Modern Methods of Text Analysis as an Effective Way to Combat Plagiarism

Serhii Myronenko[1], Yelyzaveta Myronenko[1]
linguist.s.22@gmail.com, lemony.cap12@gmail.com,
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

Summary
The article presents the analysis of modern methods of automatic comparison of original and unoriginal text to detect textual plagiarism. The study covers two types of plagiarism – literal, when plagiarists directly make exact copying of the text without changing anything, and intelligent, using more sophisticated techniques, which are harder to detect due to the text manipulation, like words and signs replacement. Standard techniques related to extrinsic detection are string-based, vector space and semantic-based. The first, most common and most successful target models for detecting literal plagiarism – N-gram and Vector Space are analyzed, and their advantages and disadvantages are evaluated. The most effective target models that allow detecting intelligent plagiarism, particularly identifying paraphrases by measuring the semantic similarity of short components of the text, are investigated. Models using neural network architecture and based on natural language sentence matching approaches such as Densely Interactive Inference Network (DIIN), Bilateral Multi-Perspective Matching (BiMPM) and Bidirectional Encoder Representations from Transformers (BERT) and its family of models are considered. The progress in improving plagiarism detection systems, techniques and related models is summarized. Relevant and urgent problems that remain unresolved in detecting intelligent plagiarism – effective recognition of unoriginal ideas and qualitatively paraphrased text – are outlined.

Keywords:
Literal and intelligent plagiarism, extrinsic detection, techniques, target models, backbone neural architectures.

1. Introduction

Undoubtedly, plagiarism is one of the challenges for the modern research society and the world it lives in, where information technologies are rapidly developing. "Plagiarism" takes its root from the Latin "plagium", which stands for "robbery" or "abduction". Over time it has become more refined, but in essence, remained unchanged [1]. Deliberate appropriation of other people's ideas and intellectual work in any field to publish them as one's work is no different from robbery. Plagiarism, which relates to the IT sphere, is divided into text plagiarism and source code plagiarism [2]. Our research will focus on text plagiarism, particularly on technologies for automatic comparison of original and appropriated text.

Among text plagiarism, there is literal, which consists of exact copying without changes, and intelligent, in which plagiarists try to make changes to the document in a subtle way to disguise the original text. There are several ways of intelligent plagiarism and among them is translation, idea adoption and text manipulation (or so-called paraphrased plagiarism, which is based on paraphrasing the text, replacing the original words, expressions or signs, but retaining the main idea of textual information) are distinguished [3].

All existing plagiarism detection systems are divided into extrinsic and intrinsic. In extrinsic detection, a text document tested for plagiarism is compared with the corpus of the source documents [4]. Intrinsic plagiarism detection does not require analyzing suspicious text to compare it with the original text sources. In this processing of the document, the style of the author's writing is analyzed, and the diversity of vocabulary, i.e., various stylometric features, are used to detect textual plagiarism [5].

The objective of this study was to analyze popular text analysis approaches and determine which one would be efficient in the plagiarism detection system. Paraphrase detection capabilities were of great interest in this research, because apart from the apparent intellectual property infringement detection capabilities, paraphrase detection technology can assist researchers in analyzing already existent advances in their vector of research by identifying the essence of the text and linking it with relevant sources of already concluded studies.

In this study, we will consider techniques related to external detection. Such techniques include: string-based – performs the most straightforward comparison on character level or word level [6]; vector space – compares lexical and syntactic components of the document transferred into vector space [7]; syntax-based – uses the syntactic features of the language in the document [8]; semantic-based – performs the definition of classes of words, synonyms, antonyms, hypernyms and hyponyms [9]; structural based – focuses on the organization of the text, in particular headings, sections, subsections, paragraphs, sentences [10]; citation-based – is

Manuscript received August 5, 2022
Manuscript revised August 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.8.30

characterized by the analysis of documents based on the citations used in the text [11].

The active introduction of technologies for automatic detection of plagiarism began in the 90s of the twentieth century [1, 5]. Over the past thirty years, many detection methods have been implemented to detect literal plagiarism successfully. However, for intelligent plagiarism, in particular paraphrase identification, this area requires further development and implementation of additional paraphrase identification models to solve and improve unresolved issues.

