mathematics

MDPI

Article

Important Arguments Nomination Based on Fuzzy Labeling for Recognizing Plagiarized Semantic Text

Ahmed Hamza Osman * and Hani Moaiteq Aljahdali

Department of Information System, Faculty of Computing and Information Technology in Rabighn, King Abdulaziz University, Jeddah 21911, Saudi Arabia
* Correspondence: ahoahmad@kau.edu.sa

Abstract: Plagiarism is an act of intellectual high treason that damages the whole scholarly endeavor. Many attempts have been undertaken in recent years to identify text document plagiarism. The effectiveness of researchers' suggested strategies to identify plagiarized sections needs to be enhanced, particularly when semantic analysis is involved. The Internet's easy access to and copying of text content is one factor contributing to the growth of plagiarism. The present paper relates generally to text plagiarism detection. It relates more particularly to a method and system for semantic text plagiarism detection based on conceptual matching using semantic role labeling and a fuzzy inference system. We provide an important arguments nomination technique based on the fuzzy labeling method for identifying plagiarized semantic text. The suggested method matches text by assigning a value to each phrase within a sentence semantically. Semantic role labeling has several benefits for constructing semantic arguments for each phrase. The approach proposes nominating for each argument produced by the fuzzy logic to choose key arguments. It has been determined that not all textual arguments affect text plagiarism. The proposed fuzzy labeling method can only choose the most significant arguments, and the results were utilized to calculate similarity. According to the results, the suggested technique outperforms other current plagiarism detection algorithms in terms of recall, precision, and F-measure with the PAN-PC and CS11 human datasets.

Keywords: similarity; plagiarism; semantic; SRL; fuzzy labeling

MSC: 68P20; 68P10; 63E72; 68U15

Citation: Osman, A.H.; Aljahdali, H.M. Important Arguments Nomination Based on Fuzzy Labeling for Recognizing Plagiarized Semantic Text. *Mathematics* **2022**, *10*, 4613. https://doi.org/10.3390/math10234613

Academic Editors: Xiang Li, Shuo Zhang and Wei Zhang

Received: 20 October 2022
Accepted: 30 November 2022
Published: 5 December 2022

1. Introduction

The evolution of, and rapid access to information through, the Internet has contributed to various data protection and ethical problems. "The act of using another person's words or ideas without giving credit to that person" is known as plagiarism. It can generally be considered as anything from basic copy-paste, in which information is simply copied, to higher levels of complexity, in which the text is distorted by sentences, translations, idea adoptions, etc. [1]. Plagiarism could be considered to be more versatile than simple, and more nuanced than trivial copying and pasting. There have typically been the following different forms of plagiarism: straight-line plagiarism, basic footnote pestilence, nuanced footnote plagiarism, plagiarism, quotation-free plagiarism and paraphrasing. The plagiarism problem includes plagiarized media, magazines and Internet tools. Longitudinal research has been undertaken in order to show students' secret patterns of plagiarism, and to analyze academics' experiences of plagiarism. On the other hand, detection of plagiarism tasks can narrowly be divided into two, namely extrinsic detection and intrinsic detection [2–4]. In extrinsic detection, the suspected document is compared with a sample that is either offline or online, whereas if the suspected document is detected internally, structural and stylometric information is used to evaluate this document, which is inserted into a report without any record of a reference source. Many online plagiarism inspectors

use a method that normally consists of World Wide Web surveys, and studies indicate that the most commonly available tools to detect plagiarism cannot detect structural changes and common paraphrases imposed by the users who plagiarise [3]. Our empirical research has shown that university teachers want computerized approaches to detect the plagiarism of ideas. The quality of various academic activities, including theses, dissertations, journal articles, congresses, essays, assignments and so on, is crucial to assess.

A method called paraphrasing may be used to change the organization of a statement or swap some of the original words with synonyms. It is also plagiarism if there are no any correct citations or quotation marks. Due to the changes in fingerprints between the original and copied documents, the approaches utilized in existing detection technologies are unable to identify the use of plagiarism. These cases are far more difficult to identify, since semantic plagiarism is frequently a blurry process that is challenging to find and even more challenging to curtail. One of the key problems in the field of plagiarism detection is how to effectively and efficiently distinguish between plagiarized and non-plagiarized papers [5–7].

Some pirated studies such as [8] at least mention the original version. However, manually checking for plagiarism in a suspicious paper is a very challenging and time-consuming task involving various source materials [9]. It is therefore thought of as a big advance in this regard to use computer systems that can perform the procedure with the least amount of user interaction. The technologies for detecting plagiarism that have been proposed so far are highly capable of catching various types of plagiarism; however, detecting whether plagiarism is present in a text relies on human monitoring [10].

Comparing the copied document with the original document is a common practice in plagiarism detection approaches. Character-matching strategies can be used to identify either completely or partially identical strings. The currently used technique for paraphrasing acquisition uses machine learning and crowdsourcing [11]. This approach focuses on the following two issues: gathering data through crowdsourcing and gathering samples at the passage level. The crowdsourcing paradigm is ineffective without automatic quality assurance and, without crowdsourcing, the cost of building test corpora is too high for practical use. Additionally, a citation-based approach is applied. This technique is employed to detect academic texts that have been read and utilized without citation [11]. The current work offers a method for human semantic plagiarism detection based on conceptual matching and arguments nomination using a fuzzy labeling method, which detects plagiarism through copy-and-paste, rewording or synonym replacement, changing word structure in sentences, and changing sentences from the passive to the active voice and vice versa.

Research addressing the automated detection of suspected plagiarism instances falls under the category of plagiarism detection methods. Methods for examining textual similarity at the lexical, syntactic, and semantic levels, as well as the similarity of non-textual content elements such as citations, illustrations, tables, and mathematical equations, are frequently presented in studies. Research that addresses the evaluation of plagiarism detection algorithms, such as by offering test sets and reporting on performance comparisons, was also examined, as it focuses mostly on gap filling. Studies on the prevention, detection, prosecution, and punishment of plagiarism in educational institutions fall under the category of plagiarism policy. This research analyzed the occurrence of plagiarism at institutions, examined student and teacher attitudes about plagiarism, and discussed the effects of institutional rules.

This research is interrelated and necessary to conduct a thorough analysis of the phenomenon of academic plagiarism. Without a strong structure that directs the investigation and documentation of plagiarism, using plagiarism detection tools in practice will be useless. Research and development efforts for enhancing plagiarism detection methods and systems are guided by the information gained from examining the application of plagiarism detection systems in practice.

In order to keep up with the behavior shifts that plagiarists typically demonstrate when faced with a higher chance of getting caught due to improved detection technologies and harsher techniques, ongoing study is required. Thishis study is one of the methods used to bridge the research gaps in the field of text theft and plagiarism.

The remainder of the sections are as follows: the above-described study on plagiarism detection is described in Section 2. Fuzzy logic is the subject of Section 3. Section 4 provides a detailed explanation of the method's basic concept. In our suggested strategy, we employed an experimental design that is described in Section 5. Section 6 presents the corpus and dataset, as well as similarity detection, and the results and discussion are in Section 7.

2. Related Works

There are two stages to detecting plagiarism: source document retrieval (also known as candidate retrieval) and comprehensive comparison between the source document and the document under examination. In the last five years, many researchers have focused on the retrieval of sources and presented solutions for it, because of the recent breakthroughs in plagiarism detection.

Recently, two approaches to recognizing extrinsic plagiarism were suggested by Arabi and Akbari [12]. Both approaches use two steps of filtering, based on the bag of words (BoW) technique at the document and sentence levels, and plagiarism is only looked into in the outputs of these two stages, in order to reduce the search space. Semantic matrices and two structural are created using a mix of the WordNet ontology and the weighting TF–IDF methodology, as well as the pre-trained network method of words embedding Fast Text. Then, the TF–IDF weighting method is used in the second technique to detect similarities in suspicious documents and sentences.

Research [13] has found that accessing plagiarism sources using external knowledge base sources increased semantic similarity and contextual importance. Other than examples where the text had been duplicated verbatim, the researchers employed a closest neighbor search and support vector machine to find potential candidates for other sorts of plagiarism. Using encoded fingerprints to create queries, a researcher has presented candidate retrieval for Arabic text reuse from online pages and provided the optimal selection of source documents [14]. Cross-lingual candidate retrieval utilizing two-level proximity information was suggested, in addition to prior work on candidate retrieval from the same language. With the suspect (or query) document segmented using a topic-based segmentation algorithm, the researchers next utilized a proximity-based model to find sources related to the segmented portions of the suspicious document. There is still room for improvement in the second phase of plagiarism detection, according to a study of the current trends in plagiarism detection research [12,15,16]. More languages and machine learning approaches need to be explored in the field of cross-language plagiarism detection, as shown by recent studies [17–19].

The detection of disguised plagiarism has been the subject of many studies [20–22]. WordNet-combined semantic similarity metrics were utilized to identify highly obfuscated plagiarism instances in handwritten paraphrases and simulated plagiarism cases [5–7,23–25]. Adding an intermediary phase between candidate retrieval and comprehensive analysis allowed for visual assessment of highly obfuscated plagiarisms, and this additional step included an expanded Jaccard measure to deal with synonyms/hypernyms in text fragments [5]. Studies have examined and evaluated approaches using both content-based and citation-based plagiarism detection in academic writing [26]. Citations and references were shown to be an effective addition to existing plagiarism detection techniques. Document plagiarism detection has been studied as a binary classification task in recent works [27]. Naive Bayes (NB), support vector machine (SVM), and decision tree (DT) classifiers have been used to determine whether or not suspicious-source document pairings included plagiarism [28]. Part of speech (POS) and chunk features were used to extract monolingual features from text pairings, concentrating on modest yet effective syntax-based features.

