp...
9120	O importante e evitar a doença, ese e o unico ...
54254	Boa noite Profa. gostaria de saber qual é a atividade...
9120	E é bem assim, ao redor de um 60 por cento dos...
45287	Boa noite pelas grandes vantagems que tem o al...
8994	Como se reconhece que uma criança está com ictericia...
24069	Resposta: Mito. é um problema cultural que vem...

Table 2
Authorship recognition feature set.

Type	Size	Description
Word	2	Average length of tokens and short tokens
Character	2	Total digits and uppercase letters
Letters	26	Letters frequency
Digit	10	Digits frequency
Vocab Richness	2	Hápax (legomenon dyslegomenon)
Stopwords Type	220	Stopwords frequency
Punctuation	12	Punctuation occurrence

4.3. Preprocessing

The preparation and loading of AVASUS data were the first steps for the model. The objective was to organize the data for the preprocessing step performed later. In this preprocessing step for data standardization, null data, HTML tags, and whitespace were removed. Shortly afterwards, stopwords were removed. Most of these were functional words commonly used in writing, such as articles, prepositions, pronouns, and conjunctions that have no meaning/relevance within the text and appear at a high frequency within the corpus. Thus, they can be removed without affecting the context of what was written (Uysal & Gunal, 2014).

The importance of stopword may vary depending on the data set. In the case of authorship recognition Stopwords, it may indicate a writing preference for many authors. Some of them might make use of more or less words. Thus, in the feature analysis step, Stopwords are used to analyze their relevance.

4.4. Feature extraction

An extraction analysis was performed to determine the contribution of distinct characteristics for each test setting. Following a previous study (El & Kassou, 2014), the set of features was named as Stylometric Features. With the use of words, digits, and punctuation that capture an author's writing style, the default sets of 220 Portuguese Stopwords and 12 punctuation marks were applied. Table 2 has the feature sets separated into type, vector size and description:

The stylometry vector is formed by a set of extracted resources, as it will be analyzed in topic 5.2, and has a total length of 274. Fig. 3 shows the arrangement of the data in the stylometric vector that is used as an input to the classifier.

The detailed description of the keys in Fig. 3 are:

1. Average length of tokens in text is an attribute that returns a value for the sum of token length divided by the total number of tokens;
2. Average short token (or Total short tokens) returns the average total short tokens in each text, a short token has a maximum of 4 characters;
3. Numeric digits average (or Total digits) return the average of the total number of digits (0 to 9);
4. Uppercase Average returns uppercase letters entered by the author within a text;
5. Letter frequency returns how often letters are used in the text;
6. Digit frequency returns how often digit are used in the text;
7. Hápax legomenon and Hápax dyslegomenon are functions that return the average use of rare words that repeat once or twice, respectively;
8. Portuguese stopwords (or stopwords frequency) returns the average usage of each Stopword within the text;
9. Punctuation occurrence returns the average punctuation used within the text. Thus, the most frequent punctuation marks returned a higher average. This vector has a size of 12 punctuation marks (single quotation mark, colon, comma, underline, exclamation mark, question mark, semi-colon, period, double quotation mark, open parenthesis, close parenthesis, and dash).

Key:

#1 - Average length of tokens;
#2 - Average short tokens;
#3 - Numeric Digits Average;
#4 - Uppercase Average;
#5 - Letters frequency;
#6 - Digit frequency;
#7 - Hápax legomenon/dyslegomenon;
#8 - Portuguese Stopwords;
#9 - Punctuation Occurrence.

Fig. 3. Stylometry vector.

4.5. Training and classification

A reasonable amount of classification algorithms require a training step that is useful for preparing the machine and constantly improving its forecasting skills. This step often requires computers high operational performance and a long time to carry out, which makes the continuous occurrence of the step difficult. Hence it is fundamental that the classifier training only occurs when there is a need for new training. Therefore, for training, a set of 80% of the data was used and, for validation, a set of 20%. The data used for both training and validation are formed by various stylometry vectors.

The main algorithms chosen for this step were Support Vector Machine (SVM) and Logistic Regression (LR). The SVM classifier for its extensive application in the literature, as well as for its satisfactory results in text analysis. The LR algorithm is a machine learning technique that, despite its simplicity, works properly when it has many features. It was chosen since it models the probability of authorship, in which the authors are classified as known or unknown. The number "1" represents known and "0", unknown authors. Thus, the algorithm works well for the problem because the dependent variable is normally dichotomous. Lastly, both achieved better results compared to other algorithms, as it will be presented in Section 5. SVM acts faster and LR maximizes class probability, both have comparable performance when working with a large amount of text.

