This journal provides immediate open access to its content under the Creative Commons BY 4.0 license. Authors who publish with this journal retain all copyrights and agree to the terms of the above-mentioned CC BY 4.0 license

DOI: 10.2478/seeur-2022-0100

A SURVEY ON AUTHORSHIP ANALYSIS TASKS AND TECHNIQUES

Arta Misini
Ph.D. Candidate, Faculty of Contemporary Sciences and Technologies,
South East European University, Tetovo, North Macedonia
am29747@seeu.edu.mk

Arbana Kadriu
Full professor, Faculty of Contemporary Sciences and Technologies,
South East European University, Tetovo, North Macedonia
a.kadriu@seeu.edu.mk

Ercan Canhasi
Faculty of Computer Science, University "Ukshin Hoti" Prizren, Kosovo
ercan.canhasi@uni-prizren.com

ABSTRACT

Authorship Analysis (AA) is a natural language processing field that examines the previous works of writers to identify the author of a text based on its features. Studies in authorship analysis include authorship identification, authorship profiling, and authorship verification. Due to its relevance, to many applications in this field attention has been paid. It is widely used in the attribution of historical literature. Other applications include legal linguistics, criminal law, forensic investigations, and computer forensics. This paper aims to provide an overview of the work done and the techniques applied in the authorship analysis domain. The examination of recent developments in this field is the principal focus. Many different criteria can be used to define a writer's style. This paper investigates stylometric features in different author-related tasks, including lexical, syntactic, semantic, structural, and content-specific ones. A lot of classification methods have been applied to authorship analysis tasks. We examine many research studies that use different machine learning and deep learning techniques. As a means of pointing the direction for future studies, we present the most relevant

methods recently proposed. The reviewed studies include documents of different types and different languages. In summary, due to the fact that each natural language has its own set of features, there is no standard technique generically applicable for solving the AA problem.

Keywords: Authorship Analysis, Text features, Stylometry, Classification, Authorship-related tasks.

INTRODUCTION

Using the author's previous works, experts in the Natural Language Processing (NLP) field may assign authorship to a text based on its characteristics. This field of study is known as Authorship Analysis (AA). In addition, it is associated with Information Retrieval (IR) since its primary concern is extracting information from people's writings. The AA task is the problem of figuring out who the real author of an anonymous text document is by looking at the textual features (El, Sara El Manar and Kassou, 2014).

Research involving authorship analysis comprises one of three tasks: authorship attribution or identification (Hoshiladevi Ramnial, 2016), (Varela et al., 2016), (A. S. Hossain et al., 2020) defines the author of a given work; authorship profiling or characterization (Khdr & Varol, 2019) determines the author's demographic traits (age, gender, or educational level); authorship verification or similarity detection (Brocardo et al., 2013), (Potha & Stamatatos, 2014) determines whether the presumed author was the one who wrote the examined document.

The new technological development has led to the emergence of several communication forms. Those forms have the advantage of anonymity, which makes this research much more meaningful. As a result, the objective of this paper is to explore the techniques used for determining the authorship of a specific document.

The essence of this task is the creation of a set of features that are generally constant across several texts written by the same author. Feature extraction is crucial in the authorship attribution task, where it plays a critical role. Most AA techniques use a set of stylometric features to identify the author of a piece of text (Lagutina et al., 2019). Classifiers employ these features to make inferences about the identity of the unknown author based on a given text. There are two main categories of AA strategies: profile-based and instance-based approaches. Profile-based approaches combine the author's training documents into a single text. The document is then analyzed to extract its stylometric characteristics. On the other hand, instance-based ones treat each document as its unique instance and include the documents individually in the training process (Hoshiladevi Ramnial, 2016), (A. S. Hossain et al., 2020). Stylometry

analysis is the basis for authorship attribution since it assumes that a fixed set of traits may sufficiently define the author's writing style. A writer's style is the set of linguistic habits consistently evident in their work. In this context, the study of unique writing habits and particular linguistic structures is referred to as stylometry. The writer's style may be captured by analyzing some measurable characteristics, such as the word frequency, the average length of sentences, or *n*-grams.

