International Journal of Medical Toxicology & Legal Medicine Volume 27, No. 5, 2024
e-ISSN: 0974-4614
p-ISSN: 0972-0448
https://doi.org/10.47059/ijmtlm/V27I5/104

An NLP-Based Plagiarism Detection Approach for Moderate-Length Sentences

Dr. Vikas Pandey[1], Dr. Shikha Pandey[2], Dr. Pawan Kumar Patnaik[3]

[1]Dept. of Information Technology, Bhilai Institute of Technology, Durg, India
[2,3]Dept. of Computer Science and Engg., Bhilai Institute of Technology, Durg, India
Email: vikas.pandeymtech2009@gmail.com, drshikhapandey2022@gmail.com, pawanpatnaik37@gmail.com

Received: 18.10.2024 Revised: 11.11.2024 Accepted: 24.12.2024

ABSTRACT

In the realm of plagiarism detection, a key challenge lies in evaluating semantic similarity between obfuscated sentences, particularly within moderate length sentences comprising only 10-20 words. To address this issue, a novel technique called Typed Dependencies Relationship (TDR), rooted in Natural Language Processing, has been introduced for identifying plagiarized content within moderate sentences. This method was tested on existing datasets of sentences and was compared against three leading plagiarism detection methods. The findings indicate that this method showing good results with keep the meaning of sentences with intricate linguistic structures.

Keywords: Plagiarism detection, NLP, Typed Dependencies Relationship

1. INTRODUCTION

Modern human communication often takes the form of brief text snippets, such as news headlines, messages, and tweets. Despite their brevity, these snippets carry significant meaning and have wide-ranging applications across various fields, including natural language processing (NLP). Analyzing this moderate length-form content can reveal crucial insights relevant to many aspects of contemporary life. It also holds considerable importance in the realms of academics especially in research. Plagiarists often cleverly modify short to moderate-length texts or sentences in research without properly crediting the original author. However, detecting plagiarism in these brief texts, which typically contain only 10-20 words and significant syntactical components, poses a significant challenge. To combat this, plagiarism detection involves identifying similarities across words, sentences, paragraphs, and documents, with word matching being the initial step. This process lays the groundwork for detecting similarities at higher levels. The goal is to develop a practical method for calculating the similarity between moderate length messages, usually around one sentence long, which can enhance plagiarism detection tools performance.

In this paper section 2 describe about the related existing work which identify the research gaps. Section 3 elaborate about system framework outline, experimental work and performance comparison with other researcher and section 4 elaborates about conclusion.

2. LITERATURE SURVEY

The current work which utilizing Natural language processing approach for plagiarism discovery also, presents the methods which is examined by this study. The majority of the exploration work did in different geographic areas has been discovered as scattered and non-uniform and have been utilizing Natural language processing strategies as clear from writing since 1990s. In most plagiarism detection frameworks, document pre processing is the fundamental stage of source detection. In pre-processing stage text are prepared in generalized formats and in document source retrieval decreases and filtered the search area. This is especially essential when the search area included large amount of documents.

The task of detecting syntactic and semantic similarity on texts is a basic issue in Natural Language Processing (NLP) due to its importance on a variety of applications of information retrieval system at any level such as document level, paragraph level or sentential level.

To detect the paragraph level of similarity is very simple because a lot of words have been found to analysis due to its length. Many research work already done for paragraph plagiarism detection in (Pandey *et al*, 2018), (Vani and Gupta, 2016), (Mohammed et al, 2019), (Chong M.,2013, (Alhzarani et al.,2010-2012,2015) used to extract syntactic features of paragraph and produce the milestones results in paragraph level of plagiarism detections.

In order to narrow down the research focus, another perspective is sentence level plagiarism detection. (Lee, 2011) discussed a strategy called Semantic Text Similarity (STS) to estimate the semantic similarity between short length sentences or messages, based on both semantic content and word order arrangement. In the paper comparison is done on semantic similarity, extracting information from a knowledge base and corpus statistics. The paper also proposes a method to account for the effect of word order on sentences. The STS strategy achieved a respectable Pearson correlation coefficient for 30 pairs of sentences and outperformed previous results.

Subsequent works is in progress on different aspect of plagiarism. (Hurtik et al, 2015) discussed on visual ads plagiarism which refers to the unauthorized use or reproduction of visual elements, such as charts, graphs, tables, diagrams, images, and illustrations, created by someone else, without proper attribution or permission. This type of plagiarism occurs when researchers present these visual aids as their own work, there by misrepresenting the originality and authenticity of their research. This paper focus on similarity comparasion in paragraph level and find out the syntatic and semantic type of plagiarism in paragraph level. (Shikha *et al* ,2024) discussed about various methods to check plagarism between paraghaphs in suspecious research papers and also provides guidelines for paraphrasing.