5. Evaluation and discussion of results

5.1. Features Evaluation

Features Evaluation was performed to explore which feature types are most important to the overall classifier. In such a way, statistical investigation of each feature extracted from the database is necessary. The choice of characteristics for the best ranking was made from this point of view, thereby identifying the important input resources. One of the datasets was selected to verify the features extracted through a BoxPlot, a method that graphically represents numerical data groups by their quartiles (Frigge, Hoaglin, & Iglewicz, 1989). Fig. 4 displays a BoxPlot, in which the distribution of some attributes extracted from the texts are exposed, such as the average length of tokens, average of short tokens, average of total digits, average of uppercase letters, and Hápax. These aspects return a unit value for vector assembly.

Figs. 5, 6, 7, and 8 show BoxPlots with the distribution of each feature vector that returns more than one value for mounting the stylometry vector. The characteristic extracted from the letter frequency, for example, returns a vector of size 26. Still, the constructed BoxPlot shows the distribution of each letter. The same goes for the other

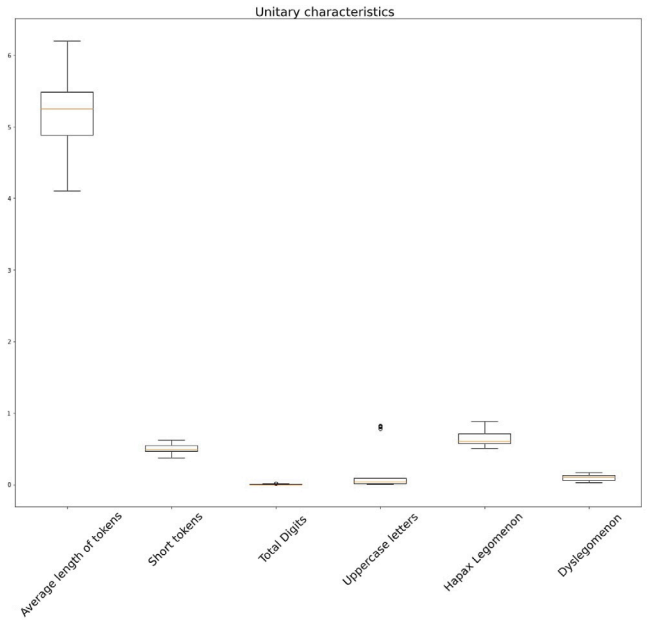

Fig. 4. Distribution of unitary characteristics Boxplot.

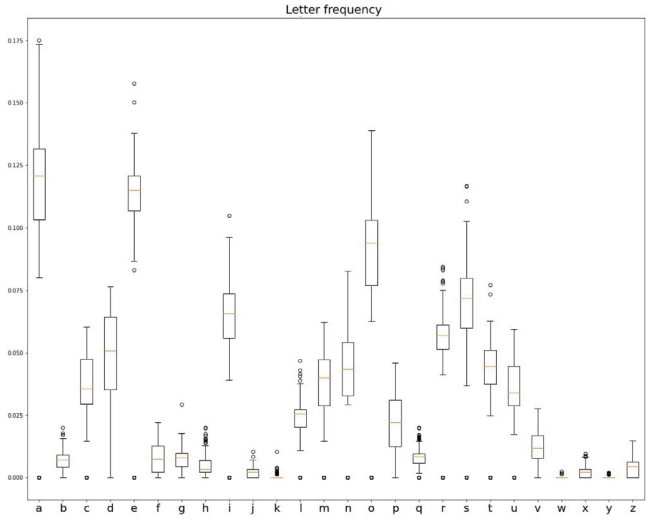

Fig. 5. Frequency distribution of letters Boxplot.

vectors. The frequency of digits shows the distribution of each, stop words display the distribution of each digit. In the box plot graph, 35 out of 220 are depicted as there was no relevant variation in the rest of the stop words, and the punctuation displays the distribution of the 12 punctuation marks. As it can be observed, the extracted data is relevant for classifier training.

While having many relevant features and making the training set robust, Stopwords did not have much variance. Such an aspect proved not to be useful in the experiments.

In BoxPlot of Fig. 4, it is observed that the total of digits had little variance, which may indicate a small amount of the use of numeric digits in the texts. In Fig. 5 it can be seen that there is a higher frequency of use of the letters: A, E, I, O, D and S and lower frequency in the letters: K, W and Y. Fig. 6 the digits from 5 to 9 there is a very large discrepancy, so this data turns out to be of little use, this same situation occurs in Figs. 7 and 8 with the punctuation marks, single quotation mark, underline, semi-colon and double quotation marks. Fig. 8 also shows that the most used punctuation marks in AVASUS

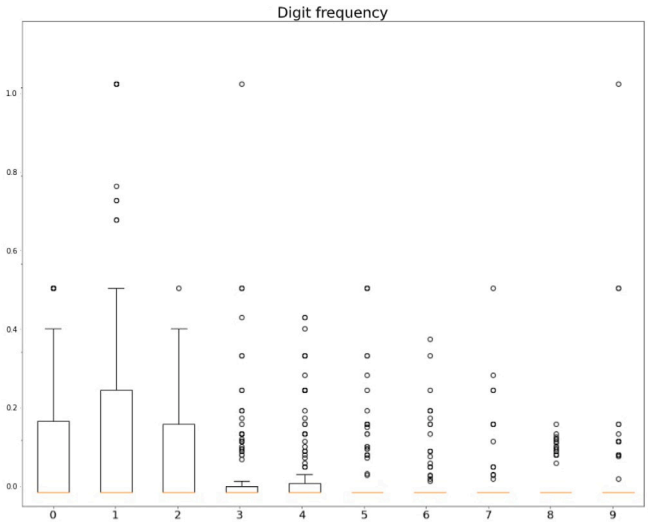

Fig. 6. Box plot of digit frequency distribution.

