Identifying the Author: A Comprehensive Review of Authorship Attribution Methods and Their Practical Uses

***Abstract*—This paper presents a thorough examination of various techniques employed in authorship attribution and their pragmatic applications. The significance of authorship attribution has escalated in the digital era owing to the extensive employment of text-oriented communication. The article provides a comprehensive survey of diverse methods employed for authorship identification, encompassing conventional stylometric techniques as well as contemporary machine learning methodologies. This paper examines the benefits and drawbacks of various techniques, as well as their potential utility in forensic investigations, plagiarism detection, and related fields. The paper additionally underscores the significance of discerning suitable characteristics and data sets in order to attain precise outcomes. In general, the review offers significant perspectives on the present status of authorship attribution investigation and its pragmatic consequences.**

***Index Terms*—Natural Language Processing, Authorship Attribution, Machine Learning, Language Models, Authorship Identification**

I. Introduction

The identification of authorship is a complex and significant undertaking within the realm of natural language processing, with numerous pragmatic implications in domains such as forensic investigations, detection of plagiarism, and literary analysis. The precise identification of the author of a written work can yield significant insights into various aspects such as the author's writing style, impact, and purpose. Throughout time, a diverse array of techniques has been devised for the purpose of authorship attribution, spanning from rudimentary statistical methods to intricate machine learning algorithms. Diverse textual genres, such as novels, electronic mail, and social media entries, have been subjected to these methodologies. The task of identifying the authorship of a given piece of content has become increasingly difficult due to the proliferation of digital content creation and sharing on the internet. The aforementioned circumstance has resulted in the emergence of novel methodologies and instruments that are tailored to address the distinctive attributes of digital text. This paper provides a comprehensive overview of the current state of authorship attribution, encompassing both conventional and online-specific techniques. The present discourse entails an analysis of the merits and demerits of individual methods, accompanied by instances of their pragmatic applications. The objective of our study is to furnish scholars and professionals with a thorough exposition of the domain while emphasizing the gaps in knowledge that require additional investigation.

II. Literature Review

The paper Whodunit [1] states authorship attribution is a process that entails the identification of the writer of a particular text. This method is applied in diverse contexts, such as the identification of the origin of anonymous threats and the detection of plagiarism. Conventional techniques are predicated upon features that are customized for a specific dataset, whereas pre-trained language models are not tailored for this particular undertaking. The present study introduces Contra-X, a novel approach that utilizes a contrastive objective to fine-tune pre-existing models and acquire author-specific representations.

The Paper Small-Scale Cross-Language [2] investigates three distinct methodologies for the purpose of cross-lingual authorship attribution, which pertains to the identification of the original author of a given document from a pool of potential candidates in a language that differs from the original text. There exist three distinct methodologies for this task, namely: employing language-agnostic characteristics in conjunction with a support vector machine, leveraging a pre-existing multilingual BERT model, and utilizing machine translation to convert one feature set into the other. The authors employ datasets comprising of bilingual writers who compose in English and one of five additional languages.

In paper, [3] The task of authorship attribution holds significant relevance in contemporary society, with its outcomes finding utility in domains such as forensic analysis, information security, computational linguistics, and safeguarding intellectual property rights. The assignment can be classified into distinct categories, namely authorship identification, verification, clustering, and profiling. To address this problem, a range of methods, such as fastText, SVM-GA, CNN, LSTM, BERT, KNN, DT, RF, LR, and NB, are employed to ascertain the veritable author of contested texts.

The paper [4] examines computational methodologies for the purpose of authorship attribution, which pertains to the identification of the author of a given text. The discourse delves into two overarching methodologies, namely classification-based and similarity-based approaches. The classification-based paradigm employs a supervised machine learning technique to create a classifier, with the prevalent use of deep learning methods in contemporary times.

The present paper Integrating Roberta [5] delves into the domain of authorship attribution within the realm of natural

language processing, with a particular emphasis on the intricacies involved in discerning the authors of succinct social media messages. The proposed approach by the authors involves the amalgamation of the RoBERTa [6] language model's fine-tuning with a neural network model that caters to user writing styles. The authors of the study have proposed the implementation of distinct encodings for tweet constructs, and have utilized a deep neural network to acquire multi-view representations of user-generated content.

