RESEARCH AND DEVELOPMENT OF PLAGIARISM DETECTION METHODS BASED ON DEEP NEURAL NETWORKS

Nguyen L.T., master's student gr. 567311

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

German Yu.O. – PhD in Technical Sciences

Annotation. This paper presents a comprehensive review of plagiarism forms in academia, including copy-paste, paraphrasing, mosaic, and idea plagiarism. It analyzes challenges posed by automatic paraphrasing tools, cross-language plagiarism, and multi-source plagiarism. The limitations of traditional detection methods such as string matching, n-gram, and TF-IDF are evaluated. Modern approaches based on deep learning (RNN, LSTM, BERT) and stylometry are discussed, leading to the formulation of a plagiarism detection problem that integrates semantic understanding and stylistic analysis.

Keywords. Plagiarism, TF-IDF, BERT,PhoBERT, RNN, LSTM, RCES, 1D-CNN, PCA, LLM.

In the context of the information explosion and the rapid development of writing tools, plagiarism in academia and content creation is becoming increasingly complex and sophisticated. Plagiarism is not simply the act of copying verbatim, but encompasses many other variations, from interpreting ideas and combining content from multiple sources to using automated tools to alter the surface appearance of the text. Each form of plagiarism poses unique challenges to detection systems, requiring approaches that go beyond traditional lexical matching techniques.

The most common and easily recognizable form of plagiarism is copy-paste plagiarism, in which a writer copies verbatim a passage of text from another source such as books, newspapers, scientific articles, or websites without citing the source. According to Helgesson and Eriksson (2015), plagiarism is defined as the act of taking another person's text and presenting it as one's own, or appropriating another person's ideas and language [1]. In the RCES classification, this is the case where the writer blatantly uses another person's entire work or copies the layout and structure of the text from a single source without alteration [2]. The basic identifying characteristic is the exact match of words, sentence structure, word order, and punctuation. Copied passages often create stylistic inconsistencies, becoming an important indicator when analyzing. Early detection methods such as exact string matching with Knuth-Morris-Pratt, Rabin-Karp, or Boyer-Moore algorithms, text fingerprinting, and n-gram methods have been widely used. Recently, deep learning with one-dimensional convolutional neural networks (1D-CNN) and feedback neural networks (RNN) has shown effectiveness in detecting short copied phrases of 5 to 7 words thanks to its ability to automatically learn local and sequential features [3].

At a more sophisticated level, paraphrasing plagiarism is a form of writing where the writer rephrases another person's ideas using different words and sentence structures but retains the core content without citing the source [1]. According to RCES, this is a case of trying to “disguise” by copying from multiple sources, cross-editing sentences or changing keywords and sentence structures [2]. Online tools like QuillBot help perform paraphrasing quickly and smoothly, posing a major challenge to detection systems that rely solely on vocabulary. The TF-IDF method with its bag-of-words model is completely ineffective because it treats synonyms as independent, causing the vector representations of two sentences with the same meaning but different phrasing to be far apart in space. Detecting this form requires a deep understanding of semantics. Modern deep learning models, especially the Transformer architecture with its self-attention mechanism, have shown superior capabilities. Models like BERT (Bidirectional Encoder Representations from Transformers) and its Vietnamese variant PhoBERT can create sentence representation vectors such that two sentences with the same meaning are close together in vector space, regardless of lexical differences.

An equally complex form of plagiarism is mosaic plagiarism or patchwriting, where the writer combines phrases and sentences from various sources to create a new passage. Each small part may be copied verbatim or slightly modified, creating a textual mosaic. There is no single source to compare it to, and the synthesized passage can flow smoothly if skillfully connected. According to Smodin, this is a form of plagiarism where the writer combines multiple works into one before inserting it into their essay [4]. The most important indicator of plagiarism is the stylistic inconsistency between parts of the same text. Each author has their own stylistic imprint in terms of word usage, sentence structure, sentence openings, and the frequency of prepositions, conjunctions, and pronouns. Stylometry, which studies and explains types of texts in terms of linguistic style and intonation [5], extracts quantitative characteristics such as average sentence length, average word length, type-token ratio, punctuation frequency, and frequency of different word types to create feature vectors for each passage. Euclidean distance, Manhattan distance, and cosine similarity measurements are used for comparison. Principal Component Analysis (PCA) techniques help reduce data dimensionality and visualize paragraphs, highlighting sections with distinct styles that stand out from the main group.