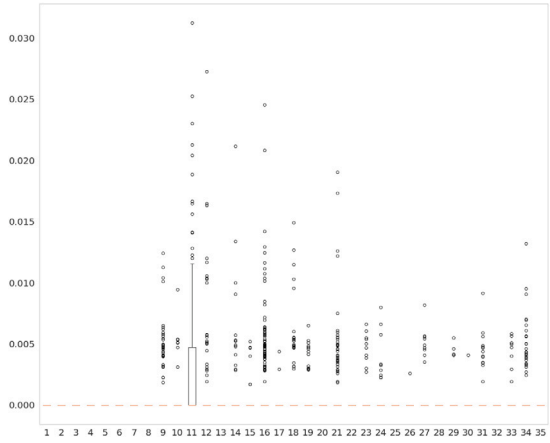

Fig. 7. Box plot of stop words frequency distribution.

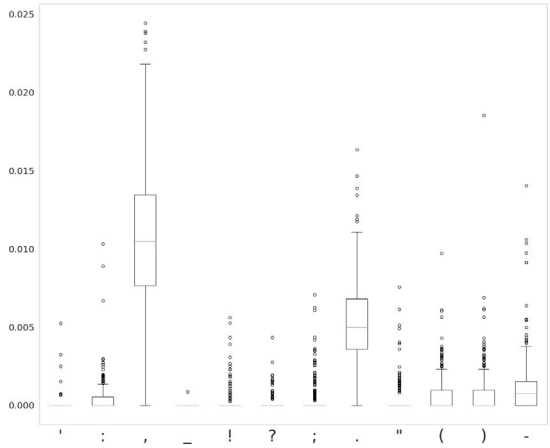

Fig. 8. Frequency distribution of punctuation marks box plot.

texts are comma and period. These characteristics are useful because some users use more or less punctuation marks, letters or numbers in their sentences, so the information contained in the entire text is considered.

Table 3
Database test sets.

Tests	Characters (n_c)	Authors	Publications (n_p)
1	$500 \leq n_c \leq 1000$	8	$20 \leq n_p \leq 35$
2	$1000 \leq n_c \leq 3000$	14	$10 \leq n_p \leq 15$
3	$100 \leq n_c \leq 500$	8	$150 \leq$
4	$100 \leq n_c \leq 300$	8	$50 \leq n_p \leq 60$
5	$100 \leq n_c \leq 300$	9	$70 \leq n_p \leq 200$
6	$30 \leq n_c \leq 60$	7	$100 \leq n_p \leq 250$
7	$30 \leq n_c \leq 60$	11	$25 \leq n_p \leq 60$
8	$300 \leq n_c \leq 700$	9	$28 \leq n_p \leq 40$
9	$400 \leq n_c \leq 800$	13	$20 \leq n_p \leq 30$
10	$30 \leq n_c \leq 50$	11	$19 \leq n_p \leq 37$
11	$700 \leq n_c \leq 2000$	15	$10 \leq n_p \leq 12$
12	$1000 \leq n_c \leq 2000$	8	$10 \leq n_p \leq 15$
13	$800 \leq n_c \leq 1500$	10	$10 \leq n_p \leq 25$
14	$50 \leq n_c \leq 90$	12	$14 \leq n_p \leq 25$

Table 4
Classification results with multiple settings - SVM.

C - Parameter/Kernel	Accuracy (%)
C = 1/Kernel = RBF	92
C = 5/Kernel = RBF	89
C = 10/Kernel = RBF	86
C = 50/Kernel = RBF	86
C = 1/Kernel = Linear	84
C = 5/Kernel = Linear	84
C = 10/Kernel = Linear	84
C = 50/Kernel = Linear	84

5.2. Classifier results

As a consequence of comparing and choosing the best classifier for the problem, classic classifiers were used to their broad acceptance in the literature (chapter 3 in Rocha, 2019) and also some classifiers with high authorship attribution rates such as LRand also some classifiers with high authorship attribution rates such as LR (Aborisade & Anwar, 2018), Gaussian Process Classification (GPC) and Bernoulli Naive Bayes (BNB). In Bernoulli's multivariate event model, resources are independent booleans (binary variables) that describe inputs. Like the Naive Bayes Multinomial model, Bernoulli's model is popular for document classification tasks (McCallum, Nigam, et al., 1998) and has proven to be more efficient than Naive Bayes Multinomial (see Table 5).