This paper [7] investigates the task of authorship attribution in unconventional settings, with a particular emphasis on forums situated on the dark web. These forums may be utilized by extremist groups to disseminate their ideology and potentially pose a security risk, given the anonymity afforded to their members. The proposed approach by the authors involves the utilization of pre-trained language models such as transformers and KERMIT to address the aforementioned challenge. The research conducted a comparative analysis of different models and features and concluded that syntactic models utilizing KERMIT exhibit superior performance compared to transformer-based models on the given dataset.

According to the paper, BertAA [8] Prior methodologies have been dependent on feature engineering. However, BertAA introduces a fine-tuning technique for a pre-existing BERT language model, leading to exceptional outcomes on datasets comprising as many as 100 authors. The integration of stylometric and hybrid features in an ensemble architecture yields an enhancement in the macro-averaged F1 score. A performance standard has been established for the complete IMDb dataset, which has yet to be comprehensively analyzed in the context of AA.

Following the paper Network Motifs Identification [9] the application of complex network analysis has been employed in the examination of tangible systems such as cells and the Internet. In recent times, significant progress has been made in this field as a result of the identification of intricate patterns in actual systems. Texts can be depicted as networks wherein the nodes symbolize words and the edges symbolize syntactic, semantic, or empirical connections. The utilization of word co-occurrence networks has been observed in various tasks, including but not limited to extractive summarization and authorship attribution.

In this paper [10], the authors suggest the utilisation of distributed document representation for authorship attribution. The authors employ the Doc2vec approach to construct a vector representation at the document level. The researchers conducted experiments utilizing relatively modest corpora and assessed the efficacy of their approach in comparison to prior studies. They also scrutinized the impact of utilizing word n-grams as input data types.

This paper Syntactic N-grams [11] proposes the utilization of syntactic n-grams (sn-grams) within the field of natural language processing. These sn-grams are derived from the arrangement of constituents in syntactic trees, as opposed to the superficial textual representation. N-grams preserve linguistic data while disregarding language-specific superficial structures. These can be utilized in any task that typically employs conventional n-grams, albeit necessitating parsing, which is a time-consuming process.

III. Dataset

A. Whodunit

The dataset comprises ten distinct blogs that have been written by a total of 50 authors. The corpus comprises 2,500 articles in entirety, with a mean of 250 articles per writer. The Blog50 dataset is an expansion of the Blog10 dataset, comprising 50 distinct blogs composed of a total of 100 unique authors. The dataset comprises a total of 5,000 articles, with an average of 100 articles per author.

The dataset [12] under consideration comprises 62 distinct movie scripts. The dataset is commonly referred to as IMDb62. The corpus comprises approximately 21,000 dialogues, with an average script length of roughly 22,000 words.

The TuringBench dataset [13] comprises 50,000 records, each of which encompasses 1,000 lexical units. The corpus comprises 50 distinct authors, each of whom is associated with 1,000 documents.

B. Small-Scale Cross-Language

The corpus utilized in this study comprises comments posted on Reddit by multilingual writers in English and any one of five additional languages. The utilization of a small-scale dataset offers a practical and authentic situation for the task of cross-language authorship attribution.

C. authorship attribution

The corpus comprises a collection of 1100 literary works in the Russian language, authored by 100 individuals. The dataset comprises literary works authored by Russian classics.

another dataset comprises brief comments generated by users of the VK social network. The training set was comprised of texts that had a minimum length of 50 characters.

D. Siamese

The paper [4] employs multiple datasets to conduct extensive author identification, namely PAN 2015. The PAN 2020 author verification dataset, comprising of more than 50,000 pairs of texts, is also utilized by them.

E. Integrating RoBERTa

The study employed a dataset devised by Schwartz et al consisting of roughly 9,000 Twitter users and 1,000 posts ascribed to each user, culminating in a cumulative count of 9 million posts.

F. Shedding Light

The present study employs a dataset sourced from the Islamic Network's online discussion forum, comprising a total of 91,874 posts, 13,995 discussions, and 2082 active users.

G. BertAA

Enron Email corpus: This email corpus comprises a total of 130,000 electronic mail exchanges among Enron managers, which were made publicly accessible following the company's bankruptcy.

IMDb Authorship Attribution Corpus: This Corpus contains 271,000 movie reviews by 22,116 authors, with an average of 12.3 texts per author.

Blog Authorship Attribution Corpus: It includes over 680,000 unfiltered blog posts from more than 19,000 authors, with an average of 35 posts per author.

H. Network Motifs Identification

The corpus is comprised of a collection of 40 literary works authored by eight distinct writers, sourced from the Project Gutenberg database. The literary works were released within the time frame of 1835 to 1922.