The most sophisticated level is idea plagiarism, where a writer copies the ideas, arguments, and thought structures of others but expresses them entirely in new words, even changing the presentation style. According to the Merriam-Webster definition cited by Smodin, plagiarism is the act of stealing and presenting another person's ideas or words as one's own, or using another person's work without attribution [4]. This is the most difficult form to detect with automated tools because there is no text duplication, and the line between legitimate idea influence and plagiarism is often very thin. Detecting idea plagiarism requires deep semantic understanding and highly abstract idea modeling. While large language models (LLMs) are opening up new approaches, a complete solution is still lacking, often requiring a combination of automated tools and expert evaluation.

The development of automated interpretation tools like QuillBot and Spinbot has created a significant challenge. QuillBot, for example, is considered a pioneering AI application with intelligent processing capabilities, helping millions of people save more than half their writing time by subtly improving wording while retaining the full meaning [6]. These tools can transform source text into new text with different wording and sentence structures while retaining the original meaning, with quality that is increasingly difficult to distinguish from natural writing style. Users simply copy the text into the tool, and within seconds they receive a smooth, interpreted version. This renders lexical matching-based systems obsolete, forcing methods to shift to semantic-based ones.

Furthermore, cross-language plagiarism is becoming increasingly common, especially in international academic environments where writers translate documents from English to Vietnamese and use them without citation. Modern tools such as Smodin Plagiarism Checker support more than 100 languages, allowing plagiarism checking across many different languages [4]. The resulting text may not contain any vocabulary overlap with the original. Detecting this type of plagiarism requires multilingual semantic matching capabilities, far exceeding traditional methods, and necessitates multilingual deep learning models such as mBERT and XLM-Roberta to represent text in the same vector space regardless of language.

Multi-source plagiarism also presents computational and detection challenges. Instead of copying from a single source, the writer takes sections from multiple different sources and combines them. No single pair of texts has a high enough degree of similarity to trigger an alert. Tools like Smodin can scan millions of documents and web pages in seconds to detect duplicate passages [4]. A holistic approach is needed, examining the entire questionable text, searching large databases for similar passages to each section, and using stylistic analysis to detect anomalies caused by the mixing of multiple authorial styles.

Faced with these challenges, traditional plagiarism detection methods reveal inherent limitations. Exact string matching, while fast and accurate against verbatim copy, cannot withstand minor alterations. n-gram methods allow for partial duplication detection but fail to capture semantic relationships, and are unsuitable for interpretations using synonyms or structural changes. The TF-IDF method, which incorporates cosine, is fast and can detect thematic similarity, but it assumes the word bag ignores word order, leading to serious errors (for example, the two sentences "Dog bites person" and "Person bites dog" have the same vector), and it cannot handle synonymy. In general, traditional methods share common limitations: "semantic blindness" (not understanding the true meaning), inability to detect interpretation, dependence on human-defined features, disregard for context, and susceptibility to being deceived by those familiar with their workings.

In this context, modern methods based on deep learning have brought breakthroughs. Reactive neural networks (RNNs) and their variants LSTM and GRU are designed to process sequential data, with a memory mechanism that retains previous information, helping to capture context and dependencies between words in sentences and paragraphs. LSTMs particularly overcome the gradient suppression problem, suitable for classification and prediction problems on time series of indeterminate length [3]. One-dimensional convolutional neural networks (1D-CNNs) are powerful in detecting characteristic local patterns such as copied phrases. Transformer architectures with self-attention mechanisms allow the model to consider the relationship between all words regardless of position, creating a leap forward in semantic understanding. Pre-trained models such as BERT and PhoBERT can generate semantically rich representation vectors, helping to detect interpretive sentences and even paragraphs composed from multiple sources.

Alongside deep learning, stylometry-based approaches remain crucial, particularly in mosaic plagiarism detection. The core principle is that each author possesses a unique stylistic signature, and when a text contains plagiarized passages, those passages bear the imprint of another author, creating statistically detectable anomalies. Feature systems include lexical features (sentence length, word length, part-of-speech ratio), syntactic features (punctuation frequency, word type), content features, and structural features. After extraction, each passage is represented as a feature vector, distance measurements are used for comparison, and PCA helps visualize the anomalies.

The trend of combining methods is gaining attention because no single method effectively handles all forms of plagiarism. The most promising combination is integrating deep learning with semantic understanding and stylometry with anomaly detection capabilities. In a hybrid model, deep learning handles semantic plagiarism detection (interpretation, ideas), while stylistic measurement focuses on mosaic plagiarism detection through stylistic differences between paragraphs. Another approach is a multi-tiered

model, where the first tier uses rapid methods like n-gram or TF-IDF for preliminary screening, and the second tier uses deep learning for in-depth analysis of suspected cases.