For the extraction of the style characteristics and recognition of authorship, it was employed a mixed database, composed of texts of teachers and students, forming the data sets of Table 3. These texts were published in the AVASUS forums and contained distinct sizes, requiring the organization of the database for training, so it became more consistent.

Considering the execution of the classifiers, the data was initially organized to extract the attributes, such as regulating the text size, number of authors, number of texts per author. Table 4 has the results of classifying test dataset 1 (see Table 3) with various SVM classifier specifications such as the C-parameter ranging from 1 to 50 and the kernel between linear or Radial Basis Function (RBF). It is possible to observe that the RBF kernel obtained better classification than the linear kernel and varying the C-parameter margin did not bring great advantages to the classifier, so for the other tests, the SVM configuration was C = 1 and kernel = RBF.

For the other classifiers used, changes were made to their default parameters and no major changes in the results were obtained. For KNN it resulted in a number of neighbors = 3, acquired by Elbow Curve (method to find the optimal value of K), for GPC the RBF kernel was used. The remaining classifiers were left with the default parameters of the sklearn library. The results are presented in Tables 5.

To also examine the results, a Receiver Operating Characteristic (ROC) curve was generated for the SVM classifier applied in the data set of tests 1 to 8. Fig. 9 shows the results of tests 1, 2, 3 and 4, where

Table 5
Classification results.

Tests	Classifiers/Accuracy (%)					
#	SVM	LR	NB	KNN	BNB	GCP
1	92	92	78	84	86	92
2	66	55	45	70	65	60
3	93	91	58	86	84	94
4	63	64	32	47	63	58
5	89	83	56	79	73	86
6	94	92	85	93	94	92
7	76	76	55	74	80	12
8	54	60	26	38	26	55
9	66	65	37	51	39	22
10	73	67	67	69	74	28
11	62	66	54	51	63	57
12	84	68	63	63	63	68
13	75	71	71	64	64	10
14	62	67	44	51	55	18

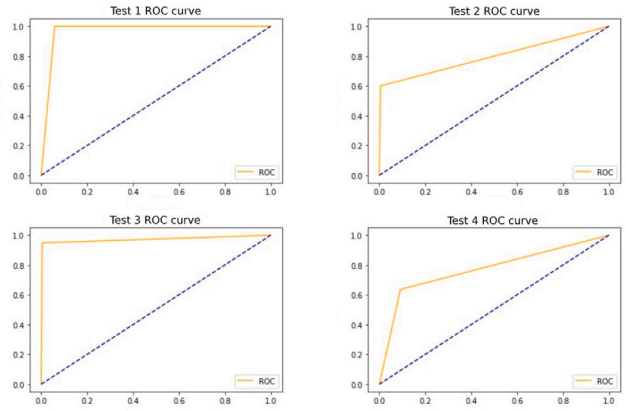

Fig. 9. ROC curves from tests 1 to 4.

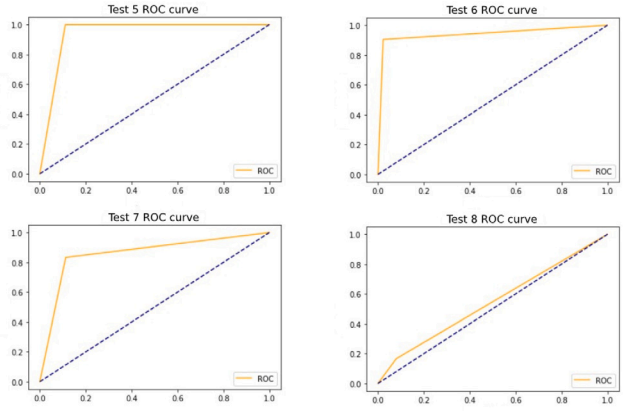

Fig. 10. ROC curves from tests 5 to 8.

it can be observed that Tests 1 and 3 have the best results because the curve is closer to 1.0 on the *y*-axis. Fig. 10 has the results of tests 5, 6, 7 and 8. Tests 5 and 6 obtained the best results and test 8 has an Area Under The Curve (AUC) close to 0.5 which means that the model has a low capacity for separation between classes (author/no-author). The use of the ROC curve provides a richer assessment of the classification model.

Based on the presented method utilized for the recognition of authorship, Table 5 shows the results obtained by the six classifiers of the fourteen test settings. In test 2, numerous characters and a few texts per author were sought, and the classification reached a median efficiency.

It is possible the number of texts per author was insufficient—the AVASUS Database no longer contained texts of this proportion with more than ten publications per author. This equally occurred in tests 11, 12, and 13; the latter obtained a more superior efficiency in some classifiers because it accepted a larger number of texts per author = 25. In tests 3 and 5, with a reasonable number of characters and more text samples per author, the efficiency achieved was high.

