Identifying fakes in social media text messages using Stylometry

1st Cristian G. Butzke
BCC/ IFC
Rio do Sul, Brazil
cristian.btzk@gmail.com

2nd Marcelo L. Brocardo
ESAG/ UDESC
Florianópolis, Brazil
marcelo.brocardo@udesc.br

3nd Wesley R. Bezerra
BCC/ IFC
Rio do Sul, Brazil
wesley.bezerra@ifc.edu.br

4rd Carlos B. Westphall
PPGCC/ UFSC
Florianópolis, Brazil
carlosbwesphall@gmail.com

***Abstract*—The authentication and authenticity of artifacts have been subverted for unsavory purposes, such as creating fake texts and other artifacts and trying to impersonate a user. Authentication has also been affected by the impersonation of some aspects of biometrics, potentially compromising the identification of subjects. In this perspective, this work proposes a continuous authentication solution that uses textual features to identify stylometric aspects using a Support Vector Machine to identify the authenticity and authorship of texts from the X/Twitter social network. As a result, the results obtained through different kernel configurations for different users are presented. Additionally, a simulator of social media publications infers the authenticity and authorship of the published text based on previously trained machine learning.**

***Index Terms*—stylometric, social media, authentication, security**

I. Introduction

Attesting the authenticity of texts on the *internet*, especially on social media, has become a challenge in recent times with Generative Artificial Intelligence (GenAI) [1]. This has led to authentication problems in electronic systems [2], deepfakes [3], [4], and other vulnerabilities resulting from the impersonation of individuals through AI [5], [6]. Problems related to access control and data vulnerability, in addition to the risks associated with their misuse, have led to the search for new protection mechanisms against improper access [7].

In this regard, one of the main challenges in application security is identifying whether an individual performing a specific action in the system is who they claim to be. The validation of a subject's identity is carried out through a process known as authentication, the most widespread means in modern systems being password-based authentication. However, for Kaur, Kumar, and Singh [8], several cases of hacking of user accounts show that it is possible to circumvent this system, regardless of how secure it is considered.

Verifying the authorship of a message involves several steps. It presents challenges in security given the introduction of Generative Artificial Intelligence; see Figure 1. In particular, the security of electronic devices and social media is an important research issue [9], [10]. Ensuring that social media authentication is not violated is not always possible, and this is due to several factors. However, the identification of authorship can be confirmed a posteriori and helps to identify fakes or frauds that may occur after security breaches in social media accounts or devices.

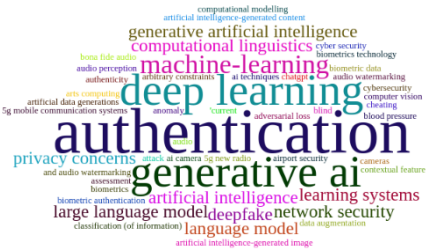

Fig. 1. Word cloud. Constructed through bibliometric study carried out by the authors in late 2024

Authentication of the subject in digital media can be a challenge. Techniques based on physical attributes sometimes diverge from continuous authentication, as they require user actions, such as pressing a biometric reader, speaking continuously, or positioning themselves in front of a camera [11]. Therefore, pattern recognition related to behavioral biometrics, on the other hand, uses actions a user performs, such as identifying walking patterns, touching the screen, analyzing writing style or eye movement to verify the user's authenticity continuously [7]. Methods based on these patterns have achieved high levels of accuracy and become more complex to transgress due to their principle being based on the use of the application.

We can infer that stylometry is an adequate solution for authenticating subjects, especially on social media. Brocardo [7] highlights an individual tendency in each user's writing style so that the extracted patterns can be used to verify the authenticity of a text. In this way, applying stylometry to a set of user messages proves to be an effective alternative for detecting account intrusions during the application's use period.

Thus, this work addresses the steps from data extraction, transformation, and loading to creating an inference model for the authorship of a text obtained from social networks, specifically the X/Twitter network. This study presents the development of a continuous authentication model based on publications by users of the social network X to prevent impersonation. It brings the following contributions:

- Propose a solution for identifying text authorship using stylometry;
- Evaluate the proposed solution against other existing solutions.