To analyze modern methods of automatic comparison of original and derivative text for detecting literal plagiarism, we will focus on basic techniques such as string-based and vector space and related target models – n-gram and vector space models. To study the methods of detecting intelligent plagiarism, we will consider such techniques as semantic-based and paraphrase identification target models: densely interactive inference network (DIIN), bilateral multi-perspective matching (BiMPM) and bidirectional encoder representations from transformers (BERT).

2. Methods

2.1 Data sources

To study modern methods of text analysis that allows detecting plagiarism, we have chosen two of its types – literal and intelligent [3] to demonstrate the difference between the selected approaches to plagiarism detection. Figure 1 demonstrates techniques of text plagiarism detection both for literal copying and intelligent plagiarism. As it can be seen from Fig.1, more sophisticated approaches are required to detect text manipulation due to the replacement of words and signs.

Among the two existing plagiarism detection systems [4, 5], we have chosen an external one because it allows covering multiple original documents collections and showing the effective direction of the search for suspected plagiarism documents.

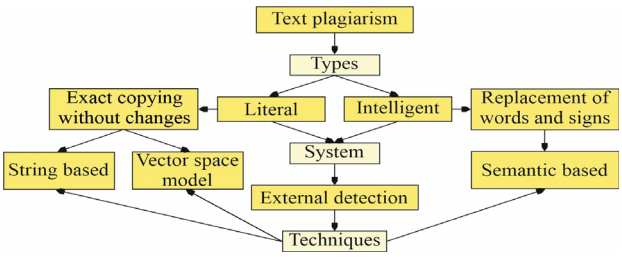

Fig. 1. Classification of researched types, systems and techniques of text plagiarism [1, 3]

In our research, among several existing techniques for literal plagiarism detection, we have chosen the basic ones – string-based and vector space model [6, 7], which later formed the basis of the most modern target models. To analyze effective text analysis methods to detect intelligent plagiarism, we focused on one of the most successful and complex techniques – the semantic-based one [9, 12, 13, 14].

As shown in Table 1, we have selected five target models, existing to date, have been among the most efficient in processing text documents in the detection of text plagiarism: N-gram, Vector Space, Densely Interactive Inference Network (DIIN), Bilateral Multi-Perspective Matching (BiMPM), Bidirectional Encoder Representations from Transformers (BERT).

2.2 Analytical approach

Our research is based on data collected by studying the target models needed to analyze their success in detecting textual plagiarism. In particular, the primary attention was paid to the first basic models and the most modern ones, which use natural language sentence matching technologies and neural architecture. A number of theoretical research methods were used and combined in this work: logical method, analysis, synthesis, classification, generalization and analogy, comparison and collation, induction and interpretation. The methodological basis of the study is formed by modern techniques that focus on lexical, syntactic and semantic textual features and model structures that aim to detect textual plagiarism.

Table 1: The list of researched target models for detection of text plagiarism, the characteristics of their algorithms and efficiency

Text plagiarism	Target Models	Algorithms of the models	The efficiency of the models, %	Reference
Literal: Exact copying without changes	N-gram	Encoding a text document into n-gram profiles and then comparing the original n-grams with n-grams, which could potentially be plagiarism	75–83	[1, 2, 6, 7, 9, 15, 16, 17, 18]
	Vector space	Transformation of words or concepts into a vector. Correlation calculations of term frequency and similarity between sentences by analyzing the vector similarity	75–90	[3, 4, 7, 10, 15, 19, 20, 21]

		of two documents		
Intelligent: Replacement of words and signs	DIIN	Conversion of a text document into high-order n-gram profiles by a neural network encoder. 2-dimensional convolution, word by word revealing of the interaction between pairs of high-order n-grams. It is based on convolutional neural networks	88.6–89.2	[2, 10, 13, 14, 22-25]
	BiMPM	Based on recurrent neural networks. The model uses BiLSTM encoder, which performs conversion of sentences into vectors and comparing them, using the cosine similarity in two directions - the original and the reverse.	88.2–88.8	[14, 19, 22-24, 25]
	BERT	The mask language model, based on transformers, combines masked tokens, and the high-level transformer encoder explores the contextual relationships between words in the text.	90.5–94.3	[12, 14, 21-25]

3. Results

To analyze the successfully implemented methods of detecting text plagiarism, we will focus primarily on the methods used to detect literal plagiarism, its peculiarities and imperfections. Then we move on to intelligent plagiarism, which in contrast to literal plagiarism, has a wide range of text properties and the detection of which requires much more effort and a set of target models. We will consider the advantages and disadvantages of these models.