When compared to traditional baselines, the suggested classifiers were shown to be more accurate in detecting plagiarism in English texts [28]. Genetic algorithms (GA) were utilized to identify disguised plagiarism in the form of summary texts using syntactic and semantic aspects. An algorithm based on the GA method was used to extract concepts at the sentence level [29,30]. Syntactic and semantic elements from the WordNet lexical database were used to include two detection levels, document-level and passage-level [30]. It was shown that a combined syntactic–semantic metric that incorporates additional characteristics such as chunking and POS tagging, as well as semantic role labeling (SRL) and its POS tagging variant, may better distinguish between various kinds of plagiarism. When it comes to spotting veiled plagiarism in a monolingual situation, deeper linguistic traits take center stage.

Paraphrasing is a technique that modifies or replaces some of the original words by their synonym, by changing the structure of the sentence. It is also considered to be plagiarism without a correct citation or quotation marks. Due to variation in the finger printouts between the original and plagiarism files, methods used in existing detection tools cannot be detected as described above. Such cases are much more difficult for people to spot, as linguistic plagiarism is often a smooth process that is difficult to find and more difficult to stop, as it often crosses international borders. Due to the plagiarism issue, there have been a number of arguments, including intellectual property (IP), ethics, legal restrictions and copyright. Intellectual property (IP) is a legal right to the production of the mind, creative and economic, as well as relevant legal fields. In particular, plagiarism is deemed wrong in a moral context, because the plagiarist takes the original author's ideas and contents and tries to deny the author's contribution, by failing to include proper citations or quotations. More legal restrictions are therefore necessary if the original author is to be able to claim their specific rights in respect of their new invention or function. There are many kinds of plagiarism, including copying and pasting, reprocessing and paraphrasing the text, plagiarism of ideas and plagiarism by converting one language to another. Plagiarism is one of the serious problems in education. The discrepancies in fingerprints between the original and the plagiarized material prevent existing detection technologies from detecting plagiarism. Semantic plagiarism is typically a hazy process that is hard to look for and even more difficult to stop, since it generally crosses international borders. The number of arguments picked up by using the fuzzy inference system(FIS) is greater than that detected using the argument weight method in [31]. Fuzzy logic may also tackle the issue of uncertainty in argument selection that affects plagiarized users. One of the most difficult challenges in the world of plagiarism detection is accurately distinguishing between plagiarized and non-plagiarized content. Plagiarism detection software currently uses character matching, n-grams, chunks, and keywords to find inconsistencies. A novel approach of detecting plagiarism is proposed in this paper. Based on semantic role labeling and fuzzy logic, these approaches will be likely to be used in the future.

Natural language processing approaches such as semantic role labeling (SRL), text clustering [32] and text classification [33] have all made use of SRL [34]. Osman et al. have proposed an improved plagiarism detection method based on SRL [5]. The suggested approach was taught to examine the behavior of the plagiarized user using an argument weighting mechanism. Plagiarism detection is not affected by all arguments. Using fuzzy rules and fuzzy inference systems, we are trying to identify the most essential points in a plagiarized text. Fuzzy logic, a kind of approximation reasoning, is a strong tool for decision support and expert systems. It is possible that most of human thinking is based on fuzzy facts, fuzzy connectives, and fuzzy rules of inference [2]. The *t*-test significance procedure was used to demonstrate the validity of the findings acquired utilizing the new method's fuzzy inference system.

The main contributions and goal of the proposed method is a thorough plagiarism detection technique that focuses on many types of detection, including copy–paste plagiarism, rewording or synonym replacement, changing word structure in sentences, and switching from the passive to the active voice and vice versa. The SRL was utilized to

perform semantic analysis on the sentences. The concepts or synonyms for each word in the phrases were extracted using the WordNet thesaurus. These three points are the main differences between our proposed method and other techniques. The second aspect is the comparison process. Whereas prior approaches have concentrated on conventional comparison techniques such as character-based and string matching, our suggested method uses the SRL as a method of analysis and comparison to capture plagiarized meaning of a text. The crucial aspect of this variation is an increase in our suggested method's similarity score employing the fuzzy logic algorithm, where none of the proposed approaches have ever been employed before.

3. Fuzzy Logic System

Many prediction and control systems, fuzzy knowledge management systems, and decision support systems have shown success with the fuzzy logic system [35–37]. For confusing and obscure information, it is often utilized. The connection between inputs and intended outputs of a system may be determined using this technique. Assumptions and approximations may be taken into account while making a choice using it. A defuzzification method and a set of predetermined rules are part of a fuzzy inference system.

Mamdani employed fuzzy logic to regulate a modest laboratory steam engine, which was first described by Zadeh [38]. It is possible to obtain decision-making models in linguistic terms, thanks to an assumption in mathematics about ambiguous reasoning. For many applications and complex control systems, fuzzy logic has recently emerged as one of the most effective methods. More than 1000 industrial commercial fuzzy applications have been successfully created in the last several years, according to Munakata and Jani [39]. Fuzzy logic's distinctive properties are at the heart of the curreendeavour.

The theory of fuzzy sets offers a foundation for the use of fuzzy logic. It became necessary to adapt traditional logic to cope with partial truths, because of its inability to deal with just two values, true and false (neither completely true nor completely false). Thus, the fuzzy logic is an extension of classical logic by generalizing the classical logic inference rules, which are capable of handling approximate reasoning. Each member of a "fuzzy set" has a degree of membership in that set, which is defined by a membership function, which is an extension of the standard "crisp set." Members of the target set are assigned a membership degree between zero and one by the membership function, which allocates a membership degree to each member of the target set [35]. Based on a set of fuzzy "IF–THEN" principles, the computer can convert language statements into actions. Conditions are linked with actions in "if A then B" fuzzy IF–THEN rules, with "if A then B" being the most common version. In the construction and modification of fuzzy logic, the rules are easily understood and simple to alter, to add new rules or to delete current rules.

By applying a membership function to the fuzzy sets of linguistic words, input values are converted into degrees of membership (in the (0;1) range). As shown in Equation (1), the equation concerning $x_i k(x_i)$ may be represented by a fuzzy set, which is obtained by holding certain variables constant μ_i and then transforming that set into a fuzzy one [40].

$$A = \frac{\mu_i k}{x_i} x_i \in X \tag{1}$$

There are fuzzy sets A and X in the world of discourse, and values vary from 1 to 0 in the fuzzy set.

In the context of fuzzification, $k(\ )$ is referred to as the kernel. "A" is the fuzzified version of "A".

$$A = \mu_1 k(x_1) + \mu_2 k(x_2) + \ldots + \mu_n k(x_n) \tag{2}$$

To execute fuzzy reasoning, the inference element of a fuzzy system combines facts collected via the fuzzification process with a set of production rules [40]. The FIS shown in Figure 1.

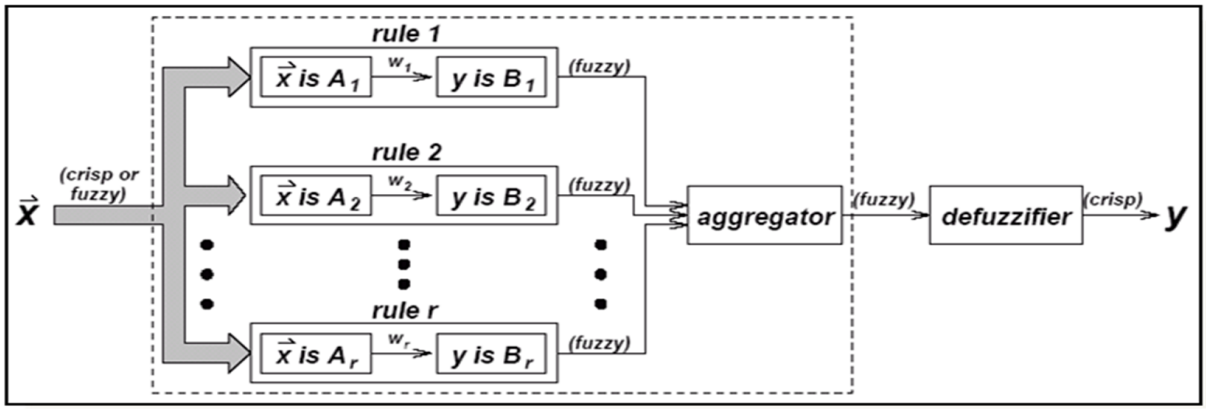

Figure 1. The fuzzy inference system phases.

Figure 1 depicts the fuzzy inference system (FIS) phases. The initial stage in the FIS is to use membership functions contained in the fuzzy knowledge base to transform explicit input into linguistic variables. The fuzzy input is then transformed into a fuzzy output by using an IF–THEN type of fuzzy rule. The final step converts the fuzzy output of the inference engine into a clean output using a membership function similar to that used by fuzzers.

4. Proposed Method

The following are the four key phases of the suggested procedure:

- The first step in data preparation is called pre-processing.
- Segmentation, stop words removal, and stemming
- Extraction of semantic role labeling
- Extraction of concepts
- The fourth step is fuzzy SRL

Figure 2 depicts the suggested method's overall design.
The next sections contain more information on each of these stages.