Research on literary works with contested authorship has played a significant role in the history of authorship analysis. The problem dates back to the statistical study of writing style. Shakespeare's plays and the Federalist Papers were the focus of the earliest studies in the 19th century. There are 85 writings in The Federalist Papers (Holmes & Forsyth, 1995), authored by Hamilton, Madison, and Jay. The collection of Shakespeare's plays (Lowe & Matthews, 1995) consists of 39 dramas. William Shakespeare was an English dramatist, poet, and actor. For a long time, historians have relied on authorship analysis to discover who wrote different literary works. The initial study aimed to identify authors of literary texts using stylometric techniques (Varela et al., 2016). Author attribution isn't just a literary problem (Phani et al., 2017), (Zhou et al., 2022), (Paci et al., 2011). The availability of text in digital formats like tweets (Sharon Belvisi et al., 2020), blogs, articles (Gupta et al., 2019), (Alsmearat et al., 2016), and other short texts (Abuhammad et al., 2021), (Khdr & Varol, 2019) has increased, making the AA a topic of open research in many fields. Authorship analysis is gaining more and more attention from researchers in recent years because of its relevance in several real-world domains, including but not limited to detecting plagiarism, cybercrime, law, online abuse, fake news (Canhasi et al., 2022), source code authorship, fraud detection, and so on. Author profiling has been used in many situations, including the identification of hate speech and the detection of cyberbullying. Due to the increase in cybercrime, authorship attribution techniques have been created to help forensic investigations (Sharon Belvisi et al., 2020).

The remaining sections of the paper are structured as follows. In Section 2, we take a look at the author-related tasks. The set of stylometric features utilized for various authorship analysis tasks is presented in Section 3. Section 4 gives an overview of the different methods used in the literature. The paper concludes briefly in Section 5.

AUTHORSHIP ANALYSIS SUBTASKS

Since authorship analysis is a decades-old topic, it has developed into three primary subtasks, including authorship profiling (Khdr & Varol, 2019), authorship verification (Brocardo et al., 2013), (Potha & Stamatatos, 2014), and authorship attribution (Hoshiladevi Ramnial, 2016), (Varela et al., 2016), (A. S. Hossain et al., 2020). Authorship attribution is an essential stylometry technique, but when accurate identification is impossible, authorship profiling helps restrict the search space by identifying variables like gender or age (Khdr & Varol, 2019). Authorship verification is different from authorship attribution in that it tries to establish whether or not the same author writes various documents. Authorship attribution, verification, and profiling are, in essence, similarity detection problems. They all aim to quantify the degree to which two papers are similar. In similarity detection tasks comparing text writing styles is a common goal.

Authorship Attribution

The process of *attributing authorship* (Hoshiladevi Ramnial, 2016), (Varela et al., 2016), (A. S. Hossain et al., 2020) focuses on determining the correct author of a contentious text. In researching the topic, identifying the author of a text has been categorized as a text classification issue (M. R. Hossain et al., 2021). By evaluating the document's writing style, authorship attribution techniques attempt to determine who of several writers is the author of the particular text. These methods operate under the premise that each writer has a distinctive pattern of word choice and sentence structure used to infer the writer of a given piece of text. Building a system for authorship attribution requires the incorporation of three components. The first component is a dataset, which has to have significant consideration given to it so that it can resonate with the author's characteristics. The data might be a collection of books, articles, or other forms of digital content. The second component is the processing of the author's text to extract stylometric features. Authorship attribution relies heavily on feature selection. The stylometric features (Lagutina et al., 2019) are the writing style markers that can help distinguish the documents from different authors. There are two methods used to extract features from the dataset. The first method, called profile-based (Potha & Stamatatos, 2014), combines the author's works into a single file and then analyzes it to extract features. The second is an instance-based method, which extracts the features by analyzing each author's document individually. Learning a classification model (A. S. Hossain et al., 2020) to determine the real author of an anonymous text is the final step in the attribution process. Studies that attempt to attribute authorship use a variety of stylistic traits and classifiers

(Hoshiladevi Ramnial, 2016), (Varela et al., 2016), which leads to a diversity of results. Authorship attribution has several practical uses, including document authentication, fighting plagiarism, resolving legal issues, and tracking an individual's identity.