3.1 Analysis of technologies for automatic comparison of original and unoriginal text to detect literal plagiarism

As shown in Figure 2, literal plagiarism consists of direct copying without changes, uses a monolingual environment and a lexical component of the language – its vocabulary, and syntactic features of the text.

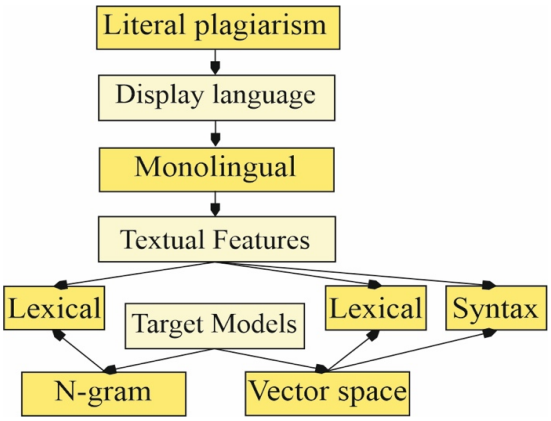

Fig. 2. Main features and target models of literal plagiarism

To analyze the efficiency of detecting literal plagiarism, we will consider the first, most common and most successful target models – N-gram and Vector Space and evaluate their pros and cons.

N-gram Model is a basic method that directly covers the grammatical structure and compares its compliance based on rows. The process takes place in three stages – first, the system searches for potential source documents, then it compares the texts to identify them, and at last, the final processing takes place to identify matches of the document being tested for plagiarism with the original document. First of all, the text document is converted to n-gram profiles, and then the original n-grams are compared with n-grams, which can potentially be plagiarism. The advantage of this method is that it can be used to automatically identify between the source and the document tested for literal plagiarism, but when the size of the document is too cumbersome, there are difficulties with identification. However, numerous experiments that allowed encoding the length of n-gram profiles using various coefficients, such as the Dice coefficient and Jaccard coefficient, have shown their efficiency. They showed that encoding the length of n-grams speeds up the search and does not affect the processing quality of the textual representation of the document. Analysis of research using n-gram technology has shown that its effectiveness in detecting literal plagiarism, which is copy-paste, ranges from 75% to 83%.

Vector Space Model is also one of the common and popular technologies for detecting literal plagiarism, which covers the lexical and syntactic component of the document language. An algebraic model based on the transformation of words or concepts into a vector is used to compare the source document with the document being studied for plagiarism, which allows tracing the relationship of words or phrases in the document. This model can recognize the frequency of terms and detect similarities between sentences in the document by analyzing the vector similarity of the two documents. It has demonstrated its effectiveness in detecting literal plagiarism not only by copy-paste but by partial paraphrasing by 75–90%. However, the vector space model is not effective in detecting sophisticated plagiarism, but it can be used for intelligent plagiarism in combination with other models.

3.2 Analysis of technologies for automatic comparison of original and unoriginal text to detect intelligent plagiarism – replacement of words and signs

Our study will analyze the technology of intelligent plagiarism, which is the replacement of words and signs, the so-called paraphrasing of the original context. As shown in Figure 3, such plagiarism involves monolingual and multilingual environments and the semantic component of language.

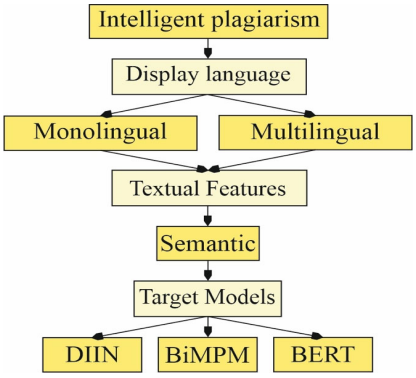

Fig. 3. Main features and target models of intelligent plagiarism

We will analyze the most prominent and successful target models to identify paraphrases and evaluate their positive for and negative sides. Paraphrase identification technologies are based on measuring the semantic similarity of text components – comparing two sentences and identifying the relationship between them. Most language models that target unidirectional modelling have been less effective than models that analyze the bidirectional context in the document. In our study, we will analyze models that are based on natural language sentence matching technology and use neural architecture – Densely Interactive Inference Network (DIIN), Bilateral Multi-Perspective Matching (BiMPM) and Bidirectional Encoder Representations from Transformers (BERT).