4.1. Data Preparation

Preparation of the data included text segmentation, stop word removal, and word stemming. Text was segmented into sentences using text segmentation software [41]. To eliminate pointless words, the stop words removal technique was used. Prefixes and suffixes were also removed using the stemming technique to uncover the base word of a term. These words were culled from the text and the rest were discarded. As a result, there may have been a decrease in the similarity of papers.

Text Segmentation: Natural language processing (NLP) relies heavily on pre-processing. Simple text segmentation is a sort of pre-processing in which text is divided into meaningful chunks. Separating text into individual phrases, words, or themes is a common practice. Steps such as information extraction, semantic role labels, syntax parsing, machine translations, and plagiarism detection all rely heavily on this stage. Boundary detection and text segmentation are used to conduct sentence segmentation. This is the most common way to denote the end of a phrase, using a period (.), exclamation point (!), or question mark (?) [42]. The first phase in our suggested text segmentation approach was sentence-based text segmentation, in which the original and comparison texts were broken down into sentence units. Due to our suggested technique's goal of comparing suspicious text with the source, we decided to utilize this method.

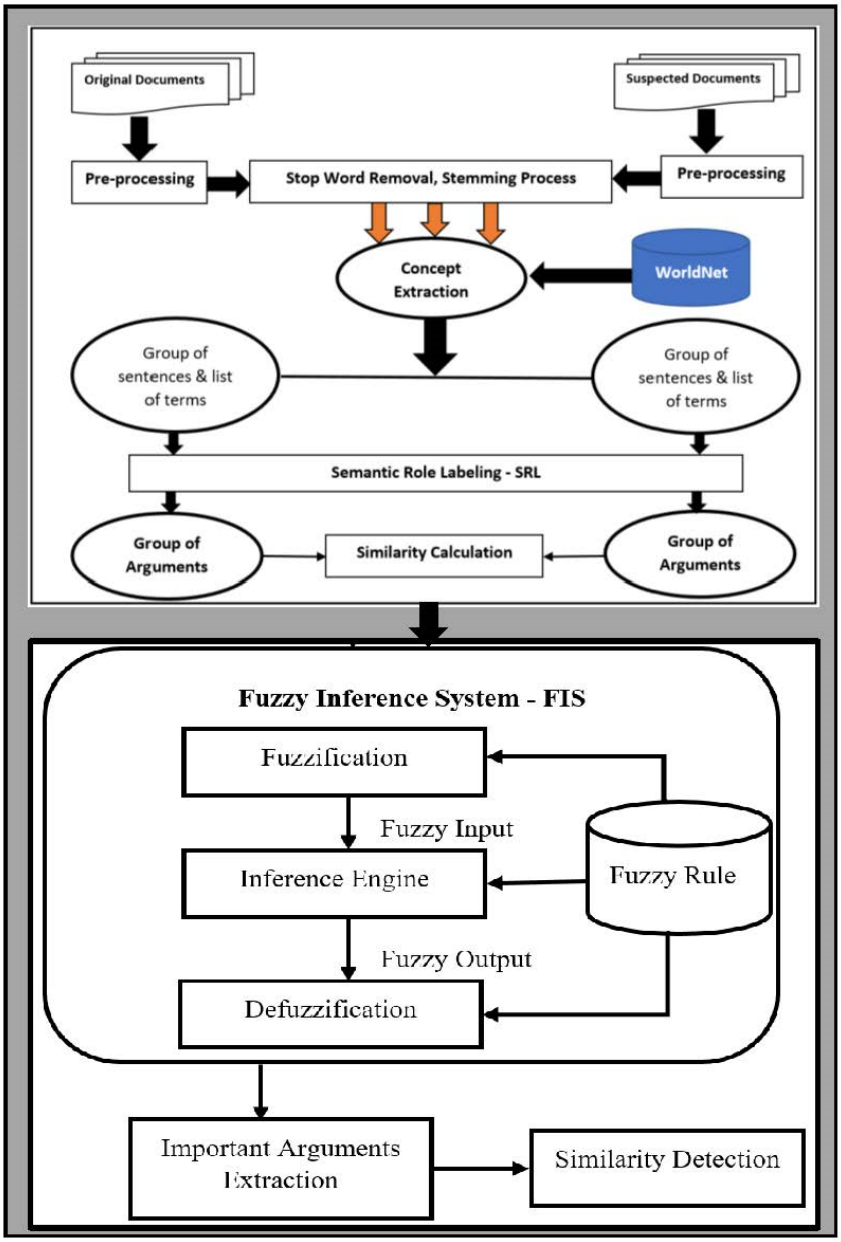

Figure 2. Fuzzy semantic approach to detecting plagiarism.

Stop Words Removal and Stemming Process: Stop words are common occurrences in written materials. Words such as "the", "and", and "a" are examples. As a result of their omission from the index, these keywords have no hint values or meanings associated with their content [43]. According to Tomasic and Garcia-Molina [44], these words account for 40% to 50% of the total text words in a document collection. Automatic indexing may be sped up and index space saved by eliminating stop words, which does not affect retrieval effectiveness [45]. There are a variety of methods for determining stop words, and each has its own advantages and disadvantages. There are a number of English stop word lists now in use in search engines. To make the system work faster, we devised a solution that eliminated all the text's stop words. The SMART information retrieval system at Cornell University, employing the Buckley stop words list [46], is the basis for our suggested technique.

Stemming is another text pre-processing step. Currently, there are several English-language stemmers to choose from that are comprehensive and in-depth. The well-known English stemmers, such as Nice Stemmer, Text Stemmer, and Porter Stemmer, are only a few

examples. A term's inflectional and derivationally related forms are reduced to a generic base form, using the Porter Stemming method. As an example, consider the following:

am, is, are ⇒ be article, articles, article's, articles' ⇒ article

Information retrieval challenges such as word form variations may be addressed with stemming (Lennon et al., 1981). It is not uncommon for a word to be misspelled or a phrase to be shortened or abbreviated, for a variety of reasons.

The stemming process produces a different word n-gram set, which is then used for similarity matching between texts using the proposed method.

4.2. Extraction of Arguments and Semantic Role Labeling

Semantic role labeling is a technique for identifying and classifying arguments in a piece of writing. Essentially, a semantic analysis of a text identifies all of its other concepts' arguments. In addition to determining "subject", "object", "verb", or "adverb", it may also be used to characterize elements of speech. Each word in the suspected and source sentences is labeled with its matching role throughout the roles labeling procedure. As a result of this research, semantic role labeling based on the sentence level was offered as a unique plagiarism detection approach.

Using semantic role labeling (SRL), a method for comparing the semantic similarity of two papers, one may determine whether the ideas in both documents are arranged similarly. In this research, ideas were labeled with role labels and gathered into groups. Groups were employed in this manner to as a fast guide to collect the suspicious portion of the text. An example of a plagiarized situation is found here:

Example (1):

Chester kicked the ball. (Original), the ball was kicked by Chester. (Suspected)

By using the Online Demo of SRL (http://cogcomp.cs.illinois.edu/page/research, accessed on 22 July 2022), the produced arguments are:

An original ("Chester kicked the ball") and suspected ("The ball was kicked by Chester") phrases analysis using SRL are shown in Figures 3 and 4:

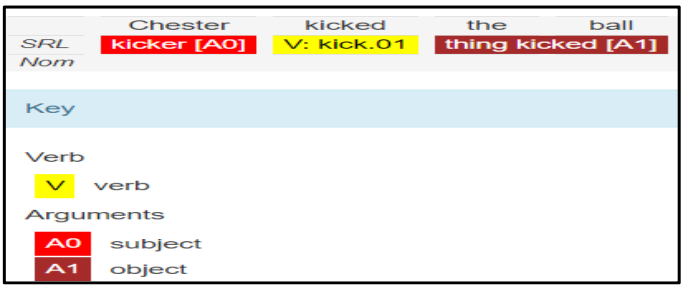

Figure 3. An analysis of the original phrase based on SRL.

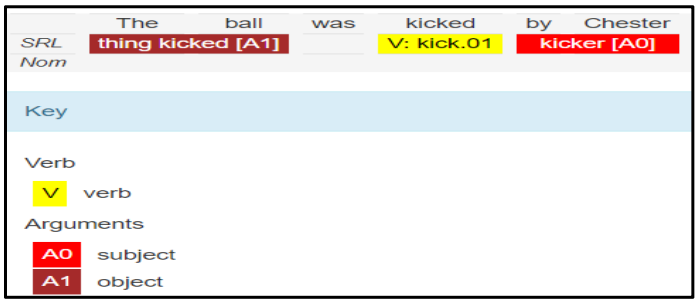

Figure 4. An analysis of the suspected phrase based on SRL.

The syntax of the two phrases above may alter depending on whether the active or passive voice is employed if synonyms and antonyms are utilized. In fact, the semantics of these two phrases are quite similar. In spite of the labels being moved about inside

the sentences, the SRL still manages to capture the arguments (subject, object, verb, and indirect object) for a sentence. Our suggested approach of plagiarism detection, based on a comparison of the sentence's arguments, is supported by this capture.

According to the SRL scheme of similarity [47], original and suspected papers were checked for similar keywords. When two words are found to be the same, we go straight to the argument label and compare the phrases in which they are conveyed. After identifying potential plagiarized phrases, this phase compares the argument labels of those sentences with the argument labels of the original phrases. In order to make an accurate comparison, the words must be compared correctly. The plagiarism ratio may be incorrect if we compare the phrases in Arg0 (subject) in the suspected text with all other arguments in the original text. For example, comparing the subject with the adjective argument (Arg–Adj) to the subject with the time argument (Arg–TMP) is an unfair general-purpose argument (Arg–O).