In this paper two practical methods for evaluating sentence similarity for short length text (4-5 word length) are introduced. The first method focuses on syntactic similarity by utilizing type dependency relationships exclusively, without incorporating a concept-based corpus or dictionary. The second approach involves semantic similarity by combining type dependency relationships (TDR) with a dictionary. An important aspect of the approach is its ability to handle text pre processing challenges, allowing sentences to be accepted without altering their structure or losing information. Both methods were found to perform effectively for moderate length sentences with efficient time complexity. In case if moderate-length sentences detection is required then it become more complicated. So here a framework is introduced for moderate length sentence (typically 10-12 words of length) which is discussed in section 3.

3. System Framework Outline

The meaning of a text arises from the words, their arrangement, and their relationships within the sentence. The suggested method detects similarities by examining both syntactic and semantic information in the texts of moderate length being compared. The proposed system framework employs measures for computing similarity between sentences in Semantic similarity (Relational similarity) level.

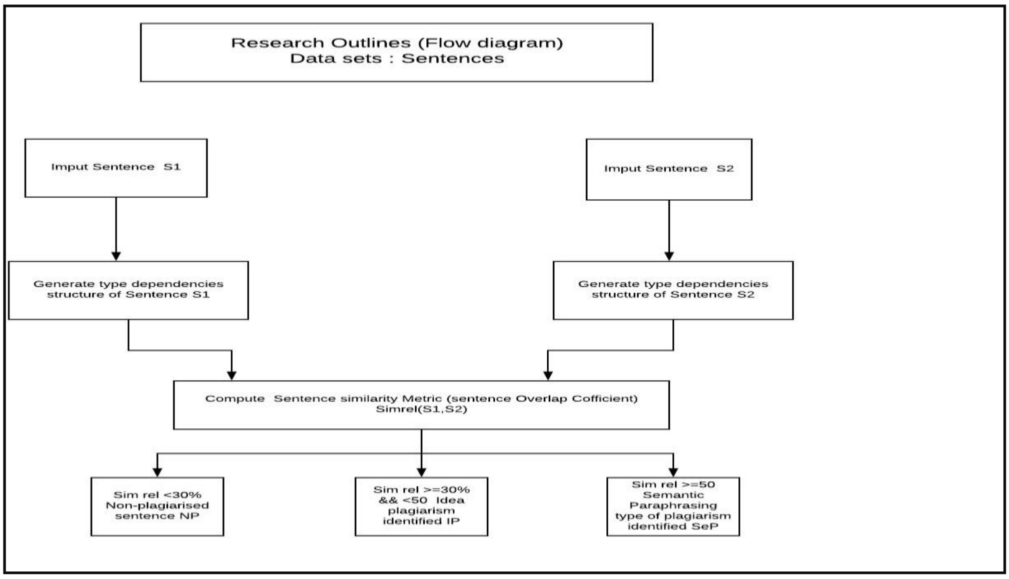

Research Outlines (Flow diagram)
Data sets : Sentences
Imput Sentence S1
Imput Sentence S2
Generate type dependencies structure of Sentence S1
Generate type dependencies structure of Sentence S2
Compute Sentence similarity Metric (sentence Overlap Cofficient) Simrel(S1,S2)
Sim rel <30% Non-plagiarised sentence NP
Sim rel >=30% && <50 Idea plagiarism identified IP
Sim rel >=50 Semantic Paraphrasing type of plagiarism identified SeP

Figure 1: Proposed System Framework

3.1 Experimental Background

The experiments were synchronized in coherence with conceptualization of four levels of obfuscation, as frequently accepted in state-of-art literature, through a crisp set of plagiarism taxonomy norms as discussed by (Alzahrani et al., 2015), (Vani et al., 2016) , (Chong et al.,2013). Instead of applying third-party natural language processing tools to support text pre-processing techniques, like use of lemmatizer, stemmer and POS-tagger, our high obfuscation plagiarism detection phase incorporates typed-dependency grammatical relations. The shift in this kind of logic deviation to detecting of high obfuscation plagiarism levels is due to the belief carried over, throughout this piece of research that, as the complexity of the structure of adjacently lying sentences increases, the information loss also increases as discussed by (Alzahrani, et al 2015), (Chong, M.