I. Distributed Document Representation

Historical and literary documents dataset: The corpus comprises of the Gutenberg Corpus, which encompasses more than 50,000 freely accessible electronic books, and the Penn Treebank.

Controlled datasets: The dataset comprises two instances that contain a total of 18,000 posts that are distributed across 20 distinct topics.

PAN/CLEF 2012 Problem A benchmark: The dataset was employed in the PAN/CLEF 2012 competition aimed at authorship attribution. It comprises a collection of written documents authored by 50 distinct individuals, with each contributor submitting 20 documents. The corpus comprises two distinct sets of documents, namely a training set consisting of 500 documents and a test set comprising 500 documents.

PAN/CLEF 2012 Problem C benchmark: The dataset was employed for the purpose of detecting instances of text reuse, comprising a training subset consisting of 250 documents and a test subset consisting of 250 documents.

PAN/CLEF 2012 Problem I benchmark: The dataset was employed for the purpose of intrinsic plagiarism identification and comprises a training set consisting of 250 documents and a test set containing 250 documents.

Reuters Corpus Volume 1 (RCV1): The corpus comprises in excess of 800,000 news articles sourced from Reuters, a prominent news establishment.

J. Syntactic N-grams

The corpus utilized in the experiments comprises textual data sourced from Project Gutenberg, a digital repository of freely available electronic books. The aforementioned texts were authored by three individuals and comprises 39 textual documents.

IV. Model and Architecture

A. Whodunit

The paper Whodunit [1] proposes Contra-X to augment the cross-entropy loss function with a contrastive learning objective to achieve this goal. The model is decomposed into a pre-trained language model and a classifier layer. During training, the model jointly optimizes the cross-entropy loss and the contrastive loss. The input length is set to 256, and the embedding length per token is 768.

B. Small-Scale Cross-Language

The paper [2] outlines a methodology that encompasses three distinct approaches for cross-language attribution, namely: (1) Language-Independent Features, (2) Pre-Trained Multi-Language Models, and (3) Translation. The Language-Independent Features methodology employs a pair of language-agnostic features that rely on syntactic information. On the other hand, DT-grams represent substructures of dependency graphs that capture the inter-word relationships. The approach of utilizing Pre-Trained Multi-Language Models involves the utilization of mBERT. The sliding window approach is employed to produce several samples from each document. The Translation methodology employs the Marian Neural Machine Translation (NMT) models, which are accessible for numerous language pairings.

C. authorship attribution

The paper [3] examines various methods used for attribution, which involves identifying the authorship of literary texts. SVM, LR, NB, DT, RF, and KNN, have been tested, and the working of these methods for advantages, such as logical justifiability and faster training time. Additionally, the paper highlights the importance of the NLP library fastText from Facebook Research for the development of vector semantic models and ML in text processing.

D. siamese

The proposed approach employs a Siamese architecture, where convolutional neural networks are utilized as sub-networks, to generate similarity measures for pairs of texts through authorship attribution. The neural network receives input at the character level and is composed of an embedding layer, four convolutional layers, and a dense layer. The utilized energy function involves the computation of the L1 distance between the outputs of the sub-network's final layers.

E. Integrating RoBERTa

The paper's [5] attribution framework under consideration comprises of two distinct modules. The initial module employs RoBERTa to derive contextualized vector representations of tweets. The employed methodology involves the utilization of a triplet loss function for the acquisition of post representations, as well as the implementation of four CNN models for the purpose of generating embeddings on various linguistic features. The feature embeddings are combined on a per-user basis utilizing a fusion function, resulting in the creation of user-style embeddings.

F. Shedding Light

The study from paper [7] proposed models put forth by the authors incorporate syntactic, lexical, and stylistic features, alongside pre-trained language models utilizing transformers. The models proposed in the study exhibit a notable degree of accuracy, and the paper offers a comprehensive account of the various stages entailed in the development pipeline of the models.

G. BertAA

The BertAA model is a variant of the BERT model that has been fine-tuned. The model is equipped with a dense layer and a softmax activation function and is trained over a limited number of epochs to effectively classify textual data into distinct author categories. BertAA has been designed to incorporate stylometric and hybrid features in order to account for content, as well as stylometric and hybrid features. The process of classification is subsequently carried out through the utilization of a logistic regression (LR) model.