To define the scope, it is essential to emphasize that this work examines the proposal's capacity to identify a subject through stylometrics. For this purpose, a dataset with posts from eight different users was used. However, it is essential to note that the proposal identifies only the user for whom it was trained; that is, it assesses whether the user who posted the evaluated text is the same as the one for whom it was trained. Therefore, the result is independent of the number of subjects other than the one evaluated, since all others are considered a negative result in the identification.

The remainder of the work is organized as follows: Section II presents the related works, the methodology adopted in developing these articles, and a description of the experiment. Following Section III, a description of the proposal's architecture and components is provided. Section IV presents a discussion of the results obtained and evaluates the proposal. Finally, in Section V, the conclusions and possible future work are presented.

II. Related Works and Methodology

A. Related Works

This section aims to report research and works published in portals, newspapers, or scientific journals whose research object is related to the themes of continuous authentication and identification of text authorship. The works were selected after the bibliographic review, with the identification of the main topics and techniques relevant to the theme.

In their research, **Kaur, Kumar, and Singh** [8] propose a study on using textual analysis as a basis for continuous authentication in social networks. The authors applied stylometry techniques to publications by several users on the Twitter platform, and the techniques used in this stage were t-score, Gini-index, Fischer score, and Correlation Feature Selection. The authors compared the performance of the KNN, Random Forest, Gradient Boosting, SVM, and Multilayer Perceptron algorithms. Based on the test results, the highest performance obtained by the SVM algorithm is notable.

The study by **Alterkvi and Erbay** [12] proposes a model for verifying the authorship of messages based on stylometric characteristics. Their model uses a database of publications on the social network Twitter, with pre-processing through the XG Boost algorithm to classify the publications selected using the Multi-Criteria Decision Method (MCDM). Four classification algorithms—Logistic Regression, SVM, Random Forest, and XGBoost—were evaluated. The logistic regression algorithm achieved the best result.

The article by **Litvak** [13] addresses the problem of authorship verification in smaller email messages. The author used the Enron email dataset, from which 52,000 emails from 52 different users were used. The author uses a Convolutional Neural Network model with a Seq2Seq model for natural language processing. The neural network developed has several convolution and pooling layers, using the max-overtime pooling algorithm The average accuracy reached 97% , a considerable metric compared to other works that used the same dataset.

The work proposed by **Castillo, Cervantes, and Vilariño** [14] aims to explore the problem of verifying the authorship of a text through a graph representation to extract relevant linguistic features. The database consists of English texts whose content addresses different topics. The authors' proposal is based on extracting textual features through centrality measures in texts represented through graphs, which are then applied to the SVM learning algorithm to build a classification model. The authors report an improvement in the score obtained by the model compared to other works.

Yang [15] presents an analysis of social media texts for the purpose of predicting age and gender. His work uses information from Reddit and utilizes the TextFooler attack for evaluation. For prediction, the BERT model is adopted for gender and age analysis. The TextFooler attack is then applied. As a result, the gender and age accuracies drop from 68% and 76% to 31.9% and 38.9%, respectively.

In their work, the authors **Xu and Fung** [16] present the possibility of cross-platform social media identification, User Identity Link (UIL). Thus, the same author can be identified across different social media platforms through stylometry, even if they are registered under a pseudonym. For this identification, a solution proposed by the authors, based on a Graph Neural Network (GNN), specifically StyleLink, is used. As a result, StyleLink performed better than the other approaches in terms of accuracy.

Yang et. al [17] uses stylometry to identify fake news. The work extracts text features and submits them to classifiers for the identification of fake news. This work evaluates the SVM, Random Forest, AdaBoost, and XGBoost approaches. The proposed model categorizes the results as propaganda, hoax, satire, and trustworthy, with XGBoost performing the best among the analyzed approaches.

