Intelligent Analysis in Text Authorship Identification

Ilya Trukhanovich and Anton Paramonov
Belarusian State University of Informatics and Radioelectronics
Minsk, Republik of Belarus
ilya.trukhanovich@gmail.com, a.paramonov@bsuir.by

***Abstract*—The problem of text authorship identification is considered. A generalized review of the current state of the problem is provided. A solution based on the modification of ensemble machine learning methods is proposed. The possibilities of applying sophisticated computational methods such as natural language processing, machine learning algorithms and stylometric analysis to identify individual writers based on their distinctive linguistic models are investigated. A hypothesis is put forward about the need for multidimensional text analysis and a new approach is proposed for identifying the authorship of the text in the form of a hybrid intelligent system.**

***Keywords*—authorship identification, semantics, natural language processing, machine learning, quantum technologies**

I. Introduction

Assigning authorship to texts that are disputed or anonymous requires the identification of text authors, which is a basic problem in linguistics. Conventional approaches depend on human analysis and linguistic knowledge, but with the development of intelligent analysis methods, there is an increasing interest in using computational tools to improve and automate this process. Researchers are looking into novel ways to reliably identify writers based on their writing traits by utilizing machine learning models, natural language processing (NLP) methods, and stylometric aspects.

The use of intelligent analytic techniques has attracted a lot of attention in the field of text authorship identification because of its potential to improve the precision and effectiveness of author identification based on writing styles. This paper explores the potential of using mining methods to identify text authorship, with the goal of highlighting the prospects, challenges, and possible uses of these modern techniques. A hypothesis is discussed about how intelligent analysis can transform text authorship identification and open the door for more trustworthy and robust attribution techniques in linguistics through a thorough investigation of the body of research and case studies.

II. Current authorship identification usage for plagiarism detection

A. General

In academics, publishing, and other businesses, plagiarism detection is an essential procedure for guaranteeing the uniqueness and integrity of written content. It is comparing a given text with a sizable database of previously published content using specialized tools and procedures to find any instances of plagiarism or unoriginal content. Plagiarism detection software can identify possible instances of purposeful or inadvertent plagiarism by comparing the content to other sources.

Detecting plagiarism is essential for sustaining academic integrity, credibility in research and publications, and a writing culture that values originality and ethics. It supports integrity, authenticity, and respect for intellectual property rights by assisting publishers, educators, researchers, and content producers in spotting and dealing with cases of plagiarism.

Plagiarism detection uses various techniques, such as citation analysis, machine learning (ML), natural language processing, text comparison algorithms, etc. and using these techniques the software analyzes the document's text structure, linguistic structure, citation style and other texts characteristics which are more accurate to its originality may be considered [1].

Programs for text scanning can be used to detect plagiarism. Using these tools, users can upload documents or enter text for research purposes. A report that displays any matching information is generated and stored in users' databases. By reading these reports, users can identify potential piracy cases and take necessary actions to correct the problem.

Properly crediting sources, properly paraphrasing, using quotations when quoting directly, committing to originality in writing and citing any borrowed ideas well is also important to help people develop good writing habits that encourage originality, stealing inadvertently prevents it.

B. Plagiarism detection approaches

Based on how they identify plagiarism, methods and systems for doing so can be categorized into many groups [2].

At the moment, there are various researches in this direction, which include analyzing current achievements [3].

These are a few of the principal approaches and strategies:

- Matching methods are useful for plagiarism detection because they can identify comparable strings in documents.
- Syntax-based methods match words based on their grammatical structures to find similarities in documents. It

works well for finding exact duplicates, but may not work well for altered texts with the same concept.
- The external method of plagiarism detection uses reference cards from which pieces can be extracted verbatim. Screens suspicious documents for adjacent or similar excerpts in the reference database and sends the results to the person in charge for review.
- Citation pattern analysis with citation patterns can be a strong marker of semantic similarity between documents when compared using this method. Comparability is determined by the size of the citations in the sample and by the similarity of their structure and/or scope.
- Style measurement analyzes an author's specific writing style to identify possible patterns of copying. This works well for spotting annotated and masked plagiarism, but may not work for highly annotated or translated pieces.
- Intrinsic plagiarism detection systems detect plagiarism just look at the text that has to be assessed; they don't compare it to other documents. Their objective is to identify any alterations in an author's distinctive writing style that might be signs of possible copying.