The reason behind this high precision may occur because these authors seem to follow a consistent style in their writings. This aspect makes the resources extracted from different texts close in their values and, consequently, makes the recognition more accurate. Stylometric features are most effective for data sets in which the authors demonstrate less similarity. Tests 6 and 7 contained only a few characters (30 to 60). Then, the results obtained in test 7 were good in five of the six classifiers. In test 6, they were good in all classifiers. This is because the GPC classifier does not classify a small number of features, and a larger number of features becomes slow in the training step. These results are promising because they indicate the proposed model has superior efficiency with small texts and, likewise, when it presents an ample number of texts per author.

5.3. Discussion

The AVASUS platform has over 590,000 (AVASUS, 2020) users enrolled. Daily, these users interact on the system through courses, completing the assignments provided and writing in the forums, which involves sharing doubts, activities and information with other colleagues and tutors. Accessing and analyzing this information allowed the following questions to be answered:

1. Of the total number of users, how many interact through the forums?
2. What is the total number of posts in the forums?
3. How many posts did each user publish? What is the total training/test volume?
4. How many posts are required to recognize authorship?
5. How many professionals would it take to recognize the authorship of these texts?
6. What is its the importance of SUS?

The answers to those questions are organized in the block below, wherein it is stated how many users posted in the forum, how many posts were published by each one, the total number of posts, the test volume in megabytes (Mb), the number of necessary posts to recognize authorship, the number of professionals in the field of education that would be necessary to analyze the texts for author recognition, and what is its substantiality for SUS.

Block:

- Number of users with forum posts = 50,939, database of 2019;
- 50,082 users published less than 10 posts;
- Total posts made by users = 255,023;
- Volume of texts that were used for training/testing = 56 Mb;
- Minimum number of posts for the algorithm to recognize authorship = 10;
- Number of professionals needed to recognize authorship = nearly 25,500 professionals.

The last item represents a hypothetical value considering a university professor of Portuguese who was observed for a few weeks with a class of 10 students and who knows how to recognize authorship using the same features as the algorithm. The paper (Silva, 2013) demonstrates it is possible for a teacher to detect each student's unique authorship marks. However, that is a challenging task to carry out. The recognition algorithm performs this task automatically and with no need of human intervention.

Table 6
Table of methods performance comparison.

Method	Authors	Texts	SVM	RL	KNN	NB
Bogdanova and Lazaridou (2014)	6	275	65	61	66	60
Otoom et al. (2014)	7	456	79, 3	–	–	84
Present study - Test 6	**7**	**1033**	**94**	**92**	**93**	**85**

Students participating in AVASUS courses must be responsible for their activities. In addition, the student who does not carry out his/her activity, who places another person to do so or does not interact and learn from the course might complete it and receive the certificate, with no one ever knowing whether that student indeed learned. Hence, with the recognition of authorship, it will not be possible for the student to place another person to interact for him/her, therefore increasing the probability of effectively learning and accomplishing the tasks by himself/herself. For SUS, this is extremely essential that students participate in the online course/training and effectively learn. Hence, the authorship recognition software is an automated way of enhancing the student's performance in activities without using a "ghostwriter" for that.

In order to verify the performance of the method applied in the paper, a search was carried out in journals that used the methods with a purpose similar to the previously described. Such a comparison is described in Table 6.

The method described in Bogdanova and Lazaridou (2014) proposes a new task of authorship attribution in several languages, in which the aim is to determine the author of a document written in the language Y, wherein it would differ from another. Further, it uses a series of stylistic cross-language characteristics for this task, such as those based on feelings and emotional markers (Pure HLF), on 275 texts divided for six authors, and on the LR, KNN, and NB classifiers. The method (Otoom, Abdallah, Hammad, Bsoul, & Abdallah, 2014) addresses the problem of authorship attribution in the Arabic texts and operates a recent set of hybrid characteristics that comprise: lexical, syntactic, structural and specific content characteristics for 456 texts and seven authors, using the NB, SVM, and other classifiers.

Concerning limitations, even if a subject has a fingerprint characteristic of writing due to their particular way of learning a language (Van Halteren, Baayen, Tweedie, Haverkort, & Neijt, 2005), the defining characteristics of this fingerprint are likely to be complex and not limited to just some features. Therefore, one possible solution would be to extend the method to employ other characteristics, such as character N-grams and word N-grams, applying them with the stylometry vector.

6. Conclusion

Technology-mediated education, especially those highly scalable, poses significant challenges. Regarding the Brazilian Unified Health System (SUS), these challenges are related to many practitioners and students in the health field, surpassing 3.5 million people. Specifically, in the case of the Virtual Learning Environment of the Unified Health System, the issue of scalability goes far beyond the context of the health workforce. The general population also uses this online learning environment.

For instance, when the Zika Virus epidemic surfaced, courses to guide pregnant women were offered, thus representing a virtual learning environment of high social demand. Either due to the need for training and qualification of healthcare professionals and students or due to the population's needs, which often seek knowledge related to prevention, self-care, or health services.