Table I presents two works with the same application context as this work. This demonstrates the relevance of applying machine learning techniques for authentication and authorship identification in social media. Another point to be noted is the use of SVM in two works and centrality measures in another, leaving only one approach with neural networks. Thus, we can infer that stochastic approaches are more suitable for solving this problem.

Although different authors have addressed identifying text authorship (see Table I), this work brings specific advances for stylometry applied to social media. Since its scope is focused on continuous authentication, an approach that allows continuous data acquisition and a well-defined flow of preparation, training, and testing makes this solution suitable for the proposed problem.

In summary, the related works reinforce the choices made in this project. SVM is a suitable choice for this project compared to other authors, as listed in the related articles. Although the

Work	Context	Technique	Result
[8]	Twitter	KNN, Random Forest, Gradient Boosting, SVM and MLP	SVM had the best performance
[12]	Twitter	Logistic Regression, SVM, Random Forest and XGBoost	Logistic Regression had the best performance
[13]	E-mails	Convolutional Neural Networks and Seq2Seq	97% accuracy
[14]	English texts	SVM	improvement in results compared to other works
[15]	Reddit	BERT model and TextFooler	accuracy decrease
[16]	X and Foursquare	GNN/StyleLink	better performance
[17]	text	SVM, Random Forest, AdaBoost and XGBoost	quatro classes, XGBoost melhor desempenho
this work	X	SVM	Fake user identification

TABLE I
COMPARISON BETWEEN RELATED WORKS.

accuracy of the convolutional neural network (CNN) was very good (97%), this approach was not used for social media texts. A stochastic approach better supports uncertainty.

B. Methodology

This section will present how the proposal was carried out, the creation of the project, and its evaluation. The necessary steps for its development and the technologies used will be described. As seen in Figure 2, the process occurs in four steps, with the third step being divided into two substeps.

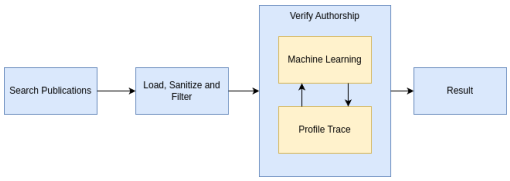

Fig. 2. Methodology. The steps are shown in blue, and the sub-steps in orange. It can be seen that the iteration between the sub-steps is continuous; that is, they feed back on each other.

Regarding the **search publications**, the publications used in the following steps of the project for feature extraction will be searched for in freely available datasets and processed later. The selected data will be **loaded, sanitized, and filtered** using the resources available through the Python language.

Relevant textual features were extracted for classification to identify a user's writing profile, such as the number of punctuations, sentences, blank spaces, sentences starting with capital letters, emojis, and hashtags. Authorship verification will use a machine learning model to classify whether a text's writing pattern corresponds to the authenticated subject's profile.

A dataset containing publications from eight accounts of known authors on the X platform was used to carry out this project. The set was obtained through the Kaggle platform [1], intended for user collaboration and data provision. The total set consists of 18,680 data from eight different authors. The most minor and most significant number of publications extracted for a user were 2,032 and 2,585, respectively.

This project follows the principles of Open Science [18]–[20] using the Open Science Framework (OSF) platform [21]. Its data and technical files can be accessed through the link[2].

[1]https://www.kaggle.com/
[2]https://osf.io/dpg7t/

C. Experiment Description

The development, training, and testing of the models were carried out on a computer with a Ryzen 7 5500x processor, a GeForce RTX 3060 video card with 12 GB of VRAM, 32 GB of DDR4 RAM with a frequency of 3200 MHz and the Ubuntu 22.04.5 LTS operating system. The coding was done in a development environment with the Visual Studio Code text editor and Python language version 3.10.12.

An important phase of the experiment is preparing the data to be used in machine learning models. This step was done using the Python programming language. The Pandas library was used to load the data, the NumPy library was used to resize vectors, and the NLTK library was used for tokenization and removal of *stopwords*. Subsequent steps, such as training and testing, used a specific library, Scikit-Learn, for machine learning and deep learning [22].