At the same time, authorship identification methods need to be modified in order to advance the field, increase efficiency, handle complicated features, adjust to changing textual environments, increase accuracy, and withstand adversarial tactics. These changes open the door to stronger and more reliable authorship attribution techniques in the fields of text analysis.

Further development in solving the problem of authorship identification seems to be the use of an integrated approach, taking into account the capabilities of various promising areas, for example, in the form of hybrid text analysis systems.

III. Modifications of machine learning methods for authorship identification and plagiarism detection

A. Ensemble methods

Advanced strategies known as ensemble machine learning integrate several models to increase prediction performance and accuracy. By combining the advantages and diversity of multiple models, these techniques produce better outcomes than any one model could on its own. For complicated situations, the main objective of ensemble approaches is to improve predictions while lowering bias and variance.

There are several types of ensemble methods.

Bagging is the process of training several models — typically of the same kind — on several training dataset subsets. By randomly selecting replacement samples from the original dataset, these subsets are formed, allowing for the appearance of some samples in a subset more than once and others not at all. For regression issues, the final prediction is determined by averaging the forecasts, and for classification problems, it is determined by majority vote. The goal of bagging is to prevent overfitting and lower variation. The Random Forest method is a well-known example of bagging applied to decision trees.

By using boosting techniques, a series of models is trained so that each model tries to fix the mistakes caused by the models before it. The models are added one after the other, with greater weight given to data samples that earlier models mispredicted, causing later models to concentrate more on challenging cases. A weighted total of all the models' forecasts makes up the final forecast. By eliminating bias, boosting creates a powerful prediction model from a number of weaker ones. Gradient Boosting Machines (GBM), XGBoost, and AdaBoost are a few examples of boosting algorithms.

Stacking is the process of training several models on the same dataset, then combining their predictions using a different model—often referred to as a blender or meta-model. The meta-model uses the predictions of the basic models, which were trained on the entire dataset, as input characteristics to determine the final prediction. By integrating the predictions of several models in a useful way, stacking can greatly increase forecast accuracy while maximizing the benefits of each underlying model.

Ensemble methods have several advantages:

- Enhanced accuracy. When dealing with intricate issues involving noisy data, ensemble approaches frequently yield better results than single models.
- Decreased overfitting. By averaging the predictions of several models or concentrating on hard-to-predict samples, strategies like bagging and boosting can lower the chance of overfitting.
- Flexibility. Ensemble approaches are applicable to a wide range of machine learning tasks, such as regression and classification.

Ensemble methods have several disadvantages:

- Enhanced complexity. Compared to single models, ensemble models are more difficult to comprehend, apply, and justify.
- Computing cost. Ensemble approaches are more expensive to run since training numerous models takes more time and computer resources.
- Overfitting. Some ensemble techniques, particularly boosting, have the potential to cause overfitting on the training set if not used properly.

In conclusion, by combining different models, clustering methods are effective machine learning tools that can significantly improve the prediction performance. They have several advantages, such as reduced overloading and increased accuracy, but also disadvantages, such as high computational cost and complexity

B. Quantum-inspired methods

Machine learning with quantum inspiration is a recent development that combines classical machine learning techniques with quantum mechanical concepts. Utilizing mathematical algorithms and insights from quantum physics, this multidisciplinary approach seeks to improve the performance and capabilities of machine learning models without the need for actual quantum computing hardware

Quantum-inspired machine learning is characterized by the use of quantum concepts in classical computational frameworks, which separates it from advanced quantum machine learning (QML) It shows current developments and practical and advanced applications in research areas, including dequantized algorithms and tensor network simulations — Strives to provide a comprehensive review of induced machine learning, clarifies its implications, and highlights the potential for further learning

Machine learning inspired by quantum includes new techniques based on digital simulations of certain quantum objects. These approaches provide new ways to train machine learning models, potentially improving productivity and efficiency.