Since 2015, AVASUS has been employed by the Ministry of Health (MS) of Brazil and the Pan American Health Organization (PAHO) to educate professionals, students, and other social actors. What is more, this platform also played a significant role in Brazil during public health

crisis periods, such as the Zika Virus epidemic (from 2015 to 2016), when thousands of children were born with microcephaly, and during the syphilis epidemic (from 2018 to 2021). More recently, it revealed itself to be essential for disseminating efforts to combat COVID-19 (2020 to 2021). In those three moments alone, AVASUS reached approximately 400 thousand more enrollments in courses related to Zika virus, syphilis, and COVID-19.

Therefore, it entails a fundamental digital tool for human education in health for, in addition to training and qualifying people, playing a social role. Thus, given its scalability due to high demand, it is necessary to ensure levels of integrity, such as authorship. In this context, computational methods combined with natural language processing, such as those presented in this research, are essential tools for identifying authorship, especially for analyzing texts in Portuguese, a language for which works on the subject are scarce. Finally, it became evident that it is possible to obtain authorship recognition of a given platform user with the implementation of intelligent software.

CRediT authorship contribution statement

Marcella Andrade da Rocha: Collection, Organizing, Review of the literature, Preparing the manuscript, Results, Manuscript review and modification, Editing and revision. **Philippi Sedir Grilo de Morais:** Manuscript review and modification, Editing and revision. **Daniele Montenegro da Silva Barros:** Preparing the manuscript, Results, Manuscript review and modification, Editing and revision. **João Paulo Queiroz dos Santos:** Manuscript review and modification, Editing and revision. **Sara Dias-Trindade:** Manuscript review and modification, Editing and revision. **Ricardo Alexsandro de Medeiros Valentim:** Organizing and coordinating, Manuscript review and modification, Editing and revision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This authorship recognition work is being developed at Laboratory of Technological Innovation in Health (LAIS) and is part of a Doctoral Thesis developed by author Marcella Andrade da Rocha with the guidance of Ricardo Alexsandro de Medeiros Valentim. This research was developed through the "No Syphilis" Project with funding from the Brazilian Ministry of Health.

References

Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and Naive Bayes classifiers. In *2018 IEEE international conference on information reuse and integration* (pp. 269–276). IEEE.

Akimushkin, C., Amancio, D. R., & Oliveira, O. N., Jr. (2018). On the role of words in the network structure of texts: Application to authorship attribution. *Physica A: Statistical Mechanics and its Applications*, *495*, 49–58. http://dx.doi.org/10.1016/j.physa.2017.12.054, http://www.sciencedirect.com/science/article/pii/S0378437117312979.

Al-Ayyoub, M., Jararweh, Y., Rabab'ah, A., & Aldwairi, M. (2017). Feature extraction and selection for Arabic tweets authorship authentication. *Journal of Ambient Intelligence and Humanized Computing*, *8*(3), 383–393. http://dx.doi.org/10.1007/s12652-017-0452-1.

Albadarneh, J., Talafha, B., Al-Ayyoub, M., Zaqaibeh, B., Al-Smadi, M., Jararweh, Y., et al. (2015). Using big data analytics for authorship authentication of arabic tweets. In *Utility and cloud computing, 2015 IEEE/ACM 8th international conference on* (pp. 448–452). IEEE.

Altheneyan, A. S., & Menai, M. E. B. (2014). Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 473–484.

Araújo, K., Valentim, R., Lima, T., Moura, D., Jr., M. O., Almeida, A., et al. (0000). The production of open educational resources as an alternative for training volunteer health workers in rural communities in Tanzania, https://wcol2019.ie p. 39.

Ashraf, S., Iqbal, H. R., & Nawab, R. M. A. (2016). Cross-genre author profile prediction using stylometry-based approach. In *CLEF*.

AVASUS (2020). *Transparência – Dados de transparência: Technical report*, BR: AMBIENTE VIRTUAL DE APRENDIZAGEM DO SISTEMA UNICO DE SAÚDE - AVASUS, https://avasus.ufrn.br/local/avasplugin/dashboard/transparencia.php.

Bhatnagar, V., & Srinivasa, S. (2013). *Big data analytics: Second international conference, BDA 2013, Mysore, India, December 16–18, 2013, Proceedings: Vol. 8302*, Springer.

Bogdanova, D., & Lazaridou, A. (2014). Cross-language authorship attribution. In *LREC* (pp. 2015–2020). Citeseer.

Brocardo, M. L., Traore, I., & Woungang, I. (2015). Authorship verification of e-mail and tweet messages applied for continuous authentication. *Journal of Computer and System Sciences*, *81*(8), 1429–1440.

Cerra, D., Datcu, M., & Reinartz, P. (2014). Authorship analysis based on data compression. *Pattern Recognition Letters*, *42*, 79–84. http://dx.doi.org/10.1016/j.patrec.2014.01.019, http://www.sciencedirect.com/science/article/pii/S0167865514000336.