Authorship Verification

Typically, *authorship verification* (Brocardo et al., 2013) is a binary classification issue that determines if two texts were created by the same author. In research, verifying whether or not two texts are the same author's work is framed as a similarity detection challenge. Several methods have been proposed for authorship verification (Koppel et al., 2009). One of them is one-class classification. In this scenario, the training process uses the instances from the contested author. After then, the model's output is compared to the test document. Because there is no concept of using negative samples during training in this scenario, the classification is a single class. By creating and using a group of pretenders, the verification problem may be reframed as an attribution problem, a strategy proposed by the many-candidates approach (Potha & Stamatatos, 2014). Unfortunately, the features determine the quality of such a group, and if we are not careful with the feature selection (Lagutina et al., 2019), the obtained results will not be accurate (El, Sara El Manar and Kassou, 2014).

Authorship Profiling

Document analysis via *authorship profiling* (Khdr & Varol, 2019) can provide demographic information such as age and gender (Alanazi, 2019) without explicitly revealing the author's identity. The goal is to infer a user's sociolinguistic details from an anonymous text using a set of predetermined classes. For a complete understanding of the author's sentiments, profiling may necessitate familiarity with the topics covered in the documents. It's possible that a more in-depth, content-specific understanding of typical themes might provide a more relevant, helpful identification of specific traits. Features (Lagutina et al., 2019) collected from the labeled data samples are profiled and examined separately for each demographic factor, as is standard practice for AA tasks (El, Sara El Manar and Kassou, 2014). The effectiveness of various classifiers or cluster models is evaluated with the use of machine learning approaches.

Extracted features and classification methods utilized for each of the aforementioned authorship-related tasks are detailed below.

FEATURE EXTRACTION

Many studies have shown that feature engineering is crucial to authorship analysis. People are habitual beings and have enduring personal characteristics. The writing style is an unconscious habit, which is different for each author. This is because each person has a way of using grammar, words, and punctuation.

Stylometric features

The study of style traits or features that can be measured statistically is known as **stylometry**. The study by (Lagutina et al., 2019) surveys the stylometric features in different authorship and NLP classification tasks including author identification, author verification, author profiling, style change detection, and text genre classification. Many different writing style features have been found in authorship-related studies. From this perspective, there are many ways to define a writer's style. A more detailed analysis shows that these style markers fall into four broad categories: lexical, syntactic, semantic, and content-specific. These feature categories can be seen in Figure 1.

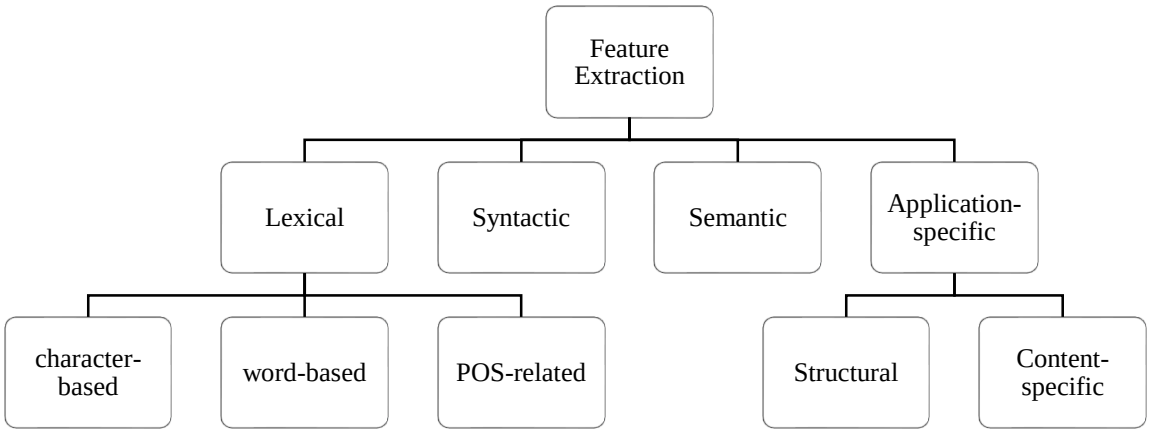


Figure 1. Features categories used in authorship analysis.