Quantum state discrimination techniques using classical algorithms such as k-nearest neighbors are quantum inspired machine learning mainly used in classification problems This method has shown the ability to reduce complexity and increase accuracy in classification problems

In addition to focusing on real-world applications, the discipline also delves into the concepts underlying quantum-inspired machine learning. This includes studying the theoretical implications of fusing quantum mechanics and machine learning and analyzing how quantum materials can improve machine learning models

Quantum-inspired machine learning is expected to further enrich the discipline as it evolves to incorporate more ideas from traditional machine learning, quantum computation, and quantum mechanics.

Quantum-enhanced machine learning offers new ways to improve machine learning models and solve challenging problems. It's an interesting hybrid of quantum physics and machine learning. Driven by theoretical research and real-world applications, the future of this interdisciplinary field is bright.

IV. Proposed example of classification model for authorship identification

A. General

We proposed method at the nexus of classical and quantum machine learning is an ensemble machine learning classifier that integrates a quantum component and distinguishes and translates features for different kinds of classifiers when summarizing the voting results in the final solution.

Authorship attribution problem is a tuple (A, K, Q), where A is the set of candidate authors, K is the set of reference (known authorship) texts, and Q is the set of unknown authorship texts. For each candidate author $a \in A$, we are given $K_a \subset K$, a set of texts written by a. Each text in Q should be assigned to exactly one $a \in A$. From a text categorization point of view, K is the training corpus and Q is the test corpus.

The approach is showed on Figure 1.

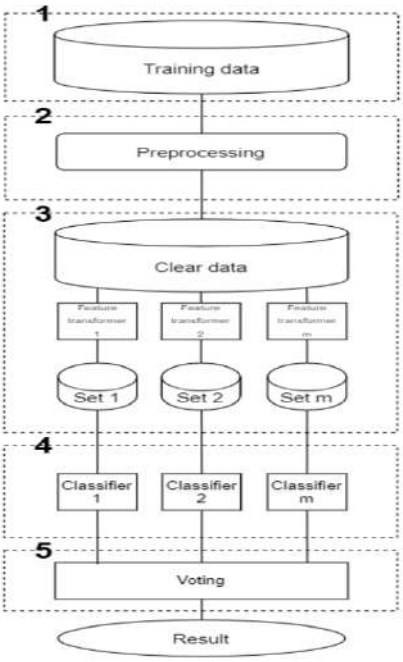

Figure 1. Proposed model architecture.

B. Description

This model has five components. Figure 1 shows them. Also this model has three main attributes.

The first attribute is feature translation and differentiation. This ensemble classifier uses a technique to translate and distinguish characteristics for various classifier types inside the ensemble. By customizing features to each classifier's unique capabilities, the ensemble model's overall prediction power is improved.

So, after preprocessing of training data we take several types of text features (lexical, statistical and so on), create feature sets and submit them to appropriate classifiers. Then we collect the results of the classifiers and pass them to the voting component. This component make decision about text ownership.

The second one is integration of quantum-inspired components. Using quantum principles or methods to process data or reach judgments as a group could constitute the quantum component of this ensemble classifier. By utilizing quantum phenomena this integration of quantum elements seeks to increase classification efficiency and accuracy.

The third one is voting results summarization. Following the predictions generated by each classifier in the ensemble, these results are combined using a voting method. The weighted majority voting algorithm is a technique for compiling expert forecasts. Expert forecasts based on weighted voting are aggregated. In order to get a final conclusion, it combines several forecasts, dynamically adjusting the expert weights according to the accuracy of each prediction.

C. Results

The model was tested in different contexts including general datasets for authorship identification and custom sets of students works.

Table I and Table II show the averaged results.

Table I
Averaged results for general datasets

Number	Classifier	Accuracy
1	Decision tree	0.38
2	K-Nearest Neighbors	0.53
3	Random forest	0.62
4	Gradient boosting	0.51
5	Proposed model	0.68

Table II
Averaged results for students works

Number	Classifier	Accuracy
1	Decision tree	0.41
2	K-Nearest Neighbors	0.46
3	Random forest	0.56
4	Gradient boosting	0.52
5	Proposed model	0.65

V. Multidimensional text analysis which includes semantics

A. General

Using statistical and computational methods, multidimensional text analysis, which includes semantics, examines the connections and patterns between different linguistic elements in a document. This method seeks to capture not just the syntactic and structural aspects of words and phrases, but also their meaning and context by integrating semantic analysis.