The training of a classification model based on Support Vector Machines can receive different configurations with distinct parameters (tuning). The parameters whose values were varied in the training stage were the kernel type (1), the regularization value C (3) and the Gamma value (2), with the configurations being changed according to their respective equations.

$$Kernels = \{Linear, RBF, PolinomialeSigmoide\} \quad (1)$$

$$Gamma = \{0.1, 1, 10, 100\} \quad (2)$$

$$C = \{0.1, 1, 10, 100, 1000\} \quad (3)$$

After verifying the best algorithm configuration for the current scenario, given by $kernel = linear, c = 1000$ (without gamma), a result of $accuracy = 100\%, precision = 100\%$ was obtained. This configuration had the best result, while $kernel = sigmoid, c = 0.1$ without gamma had the worst result ($accuracy = 70\%, precision = 70\%$). As a result, one can observe the importance of a prior evaluation of the algorithms' parameterization and how these directly influence the result obtained.

III. The proposal for continuous authentication in social media

This proposal is based on three main concepts for providing continuous stylometry authentication. The first concept is Natural Language Processing (NLP) and its techniques. Second, Support Vector Machine brings a machine learning approach to supervised models. Thirdly, stylometry is used as a factor in authorship and subject identification.

NLP techniques aim to contribute to the analysis, modeling, and understanding of language used by human beings. NLP is currently used in several everyday applications. Some examples are email platforms, voice-based assistants, search engines, and translations [23]. This research area has been widely used for command interpretation and iteration with Generative AI tools [24], [25].

For [23], extracting information from a text involves applying a text processing pipeline with procedures to solve an NLP problem. Since a language is a set of finite sentences constructed with a finite number of symbols [26], its understanding, even if complex, can be summarized in a set of steps, among them tokenization.

Tokenization is natural language processing that usually occurs in sentences and often requires minimally dividing the text into words. Good preprocessing practices involve removing digits, memorable characters, and punctuation and transforming the text into upper- or lower-case letters.

For Jurafsky and Martin [27], sentence segmentation constitutes one of the initial phases of text preprocessing. As a rule, sentences can be divided when a full stop or other punctuation marks are found. However, it is worth noting that some cases, such as the abbreviation ”dr.”, violate the rule, demonstrating the possibility of ambiguity due to some punctuation marks [23].

Another important concept is the SVM, a specific type of support vector classifier [28]. This, in turn, is part of the set of approaches brought by machine learning, identifies patterns and produces an approximate explanation for some phenomenon or event [29] and can assist in the construction of knowledge [30].

Support Vector Machines are supervised models capable of performing linear and nonlinear classifications, regression, and outlier detection [30] that use maximum margin classifiers [28]. However, it is not always possible to use this type of classifier. However, a possible solution is increasing the dimensional space of features (feature space) using a polynomial parameter based on another input parameter [30]. According to [28], an SVM is an extension of the Support Vector Classifier that expands the space using kernels, and a kernel consists of a function that quantifies the similarity between two observable data.

As for **stylometry**, this is the attribute used in behavioral biometrics and is the linguistic style used when writing, called stylometry; it is also worth noting that there is a notable divergence between the types of biometrics and their mutability [7]. For identification through stylometry, attributes such as sentence structure, semantics, and vocabulary patterns in parts of a written text are used.

This form of authentication based on behavioral factors is part of new methods of verifying a user's authenticity according to the system's use and which are still objects of study, being characterized as forms of continuous authentication [11]. However, this is somewhat complex because, according to [31], it is impossible to differentiate different people's styles through predefined attributes. Specifically in the context of this work [32] highlights that the use of styles tria as a tool to verify the authorship of a text requires the existence of quantifiable and distinctive textual characteristics as a prerequisite

These concepts are organized through a workflow with three phases of the work, the first of which is the preprocessing of information that may come from different data sources. The training and validation of the SVM model were used, and that was generated from the data provided for training. Finally, the inference that validates or not the content of the message against the previously trained model. For the latter, a graphical interface was also developed to simulate posting a message on a social network.