Lexical features

Lexical features help understand the characters and words that a writer commonly uses. The text can be analyzed at the character level, word level, and sentence level. These features include some basic measures which can be calculated for any text such as word length, character frequency, and the number of words per sentence. Khatun et al. (Khatun et al., 2019) claimed that character-level features (the number of characters, letters, the use of digits, and white spaces) are sufficient for authorship attribution. For gender identification, Alanazi

(Alanazi, 2019) used the frequency of punctuation marks since they can convey emotion in online texts.

N-gram is defined as a sequence of *n* elements (characters, words) in a text. Since the bag-of-words approach ignores the word order, word *n*-grams have been proposed in the literature as textual features. As they are language-independent and can capture contextual information are considered the most important stylistic features. Onan (Onan, 2018) has considered the *n*-gram-based representation for text genre classification. He used *n*-grams in the character and part-of-speech (POS) level. The *n*-gram approach is used for author attribution of short texts (Abuhammad et al., 2021), and literary texts (Phani et al., 2017). The definition of *n* is a crucial point in a successful AA task. In literature, researchers have determined different numbers of *n*.

The most common words in a text are called function words. Function words are considered important style markers since they are used unconsciously by authors and can capture significant stylistic patterns. The set of features used by Zhao et al. (Zobel, 2005) consists of function words only. They present function words as useful markers to identify the writing style of an author.

Syntactic features

Patterns used to construct sentences are what we call *syntactic features*. They include syntactic phrases, relations, dependencies, and writing rules. At this level, POS tagging is also used. Since authors tend to use similar writing patterns unconsciously, these features form an essential set of style-related tasks. Hriez et al. (Hriez & Awajan, 2020) have computed a list of syntactic features, including the number of noun phrases, verb phrases, and partial phrases. Stamatatos et al. (Stamatatos et al., 2001) used syntactic phrases as style markers for text genre classification, while Alsmearat (Alsmearat et al., 2017) and Alanazi (Alanazi, 2019) used them for gender identification. Also, syntactic rules have been used successfully in authorship identification and verification of literary texts (Varela et al., 2016).

Semantic features

Semantic features require more complex linguistic analysis on a deeper level. Due to the limitations of NLP technologies, semantic analysis can produce less accurate results. Consequently, there have not been many attempts to use high-level features for stylometric issues. Hriez et al. (Hriez & Awajan, 2020) used active and passive voice verbs as semantic

features to author classification. On the other hand, Hossain et al. (M. R. Hossain et al., 2021) used semantic similarity measures *on* word pairs.

Structural features

The way a writer organizes the structure of a text is expressed using *structural features*. These features include statistical measures such as the number of words per sentence and sentences per paragraph. One can also consider the maximum and minimum length of words and the *n*-gram frequency in the text. This set of structural features is studied (Hriez & Awajan, 2020). For online texts (email messages) can be helpful to consider the separation between paragraphs or the number of blank lines. Alsmearat et al. (Alsmearat et al., 2017) also considered the total number of short phrases or chunks in documents. This list of features is used for gender prediction.

Content-specific features

Content-specific features consider the particular keywords in the text. They can better capture the author's writing style within a certain text domain. Content-based information may discover some of the author's preferences. As a content-specific feature, the authors (Alsmearat et al., 2017) have experimented with the number of words in plural form. Also, the number of phrases from different topics is considered (Phani et al., 2017) in the author identification task. For text genre classification, the authors (Stamatatos et al., 2001) used content-specific keywords. The words conveying emotions are used also for gender identification (Alsmearat et al., 2016).