Researchers can find co-occurrence patterns of linguistic elements such as sentiment, topic, genre, and authorship style in multidimensional text analysis. This makes it possible to comprehend the text's features more thoroughly and provides subtle insights into the text's meaning and content.

Through the utilization of methodologies such as factor analysis, cluster analysis, and topic modeling, scholars are able to examine texts from several angles, unveiling latent patterns and structures that would not be discernible through conventional text analysis approaches. This approach will be particularly useful in identifying authorship in cases where the semantic component of the text must be taken into account.

B. Perspective of semantic analysis in authorship identification

Automatic NLP actively uses semantic analysis and subject domain ontology. The technique of comprehending natural language literature by the extraction of meaningful data from unstructured sources, such as sentiments, emotions, and context, is known as semantic analysis. It entails dissecting sentences' grammatical construction, including how words, phrases, and clauses are arranged, in order to ascertain the connections between independent concepts within a given context. These approaches are currently under active development and offer a wide range of potential opportunities [4].

A subject domain ontology is a collection of ideas and classifications inside a certain domain that illustrates their characteristics and connections. It is a type of knowledge representation that aids in the arrangement and structuring of data related to a certain area. A common vocabulary and conceptual framework are provided by ontologies for a variety of applications, including NLP.

When subject domain ontology and semantic analysis are combined, NLP systems perform better because the text is understood more accurately. Domain-specific ontologies enable NLP systems to analyze and understand text more effectively within a given context, improving the accuracy and applicability of the findings [5].

For example, in sentiment analysis, topic domain ontology can offer a systematic representation of the domain-specific concepts and relationships, while semantic analysis can assist in comprehending the meaning of words and their context. This may result in the classification and interpretation of sentiment more accurately.

Subject domain ontology and semantic analysis can be crucial parts of NLP systems. They offer an organized depiction of concepts and relationships unique to a given subject and aid in comprehending the content and context of text. NLP systems can analyze and interpret texts more accurately and relevantly by combining these strategies [6].

The applied use of semantic analysis should be considered from several approaches: software and hardware.

VI. Potential semantic analysis approaches: software

Text meaning is taken into consideration throughout the semantic text classification process. It comprehends context, sentiment, and word relationships in addition to just matching keywords. Semantic approach can use different methods.

Bidirectional encoder representations from transformers (BERT), a type of deep learning model, has been found to outperform conventional machine learning techniques in a variety of text classification tasks, such as question answering and sentiment analysis. By taking into account the complete context in which words appear, these models are able to capture intricate semantic patterns in text.

With multi-level semantic features text is classified by combining the extraction of semantic information at several levels, including keywords, local context, and global context. To collect keyword semantic information, for instance, enhanced TF-IDF based on category correlation coefficients can be employed, and neural network models with attention mechanisms, such as TextCNN and BiLSTM, can capture local and global semantic information, respectively.

Text meaning and class descriptions are compared via semantic matching. When there are few labeled examples but thorough class descriptions, this can be especially helpful in few-shot text categorization.

Text classification logic components involve classifying text using logical reasoning and rules.

In rule-based classification text is categorized using a set of pre-established logical rules. For example, the text can be categorized based on the presence or absence of specific keywords or phrases.

Logic programming and machine learning can be combined in inductive logic programming. It can be applied to build classification models that include logical rules representing domain knowledge.

Probabilistic logic models address text classification uncertainty by fusing probability theory and logic. They can make assumptions about a text's propensity to belong in a particular class based on specific feature presence.

More reliable text classification systems can result from the integration of logic and semantic components. For instance, a system may employ logical reasoning to improve categorization based on domain-specific rules after using deep learning to comprehend the semantics of text. Text categorization is known to create issues due to high dimensionality and low density of text representation, which this integration can aid with.

In essence, the application of sophisticated NLP techniques to comprehend the subtle meaning of text and the application of logical reasoning to apply classification rules are the foundations of the semantic and logic components of text classification. This combination makes it possible to create complex models that correctly classify text in a wide range of intricate situations.