H. Network Motifs Identification

The paper [9] expounds upon the methodology of constructing a complex network model which involves initial text preprocessing. The subsequent procedure involves linking words to construct a co-occurrence network. In this network, nodes are indicative of words, and links are established between nodes on the condition that the corresponding words appear in proximity at least once in the pre-processed text. The network's topological configuration is distinguished by various metrics which are utilized for the purpose of comparing classification outcomes that have been acquired through distinct networked measurements.

I. Distributed Document Representation

The paper [10] examines the topic of authorship attribution through the process of acquiring document vectors achieved through the utilization of the Doc2vec methodology. The aim of the training is to forecast a particular word based on its contextual information, whereby each word is associated with a distinct vector within a matrix. The identical approach is employed for both paragraph vectors and document vectors. The task of authorship attribution is commonly regarded as a multiclass classification problem, whereby the training set is comprised of known authors and their respective texts.

J. Syntactic N-grams

The authors of the paper [11] outline two basic steps: text representation and classification. The paper focuses on the importance of stylometric features and the need for a more elaborate representation of writing style for explaining stylistic analysis. The most effective measures for authorship attribution are lexical and character features, while the most successful stylometric methods are based on low-level information such as character n-grams or auxiliary words. The paper discusses profile-based and instance-based methods for classification.

V. Result Analysis and Findings

This paper Whodunit [1]examines the effects of contrastive learning on models of human authorship attribution. The findings indicate that the integration of contrastive learning enhances the fundamental level of performance. The most significant enhancements are evident in the Blog10 and Blog50 datasets, exhibiting a maximum increase of 6.8% for BERT and 3.7% for DeBERTa.

The study conducted on paper Small-Scale Cross-Language [2] an evaluation of various methodologies for cross-lingual authorship attribution utilizing limited datasets. The research discovered utilization of word n-grams in conjunction with SVM yields a score of 0.38 for Arabic. In contrast, English DT-grams demonstrate superior performance for DE, ES, FR, and NL, with accuracy percentages of 46.7, 46.5, 55.2, and 43.3, respectively.

In the research paper [3] training of the models was conducted on the feature space and TF-IDF. The training times of CNN and hybrid networks were observed to be faster in comparison to LSTM, BiLSTM, and BERT. The investigation additionally documented the precision scores attained by diverse models on varying quantities of authors, whereby fastText garnered the most elevated mean score of 76.3

.

The paper Siamese [4] outlines the application of Siamese Neural Networks in the context of author verification on extensive datasets. The reported accuracy scores of significance are as follows: SiamL1 and Siamcos achieved 0.980 and 0.978, respectively, for author verification in the English subset of PAN 2015. Siamcos attained a score of 0.883 for author verification in PAN 2020. Additionally, Siamcos yielded accuracy scores of 0.843 and 0.761 on ff-5K and PAN 2020.

The research conducted on paper [5] assessed the precision of various machine learning models using a dataset comprising of ten distinct clusters. The CNN-WC2+LPS model exhibited the highest level of performance among the baseline models, achieving an accuracy rate of 0.836. The investigation additionally examined models based on RoBERTa and determined that the optimal performance was achieved by concatenating the last four hidden layers as input for CNN, in conjunction with user writing styles. This approach yielded an accuracy of 0.882.

This study of paper [7] aims to evaluate the efficacy of various models in the task of authorship attribution (AA) using the Islamic Network data set. The findings suggest that BoPOS models outperformed other configurations in two-thirds of cases, implying that style is a significant feature in the context of AA. The BoPOS model yielded an accuracy score of 84.9% for three authors and 74.5% for five authors, while the TF-IDF POS model only achieved an accuracy score of 5.9% for 10 authors.

The paper BertAA [8] suggested a framework named BertAA. The study conducted a comparative analysis of the

model against a word-level TF-IDF-LR model that incorporated stemming and stop-words removal. In all conducted experiments, BertAA exhibited superior performance compared to the TF-IDF and LR benchmarks, with an average increase in relative accuracy of 14.3%. The Blog Authorship Corpus was utilised to evaluate the performance of BertAA, which yielded a state-of-the-art accuracy of 65.4% for 10 authors and 59.7% for 50 authors.

The research of paper [9] examined the utilization of network motifs and additional characteristics to attribute authorship through the application of machine learning techniques. The classification accuracy rates solely based on the frequency of directed motifs exhibited a range of 45% to 55%. The experiment yielded the highest accuracy rate when utilising the frequency of the 20 most common words across all 40 books as a classification feature. The accuracy rates for C4.5, kNN, SVM, and Bayes were 55%, 67.5%, 72.5%, and 55%, respectively.