Densely Interactive Inference Network (DIIN) Model (Fig. 4). First of all, the text document is encoded in n-gram high-order profiles using a neural network encoder. The architecture of this model is based on 2-dimensional convolution, which step by step clarifies the interaction between pairs of n-grams of a high order. Model DIIN has successfully demonstrated itself in detecting the paraphrases in text documents, especially against the background of such neural models as Embeddings from Language Models (ELMo), Enhanced Sequential Inference Model (ESIM) and Decomposable Attention Model (DecAtt). At the same time, the efficiency indexes in DIIN are among the most modern models, in which the neural architecture is also used – BiMPM and BERT, and are approximately at the same level, amounting to 88.6–89.2%.

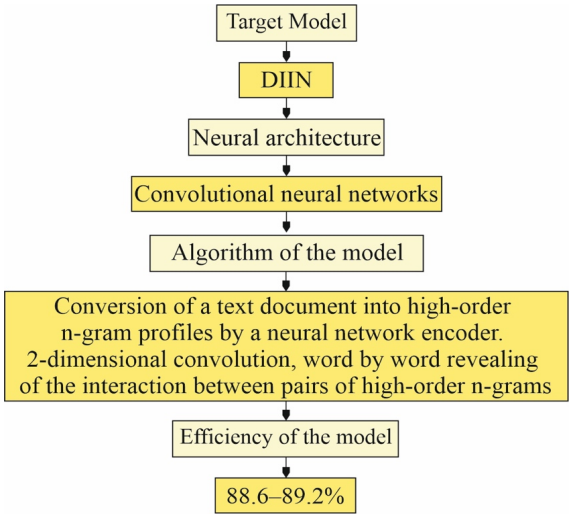

Fig. 4. The main features of the DIIN model

Bilateral Multi-Perspective Matching (BiMPM) Model. BiMPM is one of the most promising models based on natural language sentence matching technology and its main aim is to identify paraphrases. This model works on search, comparison and matching of two sentences (Fig.5). First of all, the original sentences are encoded into vectors using a neural network encoder, and their comparison is made by comparing the cosine similarity between these vectors. The comparison of two sentences takes place both in original and reverse directions. The architecture of the model for comparing two sentences is based on five layers: 1) word representation – building of a d-dimensional vector from words and symbols; 2) context representation – aims to highlight the contextual information contained in

the sentence for each time step; 3) matching layer – belongs to the key layers in BiMPM model, because on this layer there is a comparison of sentences encoded in the vector; 4) aggregation layer – allows interconnecting the analyzed vectors by fixing their length; 5) prediction layer – evaluates the correspondence distribution of two sentences encoded in the vector with a fixed length using a two-layer neural network.

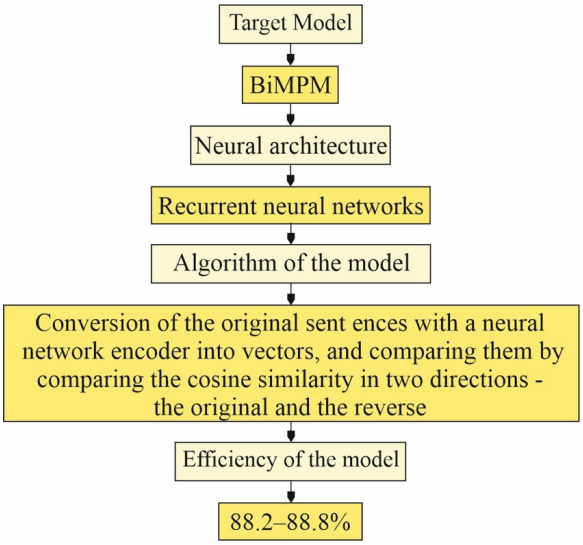

Fig. 5. The main features of the BiMPM model

The multi-perspective cosine function, which uses five layers, has shown significant efficiency in analyzing the coincidence of vectors. The results of testing the BiMPM model for its reliability in detecting and identifying paraphrases showed that among the five models that also perform well at paraphrases identification, such as Siamese-CNN, Multi-Perspective-CNN, Siamese-LSTM, Multi-Perspective-LSTM and LDC (Linguistic Data Consortium), it proved to be the most effective. The reliability index of the BiMPM model is 88.2–88.8%.