String matching [48,49] and n-gram [15] are two examples of approaches that compare each word in a suspected sentence to the original phrase. The terms "ball", "kicked", and "Chester" will be compared. Aside from the fact that this comparison is incorrect, it also consumes comparison time. Our technique compares the reasons in the suspected sentence phrase to those in the original phrase to see whether they are comparable. Subject to subject comparisons, verb to verb comparisons, etc., are all possible with our suggested SRL technique. This will reduce the number of comparisons we have to make. No comparison will be made between arguments in questionable papers and arguments from the actual source materials. When comparing original and suspected phrases, we can see that active and passive synonyms have different structures and term positions when compared to their passive counterparts, as seen in this example. These two phrases are, in fact, semantically interchangeable. The researchers have found that, despite altering synonyms within phrases, their technique managed to capture the semantic meaning of a statement. Using the WordNet concept extraction, our suggested approach of plagiarism detection may be supported.

4.3. Concept Extraction

The extraction of concepts is an important part of our detection process. WordNet [50] is used in this research as a source of synonyms and related words. It is one of the lexical semantic connections, which are relationships between words. The WordNet system quantifies semantic similarity, since the closer two words are to one another, the more similar is the structure of their connection, and the more frequent are the lexical units shared between them. Using WordNet Thesaurus as a starting point, we begin the process of identifying key concepts. The following are the steps in the procedure: Using the WordNet synset (synonym set) from the words used in the text of the document, the document's terms are mapped onto the WordNet Thesaurus database. WordNet is structured, based on the concept of synsets. A synonym set is a collection of words or phrases that have the same meaning in a certain context. Using an example of a synset from the WordNet Thesaurus database, we can better illustrate what have said so far regarding idea extraction.

Figure 5 demonstrates a synsets extraction from the terms "Canine" and "Chap" [paper: A semantic approach for text clustering using WordNet and lexical chains] and [paper title: Comparative cluster labeling involving external text sources].

In the above example, the synset of terms "Canine" and "Chap", the phrase "canine" for example, may refer to a variety of different things depending on the context: feline, carnivore, automobile, mammal, placental, and many more. The hyponymy (between specialized and more general ideas) and meronymy (between parts and holes) are examples of semantic interactions that might connect synsets together. Figure 5 provides an example of synset relations using WordNet database.

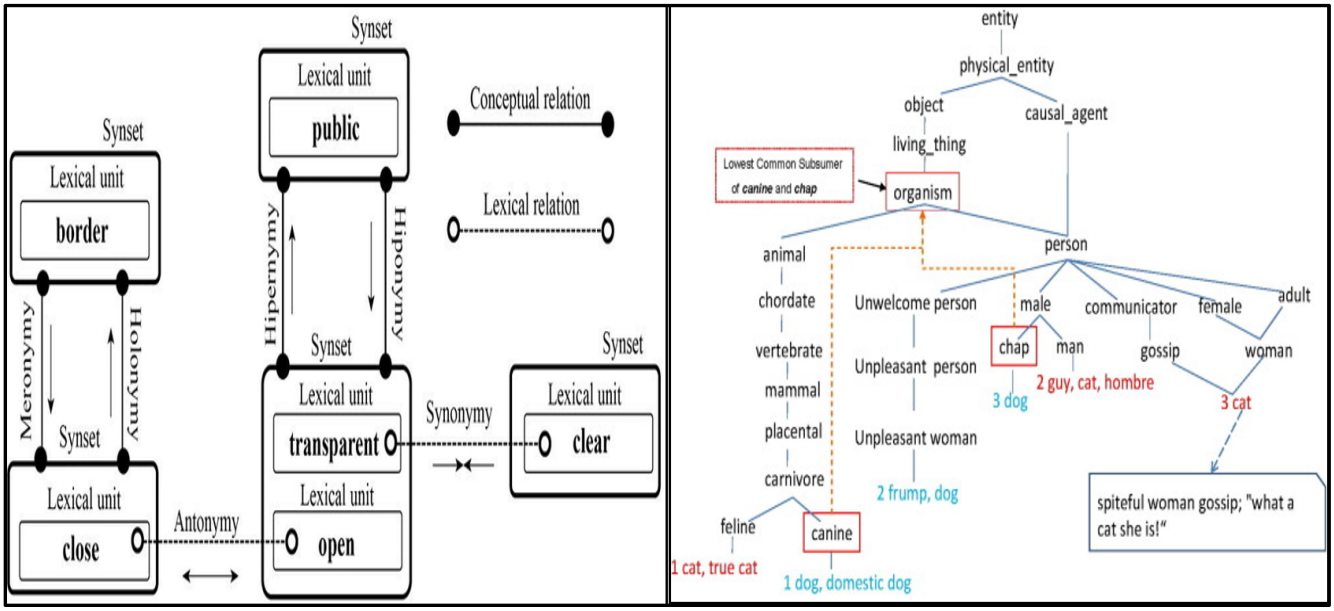

Figure 5. Terms synsets extraction.

5. Fuzzy SRL

Semantic role labeling is a method for detecting plagiarism by comparing two phrases' semantic similarities. In this section, the concept of the suggested approach is detailed.

The SRL similarity metric introduced and explained in [28] is used to determine the argument similarity score. For plagiarism detection, fuzzy is utilized as an argument selection technique to choose the most relevant arguments.

Using a fuzzy decision-making framework, Vieira [34] suggested a fuzzy criterion for feature selection. Classical multi-objective optimization has the challenge of balancing the weights of many objectives; our technique does not have that problem. Fuzzy logic elements were used into our system to determine the degree of resemblance between the suspect and the source documents. In the FIS system, we created a feature vector for each sentence (S) as follows: S = A F1, A F2,..., where A F1 represents the first argument feature, and so on. By comparing the documents, we may infer their values. Following the fuzzy logic approach, the arguments score is formed, and then a final set of high-scoring arguments is selected to combine with the similarity detection based on the comparison. Algorithm 1 below outlines the phases of our new technique.

Stemming, on the other hand, is a text pre-processing technique. Stemming is a solution to the issue of word form variation in information retrieval [34]. Spelling mistakes, alternative spellings, multi-word constructions, transliteration, affixes, and abbreviations are the most prevalent kinds of variation. Matching algorithms suffer from a lack of efficiency because of the wide range of word forms used in the information retrieval process. Using root words in pattern matching improves information retrieval significantly. This phase was completed using a Porter stemming strategy [35]. Extracting the most important words from a piece of text is an important part of our suggested technique. Because of this, our suggested method's ability to detect similarity between papers may suffer.

Algorithm 1 An improved plagiarism detection method based on fuzzy logic.

Step	Main Process	Process Detail
1	*Read original document O and suspected documents S:*	Read the original document O and suspected documents S, O and S = *{Title, S1,S2,S3, . . . ,Sn}*
2	*Apply SRL cross the original document O and suspected documents S:*	Extract all the arguments for each sentence. Collect all the similar argument in separated node.
3	*Preprocessing:*	Extract the individual sentences of the documents. Then, remove stop words. The last step for preprocessing is word stemming.
4	*Arguments similarity score calculation:*	Calculate the similarity between each corresponding argument (Verb with Verb, Subject with Subject . . . etc)
5	*Perform sentence score using Fuzzy Logic Method:*	
	A. *Construct the membership function as fuzzification:*	Define the meaning (linguistic variable) of input/output terms and determine fuzzy set used in the fuzzy inference system as described in Section 7.
	B. *Construct the fuzzy IF-THEN rules:*	Define the possible fuzzy IF-THEN rules as described in Section 7.
	C. *Defuzzification:*	Convert the fuzzy output from the inference system into a crisp output (the high score more than 0.5).
	6. Test the results before and after optimization.	Use T-test significant test to show if there is a significant improvement or not.

5.1. Membership Functions and Inference System

A fuzzy system relies on the ability to make inferences. In order to perform fuzzy reasoning, the data gathered through the fuzzification process are combined with a sequence of production rules [34]. To translate numerical data into linguistic variables and execute reasoning, fuzzy expert systems and fuzzy controllers need preset membership functions and fuzzy inference rules [35]. The magnitude of each input's involvement is represented graphically by the membership function.

Fuzzy logic-based plagiarism detection was implemented with different inputs, using a similarity score between individual arguments of original sentences and suspected sentences and one output value of y, which is a similarity score between all arguments of the original and suspected sentences. This was done in order to implement our proposed method. To demonstrate Mamdani's fuzzy inference given a collection of fuzzy rules, the goal is to provide an example. To represent each individual linguistic variable, there are many kinds of "membership functions" on the inputs and outputs in this system. The linguistic variables for x and y, for example, comprise significant and insignificant components that must be considered. These functions are shown in Figure 6.

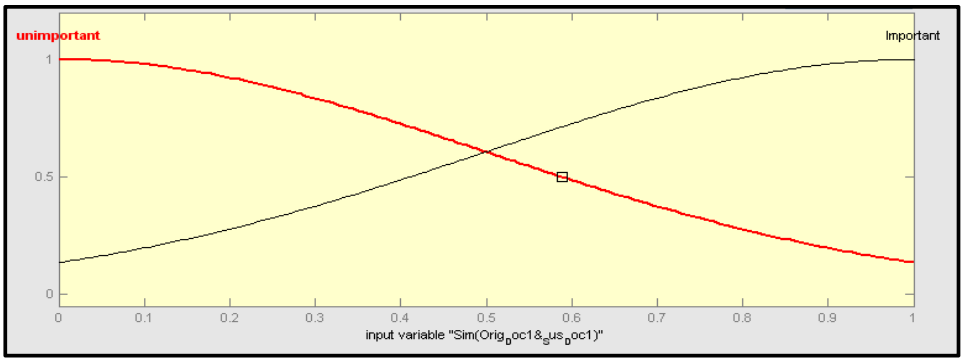

Figure 6. Input MF for fuzzy model.