For this purpose, a proper review is necessary so that it is possible to understand what features are currently available.

VII. Potential semantic analysis approaches: hardware

A. General

Apart from hardware meant for general-purpose computing and software solutions, there are also specialist hardware solutions made for text processing. The special needs of text processing jobs, like quick data processing, quick response times, and effective resource management, are catered for in these hardware solutions. These are a few instances of specialized text processing hardware.

Integrated circuits created specifically for a given purpose or task are known as Application-Specific Integrated Circuits (ASIC). ASICs may be made to work very well and with minimal power consumption on a variety of text processing applications, including sentiment analysis, keyword extraction, and regular expression matching. Real-time text analysis and NLP are two examples of text processing activities where ASICs can be especially helpful. These tasks call for high-speed data processing and low reaction times.

Field-Programmable Gate Arrays are hardware components that can be programmed to carry out particular functions. Text processing jobs requiring high throughput and low latency, like NLP and real-time text analysis, are especially well suited for FPGAs. Certain text processing activities, such regular expression matching and keyword extraction, can be efficiently and power-efficiently carried out by FPGAs through programming.

Hardware components called Tensor Processing Units (TPUs) are specially made for deep learning and machine learning applications. Matrix multiplications and other operations frequently found in machine learning models, such neural networks, are optimized for TPUs. With its ability to drastically reduce training and inference times for machine learning models, TPUs are especially helpful for large-scale text processing applications like sentiment analysis and text classification [7].

Clusters of computers specifically built for high-performance computing workloads are known as high-performance computing (HPC) clusters. Multiple networked computers, or nodes, each with its own processing capacity and memory make up an HPC cluster. Large-scale text processing jobs requiring high-performance computing resources can be handled by HPC clusters [8].

B. Remote services

Semantic text analysis cloud services offer a strong and adaptable foundation for handling and evaluating massive amounts of text data. These services include a variety of techniques and features, including as sentiment analysis, entity recognition, and NLP, for deriving insights and meaning from text. One example of a cloud service for semantic text analysis is Google Cloud Natural Language. This service recognizes the language of a given document and extracts important phrases using machine learning. Along with emphasizing entity extraction, sentiment analysis, syntax analysis, and categorization, Google's in-depth learning modules fuel the API.

Another cloud-based tool for semantic text analysis is Amazon Comprehend. With the help of machine learning, this service extracts insightful information from language in emails, social media feeds, support requests, documents, and more. By extracting text, important phrases, subjects, sentiment, and more from documents like insurance claims, it also streamlines document processing operations.

VIII. Quantum technologies in semantics

Unlike classical NLP, Quantum NLP represents linguistic aspects utilizing the quantum theory mathematical framework. By using this method, lexical meanings can be encoded as quantum states that can be processed by quantum circuits in simulators or specialized hardware.

The theoretical basis of QNLP is the DisCoCat (categorical distributional compositional) model, which uses string diagrams to convert grammatical structures into quantum processes.

Heavy preprocessing and syntax-dependent architectures are two drawbacks of syntactic analysis in traditional NLP that are addressed by the creation of quantum self-attention neural networks (QSANN). By including a self-attention mechanism, QSANN enhances the scalability and efficacy of quantum neural networks for bigger datasets, and enables their implementation on near-term quantum devices [9].

Sentiment analysis has benefited greatly from the application of QNLP, which has produced flawless test set accuracy in a number of simulations. This demonstrates how quantum computing may lead to major improvements in the comprehension and analysis of textual human emotions.

However, neither traditional nor quantum architecture currently provides a full-fledged semantic platform for solving various tasks in the field of text processing.

The considered promising developments are still partial solutions and require significant modernization.

IX. Conclusion and prospects

Summarizing all the described above, we can formulate a generalized scheme for the proposed approach of intelligent text classifier, including the role of authorship identification when attributing anonymous texts to one or another author (class).

In the scheme, after aggregating the raw data, the texts are sent to three groups for further processing.

In the first one, text processing is performed on familiar hardware using widely used proven classification algorithms such as decision trees, nearest neighbor method and so on.