The architecture can be described as follows. The section with diagonal lines in blue corresponds to the data extraction, transformation, and loading (ETL) steps. The data used were obtained from the dataset containing posts from users with a large number of followers and subsequently sanitized, modified, and adjusted to suit the following processing.

The steps of extracting stylometric features and training the machine learning model are in red, with a checkered background. The textual features are extracted from the data processed in the previous step and must be in a suitable format for input for the classification model. In this step, only the most relevant features are used. After training the model, it is validated with part of the data set used.

The last step, with a green dotted background, corresponds to the trained model theoretically used in a continuous authentication system. When the user sends a text of any nature in an application, the model performs the inference through the knowledge base and verifies whether the inputted text has characteristics corresponding to the profile identified for the user authenticated in the system.

The **ETL** (Extraction, Transformation, and Cleaning) component is responsible for extracting, manipulating, and loading the data set that will be used later. The Python language and its libraries will transform and select the data in this step. Only data that includes all the information in a format suitable for the feature extraction step will be selected. Inadequate data will be discarded.

The next step corresponds to the **application of stylometry** on the textual publications. In this step, the data obtained in the previous step will be used to extract stylometric features used in future steps to create the users' writing profiles. The features will also be obtained using Python and its libraries, including the Natural Language Toolkit for Natural Language Processing and Emoji. At this stage, the features extracted are

the most relevant according to the bibliography consulted and will serve as input for the chosen machine learning algorithm.

The **training, validation and inference** stages make up the last component of the project and will be implemented using the scikit-learn library. The learning algorithm used will be the Support Vector Machine, and at the end of the training, it should classify whether or not the supposed author wrote the evaluated text. The data will be segmented for training and testing purposes, with each stage having an appropriate amount. In the validation stage, statistical measures will be used to assess the model's accuracy. The Scikit-learn library will be used in these stages.

Using stylometric features as input for the classification model requires identifying, counting, and transforming the analyzed text. Thus, counts were performed on relevant information in the text, which was later used as input for the models.

The first set of counters used corresponds to text features, that is, information related to the way words and sentences were written. The information extracted in this process was the number of punctuation marks used, sentence counts, whitespace counts, and the number of sentences with capitalized first letters. According to the literature, this information is part of the user's writing profile and, therefore, can be used in the classification task of this work.

The second group of counters constitutes elements more specific to social networks. In this regard, the number of emojis used in the posts and hashtags was extracted, the latter being a characteristic especially present on social network X.

Subsequently, tokenization and vectorization steps were performed on the preprocessed text. The vectorization method used was TF-IDF to determine the relevance of each token to the posts.

IV. Discussion

This section presents and discusses the results obtained. The results of an evaluation of each kernel configuration used (**i**) and an accuracy analysis by the author (**ii**) are commented on. The results are concluded with an assessment (**iii**) of the authenticity of a message using the proposed model in a simulated interface.

The best results for each kernel used (**i**) are shown in Table II. The best result was obtained by the configuration that used the linear kernel, where the accuracy was approximately 90%. The polynomial, RBF, and sigmoid kernels reached 87%, 83%, and 69%, respectively.

Kernel	C Value	Gamma Value	Accuracy	Precision
Linear	1	-	0.90	0.90
Polynomial	0.1	-	0.87	0.89
RBF	1000	0.1	0.83	0.84
Sigmoid	0.1	-	0.69	0.69

TABLE II
BEST RESULTS DURING THE TESTING STAGE

Additionally, it was observed that the models presented different performances in the classification of publications, varying according to the author (**ii**). The highest accuracy was recorded for Sebastian Ruder's publications, reaching 94%, while Katy Perry's resulted in the lowest accuracy, with 86%. Figure 3 illustrates the Distribution of accuracies considering the best model configuration evaluated.