The following section describes the classification methods used for the various authorship-related tasks.

AUTHORSHIP DETECTION TECHNIQUES

Studying how to correctly identify authors is challenging, driving researchers to explore novel approaches to the problem. To define the writer's style, the researchers extracted several features from the text. Following this, they use a learning approach to determine the unknown authorship.

Several approaches have been experimented with to solve different authorship-related problems. In this way, the authors (Digamberrao & Prasad, 2018) have presented various techniques used in author identification in different languages. They concluded that no one approach is used exclusively for author identification; rather, researchers apply a variety of

techniques depending on the characteristics of the understudied language, the training data set, and the feature set. The classification was the technique used most frequently in the reviewed papers. So, we have oriented our study in this direction. The paper (Lagutina et al., 2019) reviewed different AA studies based on the document type.

This section discusses many authorship-related studies to provide a broad view of authorship analysis.

Machine-Learning classifiers

Because of the success of *machine learning* methods for automated classification, researchers have been able to focus on automatic attribution of authorship. The goal of classifiers based on the machine learning (ML) paradigm is to build classifiers to obtain a unique style function for each author.

Various ML classifiers have been used in different authorship analysis tasks. Alsmearat et al. (Alsmearat et al., 2017) have experimented with several combinations of classifiers. They used different classification algorithms for the gender identification problem. These algorithms include naïve Bayes (NB), decision trees (DT), support vector machines (SVM), and *k*-nearest neighbor (*k*-NN), among others. They achieve good results using SVMs and rule-based classifiers with stylometric features. In the same way, Khdr et al. (Khdr & Varol, 2019) have utilized NB, SVM, and DT (J48) for author profiling (gender and age prediction) in SMS messages. A higher accuracy has been achieved in gender prediction using the J48 decision tree. The authors (Maurya et al., 2016) have also used the SVM algorithm and *k*-NN with different features. These algorithms were used for author attribution. They also experimented with a different number of instances and different values of *k*. Also, in this paper, the SVM algorithm outperformed other methods. While regarding the value of *k*, 1-NN and 5-NN gave better results. A comparison with the best results reported in the literature is made by authors (Abuhammad et al., 2021). They have conducted several experiments using various combinations of features and data preprocessing methods. In tweets data, they applied the chunking technique of varying sizes and very high accuracy is achieved.

It has been popular in recent years to use a combination of style traits to train an ensemble of classifiers. It is an exciting area of study in machine learning. Ensemble learning integrates the results of individual classifiers to produce a powerful classification system. Onan (Onan, 2018) has applied this technique for genre classification. A diverse set of features and

different classifiers with ensemble learning methods has been used. He has conducted several experiments in two datasets and achieved high accuracy.

Deep-Learning techniques

Another method that has found application in authorship attribution is *deep learning*. Deep learning (DL) techniques have been getting a lot of attention lately due to their success in various fields. As end-to-end approaches, they do not require hand-crafted feature engineering, making them preferable to traditional methods. Nevertheless, it is worth noting that the availability of massive amounts of data to train these models is mainly responsible for their successful performance. When it comes to machine learning, deep learning belongs to a more extensive family of techniques that combine artificial neural networks (ANN) with representation learning.

Another class of ANNs for the author identification problem is used by (Gupta et al., 2019). They use two recurrent NNs, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). A comparison of the proposed word-index embeddings over pre-trained embeddings is made in two different datasets. A high accuracy has been achieved in one of the datasets using the GRU network with word-index embeddings.

Several experiments have been performed in the literature, evaluating and comparing ML and DL approaches. Alsmearat et al. (Alsmearat et al., 2017) have applied ANN (with several hidden layers, 500 epochs) and some other ML algorithms for gender detection in news articles (a large dataset). Due to hardware limitations, the ANN architecture has been evaluated in a smaller set of the corpus. In this work, the ML methods outperformed the ANN model. The authors (A. S. Hossain et al., 2020) utilized the SVM algorithm and MLP network (with one hidden layer) to create the authorship attribution model. They have used two activation functions for the MLP network, logistic and real. From the obtained results, the MLP with logistic activation function outperformed the SVM algorithm.