In the second one text features are encoded in quantum space, then they are transferred to hardware supporting this form of representation (quantum processors or classical devices simulators).

In the third one text is presented in the some semantic model form, and processed by specialized hardware designed for semantic analysis [10]. The semantic model can be described in the form of some abstract agent-oriented model of information processing, for example, using the CS-code [11].

The final decision about text belonging to a class (author) is made taking into account each of the above results.

The approach is showed on Figure 2.

Involvement of semantic and quantum spaces in text classification for authorship identification requires further discussion and testing, as it increases the cost of resources used.

Nevertheless, using the described components in addition to existing established methods, it is possible to increase the accuracy of the final result by analyzing not only lexical and syntactic elements, but also other components of the text. These components that may include hidden borrowings.

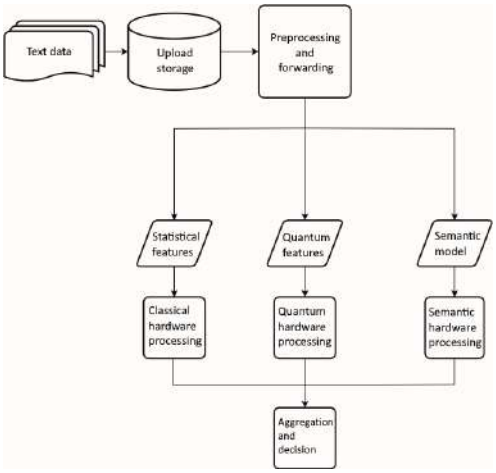

Figure 2. Potential analysing architecture.

References

[1] A. Paramonov, I. Trukhanovich, U. Kuntsevich, "Dynamic features selection in authorship identification problem," Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 309–312, 2021.

[2] M. Gavrilova. Analyzing approaches to plagiarism detection. Molodezhniy nauchno-tekhnicheskiy vestnik [Electronic periodical youth scientific and technical bulletin]. 2015, no. 1, pp. 27–35. (In Russian).

[3] E. Moskalenko, Y. Slesarev, Methods of checking the uniqueness of the text documents. Available at: https://web.snauka.ru/issues/2016/06/69137 (accessed 2023, December).

[4] V. Ivashenko, "Semantic Space Integration of Logical Knowledge Representation and Knowledge Processing Models," Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 95–114, 2023.

[5] D. Shunkevich, "Ontology-based Design of Knowledge Processing Machines," Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 73–94, 2017.

[6] V. Golenkov, N. Gulyakina, "The Main Directions, Problems and Prospects of the Development of the Next-Generation Intelligent Computer Systems and the Corresponding Technology," Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 15–26, 2023.

[7] Tensor Processing Unit : Architecture, Working & Its Applications. Available at: https://www.elprocus.com/tensor-processing-unit/ (accessed 2024, February).

[8] High-performance computing solutions. Available at: https://www.ibm.com/high-performance-computing (accessed 2024, February).

[9] G. Li, X. Zhao and X. Wang, "Quantum Self-Attention Neural Networks for Text Classification," 2022. [Online]. Available: https://arxiv.org/abs/2205.05625

[10] M. Tatur, A. Paramonov, "Open semantic technology as the foundation for new generation intelligent systems," Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], pp. 61–66, 2023.

[11] Vladimir Golenkov, *Technology of Complex Life Cycle Support of Semantically Compatible Next-Generation Intelligent Computer Systems*, V. Golenkov, Ed. Minsk: Bestprint [Bestprint], 2023.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ В ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТА

Труханович И. А., Парамонов А. И.

Рассматривается проблема идентификации авторов текстов. Делается обобщенный обзор текущего состояния проблемы. Предлагается решение на основе модификации ансамблевых методов машинного обучения. Исследуются возможности применения сложных вычислительных методов, таких как обработка естественного языка, алгоритмы машинного обучения и стилометрический анализ, для идентификации отдельных писателей на основе их отличительных лингвистических моделей. Выдвигается гипотеза о необходимости многомерного анализа текста и предлагается новый подход для идентификации авторства текста в виде гибридной интеллектуальной системы.

Received 29.03.2024