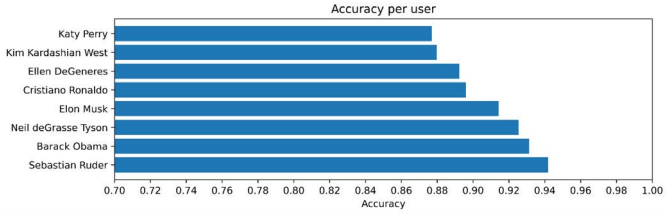

Fig. 3. Distribution of accuracy by user

Therefore, it is worth noting that the classification models demonstrate different accuracies for each user, reflecting particularities in the quantification of textual elements and writing patterns.

In order to demonstrate the applicability of the model in a real scenario, a user interface (**iii**) was developed that simulates the flow of publishing and validating content on a social network. Figure 4 shows the implemented interface, in which the user Barack Obama is authenticated in the system. When entering a text, the system applies the model trained with the publications of this user to verify whether the content corresponds to the writing profile identified for the author.

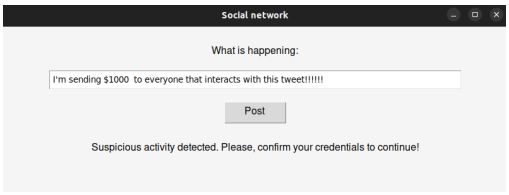

Fig. 4. Distribution of accuracy by user

In the simulation, the submitted text states that money will be donated to anyone interacting with the publication. Because it differs from the publications previously made by the supposed author, a message requesting confirmation of credentials was displayed. In a real system, a login data confirmation screen can be used at this stage, and if the information is filled in correctly, the user is allowed to continue using the system.

Three stages were included: evaluation of the tests according to the kernel, according to the user, and with inference through a simulator program. These stages provided us with an overview of the results while still focusing on the authentication results without further detailing machine learning aspects. This scope was defined to better adhere to the scope of this event.

V. Conclusion and Future Works

As seen in Table II, the results found for the tests are relevant and adequate for verifying the authors' identities. Although there is some variation, as shown in Figure 3, according

to the characteristics of each author's textual construction, it is possible to state that stylometry is adequate for continuous authentication for social media and for identifying published fakes. This is reinforced by validation using a simulator that evaluates publications in the case of one of the trained users, Figure 4.

For new development stages and future work, it is suggested that the performance of the proposed proposal be evaluated and compared to other related works. Although it is not the purpose of this publication, it is important to emphasize the need to maintain privacy in some cases (i.e., emails) where this work may be applied. Equally important is the need to carefully evaluate the future use of generative artificial intelligence in message creation and its impact on this work. Furthermore, evaluating more text characteristics and monitoring how these impact the proposal's performance will be important. As a result, it will be possible to adjust the temporal and spatial performance of the proposal according to the desired level of security in the authentication process.

VI. Acknowledgments

Anonymized

References

[1] D. Korobenko, A. Nikiforova, and R. Sharma, "Towards a privacy and security-aware framework for ethical ai: Guiding the development and assessment of ai systems," in *Proceedings of the 25th Annual International Conference on Digital Government Research*, 2024, pp. 740–753.

[2] P. Capasso, G. Cattaneo, and M. De Marsico, "A comprehensive survey on methods for image integrity," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, pp. 1–34, 2024.

[3] Z. Yu, S. Zhai, and N. Zhang, "Antifake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 460–474.

[4] W. Hutiri, O. Papakyriakopoulos, and A. Xiang, "Not my voice! a taxonomy of ethical and safety harms of speech generators," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 359–376.

[5] S. Ristovska, "Ways of seeing: The power and limitation of video evidence across law and policy," *First Monday*, 2023.

[6] T. Bianco, G. Castellano, R. Scaringi, G. Vessio *et al.*, "Identifying ai-generated art with deep learning." in *CREAI@ AI* IA*, 2023, pp. 16–25.

[7] M. L. Brocardo, "Continuous authentication using stylometry," Ph.D. dissertation, 2015.