There were two linguistic values assigned to each input in the suggested method, "important" and "unimportant". Similarity scores were calculated for input and output using the fuzzy membership function, which yielded important and unimportant scores, depending on whether the score was larger or smaller than 0.5. FIS Toolbox in MATLAB was used to calculate the membership function. Using this toolkit, non-linear processes with fuzzy rules created automatically in the FIS environment may be perfectly modelled. All of this information was entered into a computer program that determined the answer. Each rule in the system was seen as crucial to the generation of numerical forecasts. Although each argument's similarity score was used to reflect its input value, an overall score was used to show how similar each argument was to other likely suspect phrases. Section 5 explains how the arguments' similarity was determined.

5.2. Fuzzy IF–THEN Rules Construction

When dealing with an inference engine, a good understanding of the fuzzification rules is critical. The fuzzification rules base comprising the IF–THEN rules generates the linguistic parameters for the middle and yield variables outlined above. This set of IF–THEN rules extracts the most significant arguments based on our criterion. Based on the input characteristics, a popular approach for constructing rules was used to extract and create all available rules. The following equation was used to obtain the total number of rules:

$$R = f^n \tag{3}$$

where R denotes the rules; f denotes the features input; n signifies the rule's logic of possibility.

For example, in a a five-input system with two logic outputs for each input (true and false), the total number of rules created is 32. All potential rules to help the inference system to distinguish between significant and unimportant arguments were generated by Equation (3) using our suggested technique.

Our suggested technique was put to the test with over 1000 papers, yielding a massive number of rules. Although it was difficult to capture all of the created rules in the fuzzy system, it was a crucial concern. It was imperative that the number of rules created could be reduced. This issue was addressed using a mix of rule reduction techniques [34]. These rules reflect the membership function's inputs and outputs. There were around 1000 papers that were considered to be relevant arguments in the training data set for the proposed technique. Figure 7 depicts the three-dimensional fuzzy rule graphs of our suggested technique.

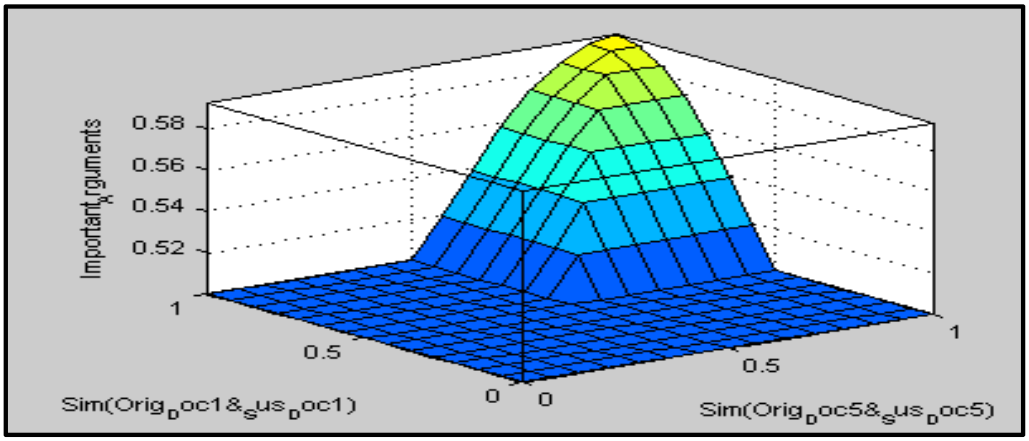

Figure 7. Important argument similarity using fuzzy model.

In order to improve the detection of similarities, the suggested method aims to choose only the strongest reasons that may have a significant impact on the plagiarism process. The fuzzy IF–THEN rule base is an important part of FIS. Prior to reducing the number of rules, all available rules were retrieved. The most essential arguments were chosen for the second round of testing comparisons. Arguments that were deemed insignificant by the FIS were not taken into consideration. After deciding on the arguments, a test was carried out. The degree of similarity relies on the number of reasons retrieved from the sentences, therefore lowering the insignificant arguments leads to an increase in the similarity score, as was discovered when comparing the findings from the first test. The CS11 human plagiarism corpus was used to obtain the matching score. In the next sections, the calculation of similarity is detailed.

5.3. Defuzzification

Defuzzification is the final phase in the fuzzy logic procedure. A final score is assigned to each argument during defuzzification, based on the inference system findings. A fuzzy set's aggregate output is utilized as the input and the outcome is a single value. Defuzzification must be finished before a single value output may be generated. According to Mogharreban [34], there are several defuzzification methods. Fuzzy reasoning systems benefit from our usage of the maximum mean defuzzification approach.

The maximum mean: The mean of maxima is computed using the distribution of output to get a single value. The equation below shows how this is done:

$$\frac{\sum_{j-1}^{q} Z_j u_c(Z_j)}{\sum_{j-1}^{q} u_c(Z_j)} \tag{4}$$

$$\sum_{j-1}^{1} \frac{Z_j}{j} \tag{5}$$

where I is the time when the distribution output hits the maximum level of $z\,j$, z is the mean of maximum, and $z\,j$ is the membership function's maximum point.

6. Experimental Design and Dataset

Experiments were conducted to determine how many sentences from the original papers were found to be plagiarized. The tests were carried out on a PAN-PC dataset [2]. According to the PAN-PC dataset [2], each of these texts was based on one or more original parts. It was decided to use the new method by looking for allegedly suspicious original texts. There were many groupings of texts, each with a particular quantity of types. When comparing the two groups, and the number of texts for each group, the first set consisted of five texts. Then, five more texts were added to the initial group, followed by 10, 20, 40, 100, and 1000. Grouping is a useful technique [5] for identifying how a plagiarized argument performs under various conditions. A total of 1000 documents were used in the studies after analyzing the arguments' behavioral patterns. As input variables in FIS, each group and each argument was selected. This results in a tally of how closely these individuals are related to one another. The input variable's values are a similarity score for each pair of arguments that are comparable. As part of the data training, the trials were carried out on the PAN-PC dataset. After that, it was put to the test on a large sample size of 1000 documents. It was discovered that important arguments may be picked using FIS. After the arguments were picked, a second round of testing was conducted. The degree of similarity was discovered to be dependent on the number of reasons retrieved from the sentences, and by lowering the unimportant arguments, the similarity score was found to be higher, as was the case with the initial testing. It was subsequently determined that the PAN-PC dataset was utilized to cross-check the results. Below, we explain how that number was determined.

The CS11 human corpus was used in an additional experiment. The problem with the PAN-PC corpus is that most of the plagiarism instances were intentionally manufactured. There are 100 instances of plagiarism in the CS11 people short answer questions corpus, according to Clough and Stevenson [36]. Examples of plagiarized texts of varying degrees of plagiarization may be found in this resource. Since the Clough and Stevenson corpus was created and built by real people rather than computer programs, it provides a more realistic picture of the actions of people who have copied work. Each document in the corpus has at least one suspiciously copied section, as well as five original sentences taken from Wikipedia. Native and non-native speaking students were asked to respond to five questions, based on the original materials. Answers were based on the instructions provided by the corpus designers, with the exception of non-plagiarized examples, and were based on actual texts with varying degrees of text overlap. Average word counts for the short sections were in the tens of words (200–300). Near-copy (19), heavy revision (19), and light revision (19) instances were found in 57 samples, while the remaining 38 samples were found to be plagiarism-free. The following are examples of questionable documents:

- Near-copy: it focuses on copying and pasting from the source text
- Light-revision: Minor alterations to the original text, such as substituting synonyms for phrases and introducing grammatical changes
- Heavy-revision: editing and rewriting in original material with restructuring and paraphrasing
- Non plagiarism: participant information was included into the writings without altering the originals

The matching arguments and the arguments included in the sentences are both taken into account when determining similarity. When comparing the two documents, the first variable identifies arguments that are similar in both, while the second identifies arguments that do not appear in either text. The Jaccard coefficient was used to determine the matching among the arguments in the original and the suspected texts.

$$similarity(C_i(ArgT_j, ArgT_k)) = \frac{C(ArgT_j) \cap C(ArgT_k)}{C(ArgT_j) \cup C(ArgT_k)} \quad (6)$$

where *Ci (ArgTk)* = ideas of the original document's argument text; *C(ArgTj)* = concepts of the suspected document's argument text.

Using the following equation, we estimated similarity between the original and the suspicious texts:

$$\mathrm{TS}(\mathrm{txt1}, \mathrm{txt2}) = \sum_{i=1,l} \sum_{\substack{j=1,m \\ k=1,n}} SimC_i(ArgT_j, ArgT_k) \quad (7)$$

where TS is the total similarity score, m = the number of arguments text in the original document, n = the number of arguments text in the suspected document, and i = the matching between the arguments text in the original text with concept i and the suspected text with that concept, along with the number of concepts.

7. Results and Discussion

Plagiarized materials were copied and pasted, synonyms were changed, and sentences were restructured in a variety of ways (paraphrasing). Three typical testing measures for plagiarism detection were utilized as described in the Equations (8)–(10).

$$\mathrm{Recall} = \frac{(\text{No of detected args})}{(\text{Total no of args})} \quad (8)$$

$$\mathrm{Precision} = \frac{(\text{No of plagirized Args})}{(\text{No of detected Args})} \quad (9)$$

$$\text{F} - \text{measure} = \frac{(2 \times \text{Recall} \times \text{Precision})}{(\text{Recall} + \text{precision})} \tag{10}$$

Using the collection of documents we chose, we ran the tests shown in Figure 8, which displays the findings. For the similarity computation, a set of documents is represented by a row.