A different method for writer identification of short texts has been applied (Sharon Belvisi et al., 2020). They used three distance measures to compute the similarity between *n*-grams vectors. Using different values of *n* in character and word *n*-grams, they evaluated three distance measures in their tweets dataset. They conducted several experiments on the data using each of the features individually. The best results have been achieved in the idiosyncratic

feature set. The cosine and Manhattan distances have given high accuracy scores using character *n*-grams.

A novel framework for author classification of poetry text in Chinese has been proposed (Zhou et al., 2022). The framework (C-Transformer) consists of three components (CNNs, multi-head attention, and Transformer). Convolutional Neural Networks (CNNs) can learn contextual information. The transformer model can catch more detailed semantic information using multi-head attention. The LDA model is used to extract topical features from the poems. The experiments were performed in four datasets with a different number of poets using six baseline models. The best results were achieved using the proposed framework. The poetry topics improved the accuracy of the model.

Given the extensive review of the AA field from different aspects such as feature extraction, methods used, and their categorization, the next and final section of this paper presents the authors' discussions and their conclusions.

DISCUSSION

The techniques used to perform the different AA tasks are discussed in this paper. AA makes use of an enormous number of different features. This paper investigated several feature categories, including lexical, syntactic, semantic, structural, and content-specific features. There have been a variety of distinct feature combinations utilized for the various AA tasks. For genre text classification, lexical and syntactic features have been used. Topic modeling has also been proposed as a successful attribution method of literary works in some publications.

The researchers conducted a variety of experiments using numerous feature categories. Lexical-syntactic, lexical-structural-content-specific, and lexical-structural combinations are evaluated. The lexical-structural combination gave higher results. Also, analyzing the text in the context of words that convey emotion has shown promising results. Due to the complex linguistic analysis, semantic features have rarely been employed. However, some researchers used those in their studies with lexical, syntactic, and structural features.

A lot of classification methods have been applied to author-related tasks. In most cases, the SVM algorithm outperformed. It is also the most used classifier and has shown good performance in small datasets. Other classifiers have also shown good performance in some cases. For author gender prediction, a combination of SVM, NB, and *k*-NN classifiers have been used. Recently, deep learning methods have been the center of much discussion because

of their usefulness in many contexts. Different neural network architectures have been constructed for the author identification task. In the literature, a comparison between DL methods and ML algorithms has been made. It has been noticed that there is no explicit boundary among algorithms that have consistently produced superior performance. It depends significantly on the amount of training data.

A large number of experiments are reported in the literature. These include combinations of feature sets and classification algorithms. From the reviewed papers, the CNN network with fastText and GloVe representations has given promising results. Another model that performed well in the author identification task was the Cosine and Manhattan distance measures. This method employed the character *n*-grams and idiosyncratic features. In the text genre classification task, the transformer framework using *n*-grams and topic modeling gave a high accuracy score. The SVM algorithm, in combination with the BOW model and emotion features, gave the highest result in the gender detection problem. An ensemble of classifiers has been trained using a variety of style markers in recent years. It is a promising field of ML research. We can combine features and methods (ensemble) to produce good results. Transformers is a novel model recently proposed for the author attribution problem. As a DL model, a transformer uses the self-attention mechanism to give distinct parts of the input data different levels of importance. The attention module gives the transformer its abilities. It is because it records the connections between each word in a sequence of words.

Some researchers in the English language have utilized topic models to attribute authorship. Topic modeling outperformed standard approaches to authorship attribution in these research studies. Emotion analysis and psycholinguistic features have been successful in author gender detection. The application of deep learning methods is seen as a potential new direction for study in this field, as they can provide high accuracy and consistency when applied to enormous corpora. Transformers-based models have been proposed recently for the author attribution problem. Language models may be used to achieve more complex levels of transformer-based systems.