[8] R. Kaur, S. Singh, and H. Kumar, "Tb-coauth: Text based continuous authentication for detecting compromised accounts in social networks," *Applied Soft Computing*, vol. 97, p. 106770, 2020.

[9] A. V. Nadimpalli and A. Rattani, "Social media authentication and combating deepfakes using semi-fragile invisible image watermarking," *Digital Threats: Research and Practice*, vol. 5, no. 4, pp. 1–30, 2024.

[10] G. Kaur, U. Bonde, K. L. Pise, S. Yewale, P. Agrawal, P. Shobhane, S. Maheshwari, L. Pinjarkar, and R. Gangarde, "Social media in the digital age: A comprehensive review of impacts, challenges and cybercrime," *Engineering Proceedings*, vol. 62, no. 1, p. 6, 2024.

[11] A. F. Baig and S. Eskeland, "Security, privacy, and usability in continuous authentication: A survey," *Sensors*, vol. 21, no. 17, p. 5967, 2021.

[12] S. Alterkavı and H. Erbay, "Novel authorship verification model for social media accounts compromised by a human," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13 575–13 591, 2021.

[13] M. Litvak, "Deep dive into authorship verification of email messages with convolutional neural network," in *Information Management and Big Data: 5th International Conference, SIMBig 2018, Lima, Peru, September 3–5, 2018, Proceedings 5*. Springer, 2019, pp. 129–136.

[14] E. Castillo, O. Cervantes, and D. Vilarino, "Authorship verification using a graph knowledge discovery approach," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 6075–6087, 2019.

[15] R. DATASET, "Evaluating adversarial stylometry using textfooler," Ph.D. dissertation, tilburg university.

[16] W. Xu and B. C. Fung, "Stylelink: User identity linkage across social media with stylometric representations," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, 2025, pp. 2076–2088.

[17] H.-C. Yang, Y.-L. Hung, and L.-C. Wang, "Stylometry-based fake news classification using text mining techniques," in *Proceedings of the 2024 11th Multidisciplinary International Social Networks Conference*, 2024, pp. 85–94.

[18] M. Hosseini, S. P. Horbach, K. Holmes, and T. Ross-Hellauer, "Open science at the generative ai turn: An exploratory analysis of challenges and opportunities," *Quantitative Science Studies*, vol. 6, pp. 22–45, 2025.

[19] P. Mirowski, "The future (s) of open science," *Social studies of science*, vol. 48, no. 2, pp. 171–203, 2018.

[20] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Wurthwein *et al.*, "The open science grid," in *Journal of Physics: Conference Series*, vol. 78, no. 1, 2007.

[21] E. D. Foster and A. Deardorff, "Open science framework (osf)," *Journal of the Medical Library Association: JMLA*, vol. 105, no. 2, p. 203, 2017.

[22] N. Ghasem Abadi, "Machine learning-based authentication of banknotes: a comprehensive analysis," *Big data and computing visions*, vol. 4, no. 1, pp. 22–30, 2024.

[23] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media, 2020.

[24] I. J. Akpan, Y. M. Kobara, J. Owolabi, A. A. Akpan, and O. F. Offodile, "Conversational and generative artificial intelligence and human–chatbot interaction in education and research," *International Transactions in Operational Research*, vol. 32, no. 3, pp. 1251–1281, 2025.

[25] B. Saha, "Generative ai for text generation: Advances and applications in natural language processing," *Journal of Computer Allied Intelligence (JCAI, ISSN: 2584-2676)*, vol. 3, no. 1, pp. 77–91, 2025.

[26] K. Chowdhary, *Fundamentals of artificial intelligence*. Springer, 2020.

[27] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[28] A. B. Manual, "An introduction to statistical learning with applications in r," 2013.

[29] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[30] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2022.

[31] F. Iqbal, R. Hadjidj, B. C. Fung, and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *digital investigation*, vol. 5, pp. S42–S51, 2008.

[32] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov, and P. Demidov, "A survey on stylometric text features," in *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, 2019, pp. 184–195.