The Bidirectional Encoder Representations from Transformers (BERT) Model. The peculiarity of BERT is that, unlike models that perceive the entered text sequentially, i.e. directed. This model, with the help of an encoder, can read the entire sequence of words in a sentence at once. The structure of the studied model is based on a masked language model that combines masked tokens and a high-level Transformer encoder, which explores the contextual relationships between words in the text. The input data is a sequence of preprocessed text represented as token, sentence and positional embeddings. (Fig. 6)

The output data for modelling in BERT are successive vectors of a given size that correspond to input markers with the same indices. Each marker is built by summing the inserted characters, segments and positions. Due to the ability of the high-level Transformer encoder to perform a multifaceted assessment of the textual similarity of embedded words, BERT differs favourably from both unidirectional and bidirectional models of automatic comparison of original and unoriginal text.

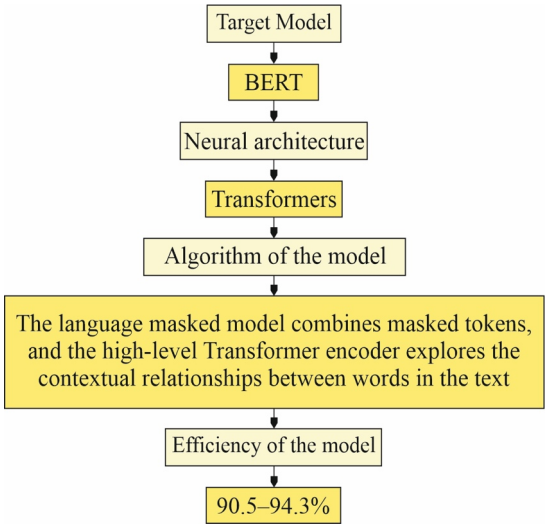

Fig. 6. The main features of the BERT model

In particular, if we compare three the most modern and efficient models for identifying paraphrases that use the neural architecture: BERT, BiMPM and DIIN, the highest percentage of reliability of the model was shown by the BERT family of models. The efficiency of this model reaches 94.3%.

4. Discussion

To date, a number of plagiarism detections system and methods have been developed to help identify not only directly plagiarized fragments of text, which the authors pretend to be the original ones, but there is also an extensive list of text verification systems that can detect intelligent plagiarism [20]. However, as evidenced by numerous studies, modern plagiarism detection systems can successfully detect literal text plagiarism if exact copying of the text without changes or changes inserted into the document is insignificant [7]. Detection of intelligent plagiarism with the help of modern techniques and target models is not that successful [17].

Figure 7 shows the results of plagiarism testing of an extensive array of scientific documents using traditional text-based plagiarism detection systems [26]. These

studies showed that it was possible to identify about 70% of plagiarized documents of plagiarism-tested scientific texts in which plagiarists were directly copying the original text. At the same time, concerning intelligent plagiarism, in particular text manipulation or so-called paraphrased plagiarism, the results obtained are not so optimistic, as among the data set in which the text was deliberately paraphrased, only about 10% of documents were revealed. For texts that have been plagiarized by translation from other languages, the positive figure is even lower – 5%. When testing the documents in which the plagiarists borrowed the idea, the results were quite disappointing because it was not possible to identify among the suspicious texts any document in which the ideas of other authors were used.

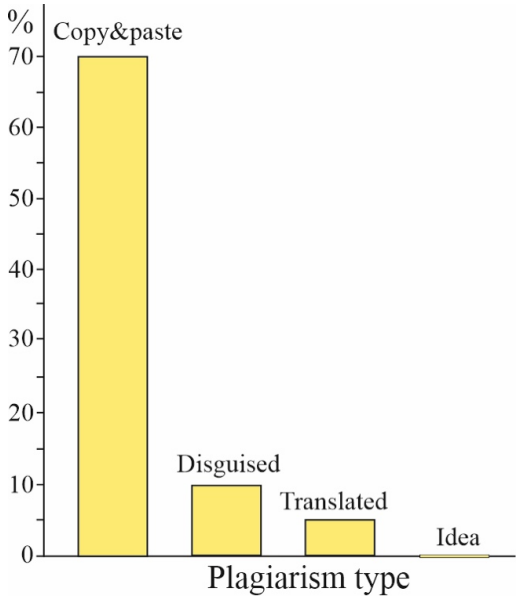

Fig.7. Indicators of plagiarism detection in scientific publications [26]