Based on the overall analysis of the current situation and challenges, the research problem of this thesis is stated as follows: Given a questionable text D, which is a sequence of sentences, and a large set of reference sources C, the task is to identify the parts of D that are highly likely to be plagiarized from one or more sources in C, and to classify the corresponding form of plagiarism as exact copying, interpretation, or mosaic plagiarism. Specific requirements include: accurate copy detection with high precision; interpretation detection even when wording changes completely but the meaning remains the same; and mosaic plagiarism detection through stylistic anomalies. The system needs to minimize false positives and omissions, and be able to explain the basis of detection to increase persuasiveness in an academic environment. Regarding constraints, the system must be able to handle English with its unique grammatical and lexical characteristics, have an acceptable processing time for medium-length texts, and not require excessively high hardware specifications.

The research hypothesis is that deep learning neural networks, especially LSTM architecture with long-term information retention and sequential data processing [5], combined with other deep learning components, can overcome the limitations of traditional methods thanks to three main capabilities: automatic feature learning from data without manual definition, semantic understanding in specific contexts, and capturing stylistic patterns to detect anomalies. The results of the deep learning neural network testing method are shown in Figure 1.

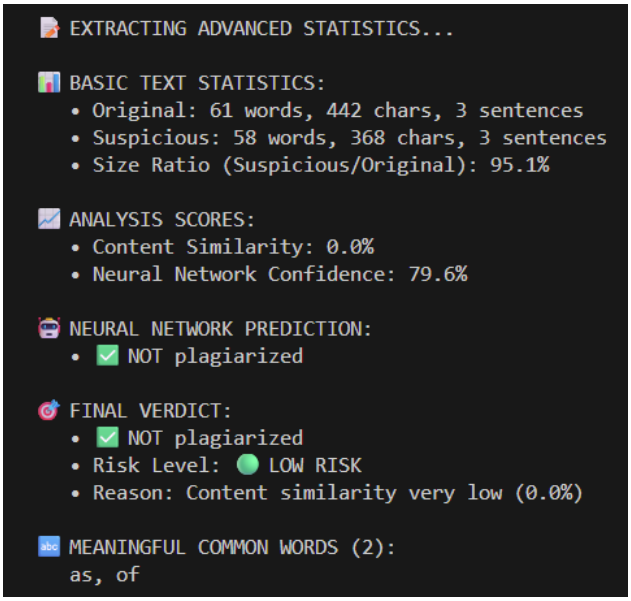

Figure 1 – The results of the deep learning neural network testing method

This paper systematically and comprehensively presents the problem of plagiarism detection in the context of modern academia. The four main forms of plagiarism – verbatim copying, interpretation, mosaicism, and idea plagiarism – are clearly analyzed, along with their identifying characteristics, levels of sophistication, and corresponding detection challenges. The development of automated interpretation tools, cross-language plagiarism, and multi-source plagiarism has shown that traditional methods based on lexical matching (exact strings, n-grams, TF-IDF) are increasingly inadequate due to limitations in “semantic blindness” and the inability to detect linguistic variations. Conversely, modern approaches such as deep learning with RNN, LSTM, and BERT architectures, and style measurement methods, offer promising prospects for more effective solutions thanks to their ability to automatically learn features, understand contextual semantics, and detect stylistic anomalies. Based on that analysis, the paper poses a specific research problem: developing a semantically and stylistically integrated plagiarism detection model capable of processing English texts with high accuracy and classifying different types of plagiarism.

List of Sources:

1. Helgesson, G. and Eriksson, S. Plagiarism in research / G. Helgesson, S. Eriksson // Medicine, Health Care and Philosophy, 2015. – Vol. 18, no. 1. – P. 91-101.

2. RCES. Plagiarism Classification / Research Center for Educational Statistics, 2020.

3. Goodfellow, I., Bengio, Y. and Courville, A. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge : MIT Press, 2016. – P. 367-370.

4. Smodin. Smodin Plagiarism Checker / Smodin LLC, 2023.

5. Trudgill, P. Sociolinguistics: An Introduction to Language and Society / P. Trudgill. – London : Penguin Books, 2000. – P. 1-3.

6. QuillBot. QuillBot Paraphrasing Tool / QuillBot, 2023.