DATASUS (2020). *Base de dados do cadastro nacional de estabelecimentos de Saúde - Profissionais do SUS: Technical report*, BR: Ministério da Saúde, ftp://ftp.datasus.gov.br/cnes.

El, S. E. M., & Kassou, I. (2014). Authorship analysis studies: A survey. *International Journal of Computer Applications*, *86*(12).

Franco-Salvador, M., Rosso, P., & Montes-y Gómez, M. (2016). A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, *52*(4), 550–570.

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, *43*(1), 50–54.

Halteren, H. V. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *4*(1), 1:1–1:17. http://dx.doi.org/10.1145/1187415.1187416.

Huang, M., & Haralick, R. (2009). Identifying patterns in texts. *IEEE Internet Computing*, 59–64. http://dx.doi.org/10.1109/ICSC.2009.22.

Hürlimann, M., Weck, B., van den Berg, E., Suster, S., & Nissim, M. (2015). GLAD: Groningen lightweight authorship detection. In *CLEF (working notes)*.

Khomytska, I., & Teslyuk, V. (2018). Authorship and style attribution by statistical methods of style differentiation on the phonological level. In *Conference on computer science and information technologies* (pp. 105–118). Springer.

Markov, I., Gomez Adorno, H., Sidorov, G., & Gelbukh, A. (2016). Adapting cross-genre author profiling to language and corpus. In *CLEF (working notes)*. pp. 947–955.

Markov, I., Stamatatos, E., & Sidorov, G. (2017). Improving cross-topic authorship attribution: The role of pre-processing. In *Proceedings of the 18th international conference on computational linguistics and intelligent text processing*.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization: Vol. 752*, (pp. 41–48). Citeseer.

Neme, A., Pulido, J., Muñoz, A., Hernández, S., & Dey, T. (2015). Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing*, *147*, 147–159.

Nóbrega, G. Â. S. d., Souza, G. F. d., Barbosa, J. G., Coutinho, K. D., & Valentim, R. A. d. M. (2016). Uma análise estatística do Ambiente Virtual de Aprendizagem do Sistema Único de Saúde: Descrição estatística dos usuários do AVASUS. In *Uma análise estatística do ambiente virtual de aprendizagem do sistema único de saúde: Descrição estatística dos usuários do AVASUS*.

Otoom, A. F., Abdallah, E. E., Hammad, M., Bsoul, M., & Abdallah, A. E. (2014). An intelligent system for author attribution based on a hybrid feature set. *International Journal of Advanced Intelligence Paradigms*, *6*(4), 328–345.

Peng, J., Choo, R. K., & Ashman, H. (2016). Astroturfing detection in social media: Using binary n-Gram analysis for authorship attribution. In *2016 IEEE Trustcom/BigDataSE/ISPA* (pp. 121–128). http://dx.doi.org/10.1109/TrustCom.2016.0054.

Rao, O. S., Raju, N. G., Latha, Y. V., & Varma, P. V. (2017). Performance evaluation of unsupervised algorithms on morpheme based authorship clustering. *Performance Evaluation*, *4*(8).

Rocha, M. A. d. (2019). *Um texto tão singular quanto a impressão digital: O uso de sistemas inteligentes para reconhecimento de autoria* (Master's thesis), Brasil.

da Rocha, M. A., da Costa, R. D., de Medeiros Valentim, R. A., & de Pinho Dias, A. (2019). Um texto tão singular quanto a impressão digital: Reconhecimento de autoria com um olhar para o Avasus/A text as singular as digital printing: Author recognition with a look at avasus. *Brazilian Journal of Development*, *5*(12), 32960–32973.

Rocha, M. A. d., Nóbrega, G. Â. S. d., de Medeiros Valentim, R. A., & Alves, L. P. C. (2020). A text as unique as fingerprint: AVASUS text analysis and authorship recognition. In *Proceedings of the 10th Euro-American conference on telematics and information systems* (pp. 1–8).

Schmid, M. R., Iqbal, F., & Fung, B. C. M. (2015). E-mail authorship attribution using customized associative classification. *Digital Investigation*, *14*, S116–S126.

Schwartz, R., Tsur, O., Rappoport, A., & Koppel, M. (2013). Authorship attribution of micro-messages. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1880–1891).

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics*, *40*(2), 269–310.

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., & Nadali, S. (2013a). Detecting deceptive reviews using lexical and syntactic features. In *Intelligent systems design and applications, 2013 13th international conference on* (pp. 53–58). IEEE.

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., & Nadali, S. (2013b). Detecting deceptive reviews using lexical and syntactic features. In *2013 13th International conference on intellient systems design and applications* (pp. 53–58). http://dx.doi.org/10.1109/ISDA.2013.6920707.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853–860. http://dx.doi.org/10.1016/j.eswa.2013.08.015.