Thus, the improvement of plagiarism detection systems, techniques and related models is an important task, especially for those dealing with intelligent plagiarism. Studies of basic techniques and models have demonstrated their reliability for detecting near literal plagiarism. Techniques for detecting intelligent plagiarism, particularly those related to text paraphrasing, need further improvement [27]. Our research focused on target models, which have recently proved to be the most effective for solving numerous problems, including identifying paraphrases in the document. They are based on natural language processing models and are essentially similar to the methods used in paraphrasing the text [15, 25]. After all, in computer translation of a text, a plethora of metrics are used to assess the quality of translation from one language to another, where sentences paraphrased due to translation have the same meaning, and their semantic similarity is preserved [10].

The models we analyzed: Densely Interactive Inference Network (DIIN), Bilateral Multi-Perspective Matching (BiMPM) and Bidirectional Encoder Representations from Transformers (BERT) shows efficiency at detecting plagiarism within range of 88–94%, which convincingly testifies to their prospects in solving numerous tasks connected with the detection of literal plagiarism in text documents. BERT and its descendant modifications of model represent family of latest and most successful natural language processing models, officially presented by Google AI researchers in 2018 [12].

5. Conclusions

Analysis of modern automated plagiarism detection technologies has shown that significant progress has been made in the sphere of literal plagiarism detection. Developments over the past thirty years of basic techniques, such as string-based and vector space combined with modern syntax-based, semantic-based, structural based and citation-based ones, have allowed to successfully test for plagiarism not only student works but also journalistic and scientific texts. However, plagiarism detection systems that aim to detect intelligent plagiarism need further improvement. Nowadays, intelligent text plagiarism recognition is not yet consistent, furthermore, there are additional legal challenges in the form of protection of the text checked for plagiarism during the analysis process [28], and in addition to the mentioned points, filling the corpora and hard drives space consumption in the plagiarism detection system itself [28].

Thus, based on the above and the performed research, it should be noted that significant prospects for the detection of plagiarized text belong to target models based on natural language processing models and neural network architecture. Among modern models, a Densely Interactive Inference Network (DIIN), Bilateral Multi-Perspective Matching (BiMPM) and Bidirectional Encoder Representations from Transformers (BERT) and its descendant models should be distinguished, as these models are characterized by the highest efficiency (88–94%) of paraphrase identification in documents. The prominent place among these models belongs to BERT because this family of models can perform bidirectional modelling and is able to read the entire sequence of words in a sentence at once as well as estimate the context from the analyzed sentence. Thus, the analysis of a number of methods showed that the method based on the BERT and its descendant models have a significant advantage over other methods and can be efficiently used in plagiarism detection systems.

References

[1]. Akanksha B., Anukruti A., Tarjni V., Desai S., Nair A.: *A Survey on plagiarism detection*. Advances in computational sciences and technology, 10(8), 2359-2365 (2017).

[2]. Vani K., Gupta D.: *Study on extrinsic text plagiarism detection techniques and tools*. Journal of engineering science and technology review, 9(5), 9-23 (2016).

[3]. Alzahrani S. M., Salim N., Abraham A.: *Understanding plagiarism linguistic patterns, textual features, and detection methods*. IEEE Transactions on systems, man, and cybernetics – Part C: applications and reviews, 42(2), 133-149 (2012).

[4]. Clough P., Stevenson M.: *Developing a corpus of plagiarised short answers*. Language resources and evaluation, 45(1), 5-24 (2011).

[5]. Maurer H., Kappe F., Zaka B.: *Plagiarism – A Survey*. Journal of universal computer science, 12(8), 1050-1084 (2006).

[6]. Gupta D., Vani K., Leema L.M.: *Plagiarism detection in text documents using sentence bounded stop word n-grams*. Journal of engineering science and technology, 11(10), 1403-1420, 2016.

[7]. Thomas S. W., Adams B., Hassan A. E., Blostein D.: *Studying software evolution using topic models*. Science of computer programming 80: 457-479 (2014).