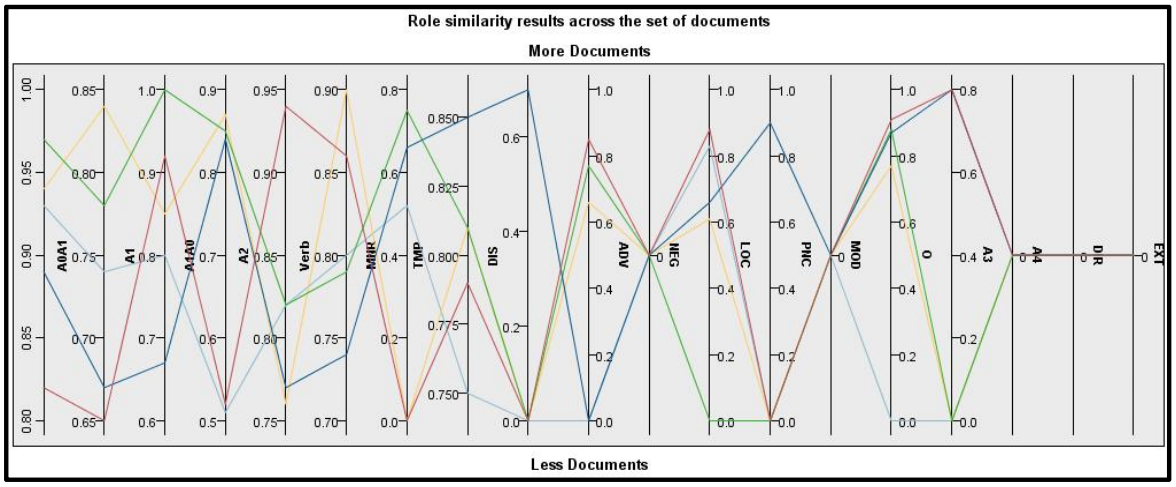

Figure 8. Role similarity scores across the set of documents.

The SRL was employed to break down the text into different arguments, examples of which are shown in Table 1 below.

Table 1. Argument types and their descriptions.

Type	Description
Arg0	agent
Arg1	direct object/theme/patient
Arg2–5	not fixed
V	verb
MNR	manner
TMP	time
DIS	discourse connectives
ADV	adjective
NEG	negation marker
LOC	location
PNC	purpose
MOD	modal verb
O	general purpose
DIR	direction
EXT	extent

Experiments employed a variety of argument types and descriptions, as shown in Table 1.

Each pair of documents is represented in Table 2 by the percentage of similarity between the suspected and original documents. Recall, precision, and F-measures all have scores over 0.58, whereas all recall measures have scores above 0.80. Table 2 shows that the scores are all larger than 0.5, which indicates that the findings are excellent, but it was still possible to enhance these scores to obtain better similarity values.

Table 2. Ranking of SRL arguments using FIS.

Fuzzy Result	O	LOC	Verb	A0	A2	A1	TMP	MNR	A0A1	PNC	A1A0	DIR	ADV	DIS	NEG	MOD	A3	A4
Fuzzy Result of 5 Documents	0.6403	0.6332	0.6278	0.6277	0.6254	0.6065	0.6009	0.597	0.5936	0.5	0.4866	0.3935	0.3614	0.3591	0.3591	0.3591	0.3591	0.3591
Fuzzy Result of 10 Documents	0.6403	0.6332	0.6277	0.6254	0.6065	0.6009	0.597	0.5936	0.5	0.5	0.4866	0.3935	0.3614	0.3591	0.3591	0.3591	0.3591	0.3591
Fuzzy Result of 20 Documents	0.5287	0.5251	0.5233	0.5186	0.5143	0.5082	0.5	0.5	0.5	0.5	0.5	0.4541	0.4411	0.4411	0.4411	0.4364	0.4357	0.4324
Fuzzy Result of 40 Documents	0.5409	0.5361	0.5359	0.5353	0.5347	0.5309	0.526	0.5252	0.5194	0.5078	0.4948	0.4749	0.4675	0.4671	0.4671	0.466	0.4632	0.4616
Fuzzy Result of 100 Documents	0.5068	0.5057	0.5055	0.5046	0.5045	0.5044	0.5041	0.5039	0.5038	0.5	0.5	0.4941	0.494	0.4922	0.4919	0.4919	0.4919	0.4918
Fuzzy Result of 1000 Documents	0.5017	0.5012	0.5012	0.5011	0.5011	0.501	0.501	0.5	0.4998	0.4996	0.4985	0.4985	0.498	0.4979	0.4979	0.4979	0.4979	0.4978
Rank	**significant**	**significant**	**significant**	**significant**	**significant**	**significant**	**significant**	**significant**	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant	Not significant

FIS has shown that the writer who plagiarizes does not concentrate on all of the reasons in a statement, therefore certain arguments are left out. Arguments like this are said to be insignificant. Table 2 shows the outcomes of the FIS cross SRL sentences.

Table 2 shows the ranking of the SRL arguments using FIS. In order to test our method, we used a variety of groupings of documents (5, 10, 20, 40, 100 and 1000). These allegedly plagiarized texts used a variety of plagiarism strategies, including copying and pasting, swapping certain phrases for their counterparts, and altering sentence structure (paraphrasing). There are two kinds of arguments. Both types of arguments have a similarity score larger than 0.5; however, the first form of argument is considered significant while the second type is considered irrelevant. Similarity scores were calculated for input and output using the fuzzy membership function, which yielded important and unimportant scores, depending on whether the score was larger or smaller than 0.5. The comparison step of the proposed technique uses a Jaccard similarity measure [37] with a threshold value of 0.5 [38–40]. For this reason, we chose 0.5 as our cutoff value. In order to enhance the similarity score, the FIS method chose the most essential reasons. On the other hand, the similarity score was reduced by minor arguments. Unimportant arguments were discarded to minimize the general resemblance of the original and suspected texts. To determine the degree to which two arguments are similar, the SRL similarity measure, developed by Osman et al. [28], is used. A table titled the "similarity scores table" shows all of the similarity ratings between the various arguments. An input to the FIS is the similarity score table. Features include the arguments and overall similarity between them, as well as the amount of original and suspected texts in the dataset that were utilized in its construction. This system's goal is to increase the similarity scores in plagiarism detection by generating many key arguments.

A common approach used by those who plagiarize is to concentrate on key phrases and then adapt their work to include them. Only the most crucial points that have a significant impact on the reader would be reworked. There are a number of target selection approaches available, all of which aim to anticipate as accurately as possible the essential objectives of the data. FIS is one of these approaches. Statistical significance test (*t*-test) results were used to demonstrate the benefits of the new strategy. These findings are shown in Table 3 and demonstrate the statistical significance of the suggested approach.

Table 3. Statistical significance testing using the *t*-test.

	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**		**Sig. (2-tailed)**
				Lower	**Upper**	
Recall-1–Recall-2	9.542×10^{-2}	3.348×10^{-2}	1.367×10^{-2}	6.028×10^{-2}	0.1306	0.000928
Precision-1–Precision-2	0.2238	5.402×10^{-2}	2.206×10^{-2}	0.1671	0.2805	0.000159
F-measure-1–Fmeasure-2	0.1521	3.661×10^{-2}	1.494×10^{-2}	0.1137	0.1906	0.000156

There are many metrics in Table 3 that may be compared using the pair of variables before and after optimization using the FIS-SRL approach, as well as their significance, using the paired samples *t*-test process. Comparing the means of two variables representing the same group at various points in time is done using the paired samples *t*-test technique. In the pair of variables statistics table, the mean values of the two variables ((Recall-1, Recall-2); (Precision-1, Precision-2); and (F-measure-1, F-measure-2)) are shown. As a paired samples *t*-test examines two variables' mean values, it is important to know their averages. The *t*-test may be used to determine whether there is a significant difference between two variables if the significance value is less than 0.05. For example, it was found that the suggested technique had significant recall (0.000928), precision (0.000159), and F-measure results in the significance field of Table 3 (Sig. (2-tailed)). This suggests that

the proposed method had significant results in all three areas. The fact that the confidence interval for the mean difference does not include 0 shows that the difference is, likewise, significant. There is also a lack of statistical significance in the F-measure, recall, and precision. Comparison of the outcomes before and after optimization indicates that there is considerable difference.

The PAN-PC dataset evaluates and compares the presented solution with other plagiarism detection systems. Figure 9 shows the comparison findings.