Silva, J. D. d. (2013). Escrita e subjetividade: As marcas da autoria no texto escolar.

da Silva, R. D., Pereira Filho, J. A., de Morais, P. S. G., de Medeiros Valentim, R. A., Coutinho, K. D., Roussanaly, A., et al. (2019). Data flow framework: A persona-based repository to modeling recommender systems. In *3rd annual learning & student analytics conference*.

Souza, B. R., Vieira, R. T., Coutinho, K. D., & Valentim, R. A. (2018). Avaliação sobre o Nível de Satisfação dos Usuários Inativos com a Plataforma AVASUS. In *Anais da VI Escola Regional de Computação Aplicada à Saúdes*.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, *62*, 2512–2527. http://dx.doi.org/10.1002/asi.21630.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, *21*, 421–439.

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., et al. (2014). Overview of the author identification task at PAN 2014. In *CEUR Workshop Proceedings: Vol. 1180*, (pp. 877–897).

SUS (2020). *Sistema Único de Saúde (SUS): Technical report*, BR: Ministério da Saúde, http://www.saude.gov.br/sistema-unico-de-saude.

Tschuggnall, M., & Specht, G. (2013). Using grammar-profiles to intrinsically expose plagiarism in text documents. In *International conference on application of natural language to information systems* (pp. 297–302). Springer.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112.

Valentim, R. A. d. M., Oliveira, A. C., Dias, A. d. P., Olieveira, E. d. S. G. d., Valentim, J. L. R. d. S., Moreira, J. A. M., et al. (2021). Educommunication as a strategy to face syphilis: An analysis of the open educational resources available at AVASUS. *DST-J Bras Doenças Sex Transm*, 1–5.

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, *12*(1), 65–77.

Vieira, G. V., de Freitas Neto, N., Coutinho, K. M. D., da Cunha Laranjeiras, L. A., de Medeiros Valentim, R. A., & Coutinho, K. D. (2017). Uma metodologia para otimizar o sistema de melhoria continuada do AVASUS com foco nas experiências do usuário. *Revista Brasileira de InovaçãO Tecnológica Em Saúde*, [ISSN: 2236-1103] *6*(3), http://dx.doi.org/10.18816/r-bits.v6i3.11129, https://periodicos.ufrn.br/reb/article/view/11129.

Marcella Andrade da Rocha. Ph.D. student at the Post-Graduate Program in Electrical and Computer Engineering (PPgEEC/UFRN), Master in Electrical Engineering and Computer and researcher at the Innovation Laboratory Health Technology (LAIS/UFRN).

Philippi Sedir Grilo de Morais. Ph.D. in Electrical and Computer Engineering at UFRN - Federal University of Rio Grande do Norte, with an emphasis on information processing and distributed systems. MSc. in Degree in Electrical and Computer Engineering at UFRN with a degree in Technology in Analysis and Systems Development at IFRN - Federal Institute for Education, Science and Technology of Rio Grande do Norte. Develops research activities within the UFRN, in Laboratory for Healthcare Innovation specifically in those fields: Mobile Devices, Ubiquitous and Pervasive Computing, Distributed Information Systems and machine learning.

Daniele Montenegro da Silva Barros. Ph.D. in the Graduate Program in Electrical and Computer Engineering from UFRN. Master in Electrical and Computer Engineering. Graduated in Technology in Analysis and System Development from IFRN (2012). Graduated in Social Sciences at the Federal University of Rio Grande do Norte (2007). Currently a researcher of the LAIS/UFRN.

João Paulo Queiroz dos Santos. Ph.D. in Electrical and Computer Engineering from the Federal University of Rio Grande do Norte (UFRN). Coordinator of the Advanced Technological Innovation Center (NAVI) at the Federal Institute of Education, Science and Technology of Rio Grande do Norte (IFRN), and researcher at the Laboratory for Technological Innovation in Health (LAIS) at UFRN. Has experience in Technological Innovation in Health, Algorithm Analysis, Machine Learning, Computing Complexity, Reinforcement Learning and Reactive Searches and Deep Learning.

Sara Marisa da Graça Dias do Carmo Trindade. Didactics by the University of Coimbra and post-doctorate in Educational Technologies and Communication by the same institution. Lecturer in the Department of History, European Studies, Archeology and Arts at the Faculty of Arts of the University of Coimbra. Integrated Researcher in the Digital Humanities Group and in the Center for Pedagogy Studies in Higher Education at the Center for Interdisciplinary Studies of the 20th Century (CEIS20-UC) of the University of Coimbra.

Ricardo Alexsandro de Medeiros Valentim. Ph.D. in Electrical and Computer Engineering from UFRN. Associate Professor at UFRN based in the Department of Biomedical Engineering, Permanent Professor at the Post-Graduate Program in Electrical and Computer Engineering (PPgEEC/UFRN) and Executive Director of the Laboratory for Technological Innovation in Health (LAIS) in Brazil at HUOL/EBSERH.