[8]. Bin-Habtoor A. S., Zaher M. A.: *A survey on text plagiarism detection systems*. International journal of computer theory and engineering, 4(2), 185-188 (2012).

[9]. Sánchez-Vega F., Villatoro-Tello E., Montes-y-Gómez M., Pineda L.V., Rosso P.: *Determining and characterizing the reused text for plagiarism detection*. Expert systems with applications, 40(5), 1804-1813 (2013).

[10]. Potthast M., Barrón-Cedeño A., Stein B., Rosso P.: *Cross-language plagiarism detection*. Language resources & evaluation, 45(1), 45-62 (2011).

[11]. Adhya S., Setua S. K.: *Text plagiarism checker using friendship graphs*. International journal of computer science & information technology, 8(4), 13-21 (2016).

[12]. Araseab Y., Tsujiibc J.: *Transfer fine-tuning of BERT with phrasal paraphrases*. Computer speech & language, 66, 101-164 (2021).

[13]. Guu K., Hashimoto T. B., Yonatan Oren Y., Liang P.: *Generating sentences by editing prototypes*. Transactions of the Association for Computational Linguistics, 6, 437-450 (2018).

[14]. Shi Z., Minlie Huang M.: *Robustness to modification with shared words in paraphrase identification*. Association for computational linguistics. Findings of the association for computational linguistics: EMNLP 2020, 164-171, 2020.

[15]. Carvalho N. R., Almeida J. J., Henriques P. R., Varanda M. J.: *From source code identifiers to natural language terms*. Journal of systems and software, 100, 117-128 (2015).

[16]. Chew Y. C., Yoshiki Mikami Y., Nagano R. L.: *Language identification of web pages based on improved n-gram algorithm*. International journal of computer science, 8(3), 47-58 (2011).

[17]. Nahas M. N.: *Survey and comparison between plagiarism detection tools*. American journal of data mining and knowledge discovery, 2(2), 50-53 (2017).

[18]. Peng X., Huang J., Hu Q., Zhang S., Elgammal A., Metaxas D.: *From circle to 3-sphere: Head pose estimation by instance parameterization*. Computer vision and image understanding, 136, 92-102 (2015).

[19]. Amine A., Elberrichi Z., Simonet M.: *Automatic Language Identification: An Alternative Unsupervised Approach Using a New Hybrid Algorithm*. International Journal of Computer Science and Applications, 7(1), 94-107 (2010).

[20]. Arrish S., Afif F. N., Maidorawa A., Salim N.: *Shape-based plagiarism detection for flowchart figures in texts*. International journal of computer science & information technology, 6(1), 113-124 (2014).

[21]. Oberreuter G., Velásquez J.: *Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style*. Expert systems with applications, 40(9), 3756-3763 (2013).

[22]. Ji Y., Eisenstein J.: *Discriminative improvements to distributional sentence similarity*. Proceedings of the 2013 Conference on empirical methods in natural language processing, 891-896. Seattle, Washington, USA (October 18-21), 2013.

[23]. Madnani N., Dorr B. J.: *Generating Phrasal and Sentential Paraphrases: A Survey of data-driven methods*. Computational linguistics, 36(3), 341-387 (2010).

[24]. Nguyen-Son Q., Yusuke Miyao Y., Echizen I.: *Paraphrase detection based on identical phrase and similar word matching*. 29th Pacific Asia conference on language, Information and computation, 504-512. Shanghai, China (October 30-November 1), 2015.

[25]. Vo N. P. A., Popescu O., Magnolini S.: *Paraphrase identification and semantic similarity in Twitter with simple features*. Association for computational linguistics. Proceedings of the Third International Workshop on natural language processing for social media, 10-19, 2015.

[26]. Gipp B., Meuschke N., Beel J.: *Comparative evaluation of text- and citation-based plagiarism detection approaches using GuttenPlag*. In Proceedings of 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11), 255–258. Ottawa, Canada (June 13-17), 2011.

[27]. Adam R., Suharjito M.: *Plagiarism detection algorithm using natural language processing based on grammar analyzing*. Journal of theoretical and applied information technology, 63(1), 168-180 (2014).

[28]. Butakov S., Dyagilev V., Tskhay A.: *Protecting students' intellectual property in the web plagiarism detection process*. The International review of research in open and distributed learning, 13(5), 1-19 (2012).