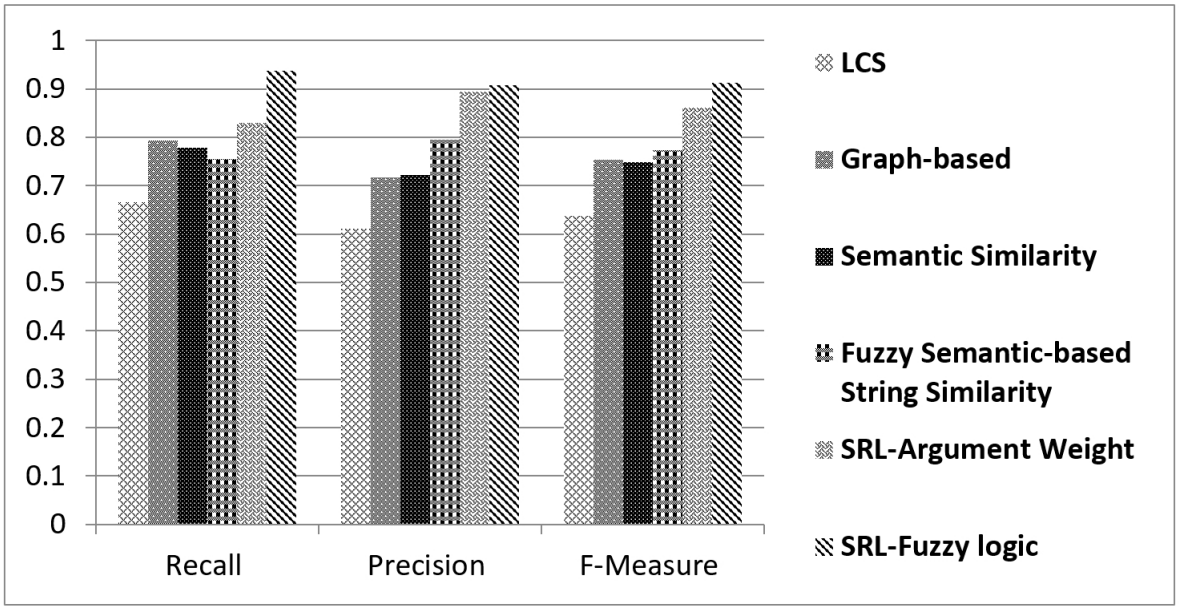

Figure 9. Contrast between the current text-based similarity detection methods.

A comparison of SRL fuzzy logic with string-similarity, LCS, graph-based methods, semantic similarity methods, and SRL argument weight methods is shown in Figure 9 [29,30,34–36]. The similarity results were shown to be improved with our new technique.

Our suggested technique was compared to Chong's naive Bayes classifier [51] with a set of all features, best features, and ferret baseline method in the Tables 4–6 for the similarity classes (heavy-revision, light-revision, and near-copy). Table 4 shows the heavy-revision plagiarism class.

Table 4. Heavy plagiarism class.

Plagiarism Detection Method	Average Recall	Average Precision	Average *F*-Measure
Naïve-Bayes Method with all features	0.333	0.211	0.258
Naïve-Bayes Method with the best features	0.667	0.526	0.588
Naïve-Bayes Method with Ferret Baseline	0.615	0.421	0.5
Fuzzy-SRL-Method	0.713	0.796	0.746

Table 5. Light plagiarism class.

Plagiarism Detection Method	Average Recall	Average Precision	Average *F*-Measure
Naïve-Bayes Method with all features	0.44	0.579	0.5
Naïve-Bayes Method with the best features	0.55	0.579	0.564
Naïve-Bayes Method with Ferret Baseline	0.419	0.684	0.52
Fuzzy-SRL-Method	0.725	0.809	0.760

Table 6. Cut-and-paste plagiarism class.

Plagiarism Detection Method	Average Recall	Average Precision	Average *F*-Measure
Naïve-Bayes Method with all features	0.267	0.211	0.235
Naïve-Bayes Method with the best features	0.5	0.474	0.486
Naïve-Bayes Method with Ferret Baseline	0.5	0.211	0.296
Fuzzy-SRL-Method	0.935	0.741	0.827

Table 5 compares the proposed technique to previous methods, based on a mild plagiarism class. Recall, precision, and F-measure were all found to be the best for the suggested technique. Table 5 shows the results for the light-revision plagiarism class.

Table 6 shows an assessment of the suggested approach and other methods on the on copy-and-paste class. We observed that the suggested technique had the highest scores for F-measure, recall, and precision.

In addition, the amount of time it takes to complete a task is also taken into consideration. This metric is often used to evaluate the effectiveness of algorithms. The temporal complexity of the suggested approach was used to assess its suitability. The suggested method was found to be in the same class as the rest of the methods. There are several plagiarism detection methods in this class, according to research by Maxim Mozgovoy and JPlag [34,35]. Even so, they observed that certain plagiarism detection algorithms have a time complexity of O(f(n)N2), where f(n) is the time it takes to compare a pair of files with a length of n and N is the collection size (number of files). Time-consuming techniques, such as fuzzy semantic comparison and semantic-based string similarity, were compared. It was demonstrated that semantic-based string similarity, LCS, and semantic-based similarity all have the same level of temporal complexity as the proposed method. Table 7 displays the results, in terms of how long each type of method takes.

Table 7. Time complexity comparison.

Algorithm	Time Complexity
Fuzzy Semantic-based String Similarity	$\mathbf{O(n^2)}$
Longest Common Subsequence (LCS)	$\mathbf{O(n^3)}$
Semantic-based Similarity	$\mathbf{O(n^2)}$
SRL-Argument Weight	$\mathbf{O(n^2)}$
Graph-based Method	$\mathbf{O(V+E)}$
Sentence-based Natural Language	$\mathbf{O(n^2)}$
SRL-Fuzzy Logic	$\mathbf{O(n^2)}$

On the other hand, the string similarity-based fuzzy semantic method, semantic similarity, the similarity based on SRL method, the similarity based on graph-based representation method, and similarity based on sentence-NLP all have higher temporal complexity than ours, as shown in Table 7. The findings reveal that the suggested technique falls within a category of detection algorithms that is generally recognized. There are three major differences between our suggested approach and previous methods:

When it comes to copying and pasting, rewording or replacing words, changing the voice of a phrase from active to passive or vice versa, or changing the word structure in phrases, are all instances of plagiarism that may be caught using the method we provide.

In contrast to earlier methods, which focused on more traditional comparison techniques like character-based and string matching, the SRL is used as a comparison mechanism to analyze and compare text to identify instances of plagiarism. According to our results on the PAN PC-09 dataset, we are able to outperform other methods for detecting plagiarism, such as longest common subsequence [52], graph-based method [31], fuzzy semantic-based string similarity [49], and semantic-based similarity [53]., Additionally, we

found that our technique outperforms other methods described by Chong [51], including naive Bayes classifier and ferret baseline, on CS11 corpora.

8. Conclusions and Future Work

The current study offers a plagiarism detection system that includes the following steps: the first and second documents are uploaded into a database, where the text is processed to be segmented into sentences, stop words are eliminated, and words are stemmed to their original forms. Next, the processed text is parsed in each document to find any arguments within, and then each argument found is represented as a member of a group, to determine how similar the groups of text are to one another. To select the best arguments from the text, the FIS has been applied. For plagiarism detection, semantic role labeling may be utilized by extracting the arguments of sentences and comparing the arguments. A FIS has been used to choose the arguments that have the most impact. When performing the similarity calculation, only the most essential reasons were taken into consideration, thanks to the use of FIS. The standard datasets for human plagiarism detection (CS-11) have been tested. In comparison to fuzzy semantic-based string similarity, LCS, and semantic-based approaches, the suggested approach has been proven to perform better.

A common approach used by those who plagiarize is to concentrate on key phrases and then adapt their work to include them. The proposed method proved that crucial points that have a significant impact on the reader should be reworked. The study aimed to anticipate, as accurately as possible, the essential objectives of the data. The results of statistical significance tests demonstrated the impact and benefits of the new strategy, compared with methods of plagiarism detection based on other strategies.

The limitation of this research must also be emphasized. This research did not cover some types of plagiarism, such as the similarity of non-textual content elements, citations, illustrations, tables, and mathematical equations, and these types are frequently discussed in studies.

To conclude, the methods of paraphrase type identification suggested in this research can be used and improved in a wide range of academic contexts. This involves not only support in identifying plagiarism, but also a focus on upholding ethical academic conduct. Genetic algorithms may be used to improve the results that can be produced, by employing the FIS in future. In addition, the above-mentioned limitation of this study is still considered as a research gap, which will be filled in the future.