The paper [10] employed machine learning techniques in conjunction with the Doc2vec approach to categorise diverse datasets according to authorship attribution. Two distinct classifiers, namely SVM and Linear Regression, were employed in the experiments. The findings indicate that Linear Regression outperformed SVM and yielded the most favourable outcomes. The method proposed for The Guardian corpus attained the highest level of accuracy, with D2V words+2-grams resulting in an average accuracy that varied between 55.34% and 94.75% across diverse categories.

The study of paper [11] used n-grams and sn-grams to identify the author of a given text from a corpus of texts from three native English-speaking authors. SVM classification was found to perform better than NB and J48, and sn-grams tended to outperform traditional n-grams. The study achieved 100% accuracy for bigrams and trigrams using various sn-grams, n-grams, and profile sizes. The SVM classifier achieved 93% accuracy for 4 grams and 87% accuracy for 5 grams.

VI. Comparative disscussion

These documents being evaluated are all related to the field of authorship attribution, which involves the identification of the writer of a given text. The initial manuscript [1] introduces Contra-X, a technique that utilizes contrastive learning to acquire extremely distinctive representations for the purpose of authorship identification. The Contra-X system is constructed using a pre-existing language model and a classifier layer, resulting in an exceptional performance on three standard datasets. The second scholarly article delineates three distinct methodologies for cross-lingual attribution, namely Language-Independent Features, Pre-Trained Multi-Language Models, and Translation. The aforementioned techniques are expounded upon in detail within the paper [2]. In the fourth paper, Zhang et al. (2021) introduce a Siamese architecture that utilizes convolutional neural networks to produce similarity metrics for pairs of texts written by either the same or different authors. The proposed approach aims to enhance the accuracy of text similarity measurement. The attribution framework proposed in the fifth paper (Wang et al., 2021) utilizes RoBERTa for text representation and a convolutional neural network classifier for author categorization. The second component of the framework employs a triplet loss function and four convolutional neural network models to produce embeddings based on diverse linguistic characteristics, with the aim of acquiring knowledge of the writing style of users. The sixth paper, as documented by Ranaldi (2022), presents various models for authorship attribution, encompassing syntactic, semantic, and discourse-based models.

In general, the aforementioned papers demonstrate the efficacy of diverse methodologies in the realm of authorship attribution. The article by Fedotova (2021) highlights the effectiveness of classical machine learning algorithms in the field of authorship attribution. These algorithms are deemed to be transparent and justifiable in contrast to the more opaque neural networks. The efficacy of pre-trained language models, such as RoBERTa as highlighted in [5], has been established in producing contextualized vector representations of textual data. The methodology of contrastive learning, as proposed in the scholarly work [1], has demonstrated noteworthy enhancements in outcomes of authorship attribution in comparison to prior approaches. The effectiveness of the Siamese architecture for authorship attribution at the character level has been demonstrated in the work proposed by Zhang (2021). Additionally, Ranaldi (2022) has shown that incorporating syntactic, semantic, and discourse-based features can also be effective for authorship attribution.

Additionally, the aforementioned papers underscore the significance of authorship attribution across different languages. The article by Murauer et al. (2021) presents a methodology that offers three distinct strategies for addressing the issue at hand. These include the implementation of pre-trained multilingual models such as mBERT, the use of language-independent features such as POS tags, and translation. The integration of user writing style information in the attribution process is also emphasized in a recent publication by Wang et al. (2021).

VII. LIMITATIONS AND SCOPE

This section aims to conduct a comparative and analytical examination of the limitations and scope of these research papers.

In the paper Whodunit [1], it is noteworthy that this approach may result in amplified bias against specific authors and reduced fairness across different classes. The article proposes potential avenues for further investigation, including an exploration of the uniformity of contrastive learning's influence on fluctuations in accuracy at the class level and the resolution of the dilemma between enhanced performance on author pairs with high similarity and reduced performance on other authors.

This paper Small-Scale Cross-Language's [2] primary constraints are the disparate outcomes of the distinct datasets resulting from the linguistic divergence between language pairs and the machine translation outcomes' reliance on the

caliber of translation models, particularly for languages with limited resources.

In the third paper [3], the proposed methodology is subject to certain limitations, such as the number of texts, the number of samples, and the characteristics of the texts. The study proposes that the training dataset ought to exclusively comprise texts that are congruent with the writing style of the author and that the identity of the author of each training text should be ascertainable.