CONCLUSION

The paper provided an overview of the studies and techniques used in the authorship analysis field. We considered a variety of features and methods that have been proposed in AA. Different forms of written text and languages are included in this paper. Each technique is analyzed separately within the context of the feature set and classification methods used.

Feature extraction is a crucial aspect of determining an author's writing style. The language of the text plays a significant role in the selected features. The volume of training data is a critical factor in classifying authors by their writing style. Much research has demonstrated that factors such as dataset size, document length, and the number of potential authors impact the output of an authorship attribution approach. The feature set that best represents the author's writing style is an essential factor for good performance if an ML algorithm is used. On the other hand, a high level of accuracy in authorship attribution using DL techniques may be possible, but only if a significant amount of training data is available.

It is also worth noting that the experiments that used more data and fewer candidate authors gave more satisfactory results. It is unquestionably a result of the extensive training set and the model's ability to learn from more samples. In conclusion, the unique characteristics of each natural language mean that no one approach can be applied universally to the AA problem.

REFERENCES

1. Abuhammad, Y., Addabe, Y., Ayyad, N., & Yahya, A. (2021). Authorship attribution of modern standard Arabic short texts. *ACM International Conference Proceeding Series*, *1*(1), 1–7. https://doi.org/10.1145/3485557.3485563
2. Alanazi, S. A. (2019). Toward identifying features for automatic gender detection: A corpus creation and analysis. *IEEE Access*, *7*(Ii), 111931–111943. https://doi.org/10.1109/ACCESS.2019.2932026
3. Alsmearat, K., Al-Ayyoub, M., Al-Shalabi, R., & Kanaan, G. (2017). Author gender identification from Arabic text. *Journal of Information Security and Applications*, *35*, 85–95. https://doi.org/10.1016/j.jisa.2017.06.003
4. Alsmearat, K., Shehab, M., Al-Ayyoub, M., Al-Shalabi, R., & Kanaan, G. (2016). Emotion analysis of Arabic articles and its impact on identifying the author's gender. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2016-July* (November). https://doi.org/10.1109/AICCSA.2015.7507196
5. Brocardo, M. L., Traore, I., Saad, S., & Woungang, I. (2013). Authorship verification for short messages using stylometry. *2013 International Conference on Computer, Information and Telecommunication Systems, CITS 2013*. https://doi.org/10.1109/CITS.2013.6705711
6. Canhasi, E., Shijaku, R., & Berisha, E. (2022). Albanian fake news detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *February*, 0–24. https://doi.org/10.1145/3487288
7. Digamberrao, K. S., & Prasad, R. S. (2018). Author Identification on Literature in Different Languages: A Systematic Survey. *2018 International Conference on Advances in Communication and Computing Technology, ICACCT 2018*, 174–181. https://doi.org/10.1109/ICACCT.2018.8529635
8. El, Sara El Manar and Kassou, I. (2014). Authorship Analysis Studies: A Survey. *International Journal of Computer Applications*, *86*(12), 22–29. https://doi.org/10.5120/15038-3384
9. Gupta, S. T. P., Sahoo, J. K., & Roul, R. K. (2019). Authorship identification using recurrent neural networks. *ACM International Conference Proceeding Series*, 133–137. https://doi.org/10.1145/3325917.3325935
10. Holmes, D. I., & Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, *10*(2), 111–127. https://doi.org/10.1093/llc/10.2.111
11. Hoshiladevi Ramnial, S. P. and S. P. (2016). Authorship Attribution Using Stylometry and Machine Learning Techniques. *Advances in Intelligent Systems and Computing*, *384*(August), 247–257. https://doi.org/10.1007/978-3-319-23036-8
12. Hossain, A. S., Akter, N., & Saiful Islam, M. D. (2020). A stylometric approach for author attribution system using neural network and machine learning classifiers. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3377049.3377079
13. Hossain, M. R., Hoque, M. M., Dewan, M. A. A., Siddique, N., Islam, M. N., & Sarker, I. H. (2021). Authorship classification in a resource constraint language using convolutional neural networks. *IEEE Access*, *9*, 100319–100338. https://doi.org/10.1109/ACCESS.2021.3095967
14. Hriez, S., & Awajan, A. (2020). Authorship Identification for Arabic Texts Using Logistic Model Tree Classification. In *Advances in Intelligent Systems and Computing: Vol. 1229 AISC*. Springer International Publishing. https://doi.org/10.1007/978-3-030-52246-9_48
15. Khatun, A., Rahman, A., Islam, M. S., & Marium-E-Jannat. (2019). Authorship