Author Contributions: Conceptualization, A.H.O. and H.M.A.; methodology, A.H.O.; validation, A.H.O. and H.M.A.; formal analysis, A.H.O.; investigation, A.H.O.; resources, A.H.O.; data curation, A.H.O. and H.M.A.; writing—original draft preparation, A.H.O. and H.M.A.; writing—review and editing, A.H.O. and H.M.A.; visualization, A.H.O.; supervision, A.H.O.; project administration, A.H.O.; funding acquisition, A.H.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by the Institutional Fund Projects under grant no. (IFPIP: 481-830-1443). The authors gratefully acknowledge the technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research work was funded by the Institutional Fund Projects under grant no. (IFPIP: 481-830-1443). The authors gratefully acknowledge the technical and financial support provided by the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Potthast, M.; Stein, B.; Barrón-Cedeño, A.; Rosso, P. An evaluation framework for plagiarism detection. In *Coling 2010: Posters*; Coling 2010 Organizing Committee: Beijing, China, 2010.
2. Potthast, M.; Stein, B.; Eiselt, A.; Barron-Cedeno, A.; Rosso, P. Overview of the 1st International Competition on Plagiarism Detection. In Proceedings of the PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, San Sebastian, Spain, 10 September 2009; Available online: CEUR-WS.org (accessed on 18 June 2022).
3. Mozgovoy, M.; Kakkonen, T.; Cosma, G. Automatic Student Plagiarism Detection: Future Perspectives. *J. Educ. Comput. Res.* **2010**, *43*, 511–531. [CrossRef]
4. Kakkonen, T.; Mozgovoy, M. An Evaluation of Web Plagiarism Detection Systems for Student Essays. In Proceedings of the Sixteenth International Conference on Computers in Education, Taipei, Taiwan, 27–31 October 2008.
5. Osman, A.H.; Salim, N.; Binwahlan, M.S.; Alteeb, R.; Abuobieda, A. An improved plagiarism detection scheme based on semantic role labeling. *Appl. Soft Comput.* **2012**, *12*, 1493–1502. [CrossRef]
6. Osman, A.H.; Barukab, O.M. SVM significant role selection method for improving semantic text plagiarism detection. *Int. J. Adv. Appl. Sci.* **2017**, *4*, 112–122. [CrossRef]
7. Osman, A.H.; Salim, N.; Elhadi, A.A.E. A tree-based conceptual matching for plagiarism detection. In Proceedings of the 2013 International Conference on Computing, Electrical and Electronic Engineering, Khartoum, Sudan, 26–28 August 2013.
8. Foltýnek, T.; Meuschke, N.; Gipp, B. Academic plagiarism detection: A systematic literature review. *ACM Comput. Surv. CSUR* **2019**, *52*, 1–42. [CrossRef]
9. Lovepreet, V.G.; Kumar, R. Survey on Plagiarism Detection Systems and Their Comparison. In *Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018*; Springer: Odisha, India, 2019.
10. Gillam, L.; Vartapetiance, A. From English to Persian: Conversion of Text Alignment for Plagiarism Detection. In Proceedings of the Working notes of FIRE 2016—Forum for Information Retrieval Evaluation, Kolkata, India, 7–10 December 2016.
11. Burrows, S.; Potthast, M.; Stein, B. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 1–21. [CrossRef]
12. Arabi, H.; Akbari, M. Improving plagiarism detection in text document using hybrid weighted similarity. *Expert Syst. Appl.* **2022**, *207*, 118034. [CrossRef]
13. Alzahrani, S.; Aljuaid, H. Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 1110–1123. [CrossRef]
14. Lulu, L.; Belkhouche, B.; Harous, S. Candidate document retrieval for Arabic-based text reuse detection on the web. In Proceedings of the 2016 12th International Conference on Innovations in Information Technology (IIT), Abu Dhabi, United Arab Emirates, 28–30 November 2016.
15. Yalcin, K.; Cicekli, I.; Ercan, G. An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding. *Expert Syst. Appl.* **2022**, *197*, 116677. [CrossRef]
16. Chang, C.-Y.; Lee, S.-J.; Wu, C.-H.; Liu, C.-F.; Liu, C.-K. Using word semantic concepts for plagiarism detection in text documents. *Inf. Retr.* **2021**, *24*, 298–321. [CrossRef]
17. Bohra, A.; Barwar, N. A Deep Learning Approach for Plagiarism Detection System Using BERT. In *Congress on Intelligent Systems. Lecture Notes on Data Engineering and Communications Technologies*; Springer: Singapore, 2022.
18. Alotaibi, N.; Joy, M. English-Arabic Cross-language Plagiarism Detection. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1–3 September 2021.
19. Roostaee, M.; Fakhrahmad, S.M.; Sadreddini, M.H. Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection. *Expert Syst. Appl.* **2020**, *160*, 113718. [CrossRef]
20. Al-Shamery, E.S.; Gheni, H.Q. Plagiarism Detection using Semantic Analysis. *Indian J. Sci. Technol.* **2016**, *9*, 1–8. [CrossRef]
21. Cader, J.M.A.; Cader, A.J.M.A.; Gamaarachchi, H.; Ragel, R.G. Optimization of Plagiarism Detection using Vector Space Model on CUDA Architecture. *Int. J. Innov. Comput. Appl.* **2022**, *13*, 232–244. [CrossRef]
22. Guillén-Nieto, V. Plagiarism Detection: Methodological Approaches. In *Language as Evidence*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 321–372.
23. Osman, A.H.; Aljahdali, H.M. Role Term-Based Semantic Similarity Technique for Idea Plagiarism Detection. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 475–484. [CrossRef]
24. Osman, A.H.; Salim, N. An improved semantic plagiarism detection scheme based on chi-squared automatic interaction detection. In Proceedings of the 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), Khartoum, Sudan, 26–28 August 2013.
25. Osman, A.H.; Salim, N.; Binwahlan, M.; Twaha, S.; Kumar, Y.J.; Abobieda, A. Plagiarism detection scheme based on semantic role labeling. In Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, 24–26 July 2012.
26. Gipp, B. Citation-Based Document Similarity. In *Citation-Based Plagiarism Detection*; Springer: Wiesbaden, Germany, 2014; pp. 43–55.
27. Luo, L.; Ming, J.; Wu, D.; Liu, P.; Zhu, S. Semantics-based obfuscation-resilient binary code similarity comparison with applications to software and algorithm plagiarism detection. *IEEE Trans. Softw. Eng.* **2017**, *43*, 1157–1177. [CrossRef]

28. Amini, P.; Ahmadinia, H.; Poorolajal, J.; Amiri, M.M. Evaluating the High Risk Groups for Suicide: A Comparison of Logistic Regression, Support Vector Machine, Decision Tree and Artificial Neural Network. *Iran. J. Public Health* **2016**, *45*, 1179–1187.
29. Pajić, E.; Ljubović, V. Improving plagiarism detection using genetic algorithm. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019.
30. Vani, K.; Gupta, D. Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. *Expert Syst. Appl.* **2017**, *73*, 11–26.
31. Osman, A.H.; Salim, N.; Binwahlan, M.; Hentably, H.; Ali, A.M. Conceptual similarity and graph-based method for plagiarism detection. *J. Theor. Appl. Inf. Technol.* **2011**, *32*, 135–145.
32. Krishna, S.M.; Bhavani, S.D. An efficient approach for text clustering based on frequent itemsets. *Eur. J. Sci. Res.* **2010**, *42*, 385–396.
33. Suanmali, L.; Salim, N.; Binwahlan, M.S. Automatic Text Summarization Using Feature-Based Fuzzy Extraction. *J. Teknol. Mklm.* **2009**, *2*, 105–155.
34. Shehata, S.; Karray, F.; Kamel, M.S. An Efficient Model for Enhancing Text Categorization Using Sentence Semantics. *Comput. Intell.* **2010**, *26*, 215–231. [CrossRef]
35. Baruah, H.K. The theory of fuzzy sets: Beliefs and realities. *Int. J. Energy Inf. Commun.* **2011**, *2*, 1–22.
36. Guribie, F.L.; Owusu-Manu, D.-G.; Badu, E.; Edwards, D.J. Fuzzy synthetic evaluation of the systemic obstacles to personalizing knowledge flows within and across projects. *Constr. Innov.* **2022**. [CrossRef]
37. Jiskani, I.M.; Cai, Q.; Zhou, W.; Lu, X.; Shah, S.A.A. An integrated fuzzy decision support system for analyzing challenges and pathways to promote green and climate smart mining. *Expert Syst. Appl.* **2021**, *188*, 116062. [CrossRef]
38. Zadeh, L.A. Fuzzy sets. *Inf. Control.* **1965**, *8*, 338–353. [CrossRef]
39. Munakata, T.; Jani, Y. Fuzzy systems: An overview. *Commun. ACM* **1994**, *37*, 68–76. [CrossRef]
40. Ibrahim, A. *Fuzzy Logic for Embedded Systems Applications*; Newnes: Oxford, UK; Elsevier: Berkeley, CA, USA, 2004.
41. Ma, W.; Tran, D.; Sharma, D. A novel spam email detection system based on negative selection. In Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, Republic of Korea, 24–26 November 2009.
42. Mikheev, A. Tagging sentence boundaries. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, 29 April–4 May 2000.
43. van Rijsbergen, C.J. A New Theoretical Framework for Information Retrieval. *ACM SIGIR Forum* **2017**, *51*, 44–50. [CrossRef]
44. Tomasic, A.; Garcia-Molina, H. Query processing and inverted indices in shared-nothing text document information retrieval systems. *VLDB J.* **1993**, *2*, 243–275. [CrossRef]
45. Frakes, W. Information Retrieval: Data Structures and Algorithm. Baeza-Yates, R., Ed.; Pearson College Div: London, UK, 1992.
46. Buckley, C.; Salton, G.; Allan, J.; Singhal, A. *Automatic Query Expansion Using SMART: TREC 3*; NIST Special Publication sp; Department of Computer Science, Cornell University: Ithaca, NY, USA, 1995; p. 69.
47. Palmer, M.; Gildea, D.; Kingsbury, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput. Linguist.* **2005**, *31*, 71–106. [CrossRef]
48. Shivaji, S.K.; Prabhudeva, S. Plagiarism detection by using karp-rabin and string matching algorithm together. *Int. J. Comput. Appl.* **2015**, *115*, 37–41.
49. Alzahrani, S.; Salim, N. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In Proceedings of the CLEF 2010 LABs and Workshops, Notebook Papers, Padua, Italy, 22–23 September 2010.
50. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
51. Chong, M.; Specia, L.; Mitkov, R. Using natural language processing for automatic detection of plagiarism. In Proceedings of the 4th International Plagiarism Conference (IPC-2010), Newcastle upon Tyne, UK, 21–23 June 2010.
52. Kent, C.; Salim, N. Features Based Text Similarity Detection. *J. Comput.* **2010**, *2*, 53–57.
53. Kent, C.K.; Salim, N. Web Based Cross Language Plagiarism Detection. In Proceedings of the Second International Conference on Computational Intelligence, Modelling and Simulation, Bali, Indonesia, 28–30 September 2010; pp. 199–204.