The research on paper Siamese [4] the comparative analysis is solely conducted utilizing the similarity-based approach. The present investigation does not delve into alternative architectures employed for single-shot tasks in the realm of image processing. Subsequent research may investigate the utilization of alternative metrics, such as hyperbolic distance, for the purpose of authorship attribution.

The present study from paper [5] has constraints and potential for further investigation encompassing assessing the model's efficacy with a greater number of authors or reduced texts from the same author. Furthermore, investigating alternative pre-existing language models and enhancing user writing proficiency by incorporating the temporal aspects of user posts may augment the precision of the model.

The paper titled "Shedding Light" [7] recommends further research to explore the language patterns employed in the dark web and establish connections between users of dark web forums and those of standard web forums.

The presented paper's [8] proposed approach exhibits favorable performance on concise texts with minimal disparities in the number of texts per author. Nonetheless, the utilization of the model is constrained to scenarios in which there exists an adequate amount of training data for each author. There is potential avenues for expansion, including additional pretraining on the target domain and investigation into alternative pre-trained language models.

This paper [9] findings indicate that motifs possess the ability to encapsulate various facets of the writing style of distinct writers. However, the precision achieved is comparatively inferior to certain conventional methodologies documented in the literature.

The present methodology of the paper [10] has constraints of this approach encompassing its reliance on annotated datasets and the requisite for substantial quantities of training data. The extent of its coverage is restricted to tasks involving attribution of authorship in textual format.

The study's [11] scope is limited to authorship attribution using sn-grams and traditional n-grams on a relatively small corpus of works from three authors. Future work could explore larger corpora, different parsing techniques, and other NLP tasks.

VIII. Conclusion

The field of authorship attribution is a multifaceted and interdisciplinary area that has garnered significant interest in recent times, owing to its diverse practical implications, spanning from literary and historical scrutiny to forensic inquiries and cybersecurity. In summary, authorship attribution is a complex domain that has gained prominence in contemporary times.

The field of authorship attribution is characterized by a dynamic and evolving landscape, with a range of methodological approaches being employed to address the challenges and limitations of existing techniques. Ongoing research efforts are focused on developing novel methods and techniques to enhance the accuracy and reliability of authorship attribution analyses. Although there is no foolproof approach to ensure absolute precision in all scenarios, the amalgamation of diverse methodologies and techniques can enhance the dependability and authenticity of outcomes related to authorship attribution.

References

[1] B. Ai, Y. Wang, Y. Tan, and S. Tan, "Whodunit? learning to contrast for authorship attribution," *arXiv preprint arXiv:2209.11887*, 2022.

[2] B. Murauer and G. Specht, "Small-scale cross-language authorship attribution on social media comments," in *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, 2021, pp. 11–19.

[3] A. Fedotova, A. Romanov, A. Kurtukova, and A. Shelupanov, "Authorship attribution of social media and literary russian-language texts using machine learning methods and feature selection," *Future Internet*, vol. 14, no. 1, p. 4, 2021.

[4] C. Saedi and M. Dras, "Siamese networks for large-scale author identification," *Computer Speech & Language*, vol. 70, p. 101241, 2021.

[5] X. Wang and M. Iwaihara, "Integrating roberta fine-tuning and user writing styles for authorship attribution of short texts," in *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5*. Springer, 2021, pp. 413–421.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] L. Ranaldi, F. Ranaldi, F. Fallucchi, and F. M. Zanzotto, "Shedding light on the dark web: Authorship attribution in radical forums," *Information*, vol. 13, no. 9, p. 435, 2022.

[8] M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida, "Bertaa: Bert fine-tuning for authorship attribution," in *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 2020, pp. 127–137.

[9] V. Q. Marinho, G. Hirst, and D. R. Amancio, "Authorship attribution via network motifs identification," in *2016 5th Brazilian conference on intelligent systems (BRACIS)*. IEEE, 2016, pp. 355–360.

[10] J.-P. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, D. Pinto, and L. Chanona-Hernández, "Application of the distributed document representation in the authorship attribution task for small corpora," *Soft Computing*, vol. 21, pp. 627–639, 2017.

[11] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing," *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, 2014.

[12] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with topic models," *Computational Linguistics*, vol. 40, no. 2, pp. 269–310, 2014.

[13] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, "Turingbench: A benchmark environment for turing test in the age of neural text generation," *arXiv preprint arXiv:2109.13296*, 2021.