attribution in bangla literature using character-level CNN. *2019 22nd International Conference on Computer and Information Technology, ICCIT 2019*, 18–20. https://doi.org/10.1109/ICCIT48885.2019.9038560
16. Khdr, A. J., & Varol, C. (2019). Age and Gender Identification by SMS Text Messages. *2018 International Conference on Artificial Intelligence and Data Processing, IDAP 2018, July 2021*, 1–5. https://doi.org/10.1109/IDAP.2018.8620780
17. Koppel, M., Schlier, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9–26. https://doi.org/10.1002/asi.20961
18. Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., & Demidov, P. G. (2019). A Survey on Stylometric Text Features. *Conference of Open Innovation Association, FRUCT*, 184–195. https://doi.org/10.23919/FRUCT48121.2019.8981504
19. Lowe, D., & Matthews, R. (1995). Shakespeare Vs. Fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, *29*(6), 449–461. https://doi.org/10.1007/BF01829876
20. Maurya, R. K., Saxena, M. R., & Akhil, N. (2016). Authorship Attribution Using Stylometry and Machine Learning Techniques. *Advances in Intelligent Systems and Computing*, *384*(January 2016), 247–257. https://doi.org/10.1007/978-3-319-23036-8
21. Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28–47. https://doi.org/10.1177/0165551516677911
22. Paci, H., Kajo, E., Trandafili, E., Tafa, I., & Salillari, D. (2011). Author identification in Albanian language. *Proceedings - 2011 International Conference on Network-Based Information Systems, NBiS 2011*, 425–430. https://doi.org/10.1109/NBiS.2011.71
23. Phani, S., Lahiri, S., & Biswas, A. (2017). A supervised learning approach for Authorship attribution of Bengali literary texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *16*(4), 15. https://doi.org/10.1145/3099473
24. Potha, N., & Stamatatos, E. (2014). A profile-based method for authorship verification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8445 LNCS*, 313–326. https://doi.org/10.1007/978-3-319-07064-3_25
25. Sharon Belvisi, N. M., Muhammad, N., & Alonso-Fernandez, F. (2020). Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features. *2020 8th International Workshop on Biometrics and Forensics, IWBF 2020 - Proceedings*, 1–6. https://doi.org/10.1109/IWBF49977.2020.9107953
26. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Automatic text categorization in Terms of Genre and Author. *Computational Linguistics*, *26*(4), 471–495. http://casa.disi.unitn.it/~moschitt/books/Book2005-reduced.pdf
27. Varela, P., Justino, E., Britto, A., & Bortolozzi, F. (2016). A computational approach for authorship attribution of literary texts using sintatic features. *Proceedings of the International Joint Conference on Neural Networks*, *2016-Octob*(May 2021), 4835–4842. https://doi.org/10.1109/IJCNN.2016.7727835
28. Zhou, A., Zhang, Y., & Lu, M. (2022). C-Transformer Model in Chinese Poetry Authorship Attribution. *International Journal of Innovative Computing, Information and Control*, *18*(3), 901–916. https://doi.org/10.24507/ijicic.18.03.901
29. Zobel, Y. Z. and J. (2005). Effective and Scalable Authorship Attribution Using Function Words BT - AIRS 2005, LNCS 3689. *Airs 2005, Lncs 3689*, 174–189. papers2://publication/uuid/97C99154-6116-4BBD-85CC-3BE690928DF5