2013). The typed dependency representation of sentences were designed by Stanford NLP group to provide a simple description of the grammatical relationships in those sentences that is easily understandable and easily used by people who have no knowledge about word relation in a sentence. So along with POS-tag category information borne by each word, the relations express intra-sentential and inter-sentential relationships borne by the words of candidate sentences. (Abdullah et. al., 2020) also attempted to exploit the use of such syntactic dependency relations (in tree formats) by focusing on extraction of main POS categories within sentences, however, with a difference of using external (WordNet) controlled vocabulary for measuring total or partial word-to-word similarity using jaccard similarity metric.

So, in order to overcome all types of limitations in sentence similarity, a new approach is introduced to check sentence similarity called Relational Similarity (or Sim_{rel}) which is discussed in section 3.2.

3.2 Experimental Setup for Relational Similarity (Sim_{rel})

In the experimental set up NLTK tool kit is used to extract noun-phase and verb-phrase chunkers as lexical components for making similarity comparisons. For instance, selecting any pair of sentences, A and B, in typed dependency relation format:

{Predicate (argumentA1, argumentA2);
Predicate (argumentB1, argumentB2)}

Where Predicate and Predicate exhibits one out of many grammatical relations (enlisted in Manning's Stanford Dependency Manual in (Schuster & Manning, 2016). The sentential words represent argumentA1 and argumentA2 components for sentence A as well as argumentsB1 and argumentB2 components for sentence B. In contrast to using cosine, jaccard, containment like similarity metrics which seem to be more suitably applicable to n-gram representations of text, the typed dependency representations of the text demand the applying of an innovatively designed relational-similarity metric for measuring contextual similarity over pair of candidate sentences, denoted by Simrel (A, B) in expression 1 such that:

$$\mathbf{Sim_{rel}\ (A, B)} = \frac{\mathbf{1*|S(A,m)\cap(B,m)|+0.67*|S(A,n)\cap(B,n)|+0.33|S(A,p)\cap(B,p)|}}{\mathbf{min(countA.countB)}} \qquad \textbf{… (1)}$$

In expression (1), S (A) and S (B) represent total set of typed dependency relations in the pairs of candidate sentences. Their intersection represents set of 'm' number of common relationships exhibiting 100% similarity, 'n' number of common relationships exhibiting 67% similarity as well as 'p' number of common relationships exhibiting 33% similarity. The total number of grammatical relations appearing in sentences A and B are denoted by countA and countB. Thus, the relational-similarity takes into account, a sum of measures of the overlapping degrees of similarities observed upon complete overlapping (total match of dependency relations), moderate overlapping (two-thirds match of dependency relations) and minimum overlapping (one-third match of dependency relations)extracted out of the input sentence pairs.

3.3 Execution instances

The computation of relative-measure of similarity can be illustrated by working upon a pair of input sentences denoted by A and B as given below:

A = "If she can be more considerate to others, she will be more popular".
B = "She is not considerate enough to be more popular to others".

A typed-dependency relation for sentence A and B is given in Table1:

Table 1: A typed-dependency relation for sentence A and B

Dep. triples (sent A)	**Dep. triples (sent B)**
mark(considerate, If)	nsubj(considerate, She)
nsubj(considerate, she)	cop(considerate, is)
aux(considerate, can)	neg(considerate, not)
cop(considerate, be)	root(ROOT, considerate)
advmod(considerate, more)	advmod(considerate, enough)
advcl(popular, considerate)	mark(popular, to)
case(others, to)	cop(popular, be)
nmod(considerate, others)	advmod(popular, more)
nsubj(popular, she)	xcomp(considerate, popular)
aux(popular, will)	case(others, to)
cop(popular, be)	nmod(popular, others)
advmod(popular, more)	
root(ROOT, popular)	**Total no of relationship: 11**
Total no of relationship: 13	
Source: (courstey.http://nlp.stanford.edu:8080/parser/index.jsp, Stanford University, <2020>)	

It may be noted that both sentences reveal same meaning (semantics) if thought from human perspective. The formula designed in expression 1 attempt to capture this human perspective by computing relative degree of context similarity (overlap) borne by the pair of sentences.

The mentioned degree of context overlaps is computed by testing upon set of following relational conditions:

1. **100% matches**: All type-dependency portions (100% matches) corresponding to the sentence pair are found matched in the following manner:

(PredicateA = =PredicateB)= TRUE && (ArgumentA1 = =ArgumentB1) = TRUE && (ArgumentA2 = = ArgumentB2) = TRUE.

In the above example, the involved predicate and argument strings hold the compound relational condition: {(nsubj = = nsubj) = TRUE && (like = = like) = TRUE && (I, I) = TRUE} as true and contribute to the count of relations exhibiting total overlap.

2. **67% matches**: Any two of the three typed-dependency portions (67% matches) corresponding to the sentence pair are found matched in the following combinations:

either {(PredicateA = = PredicateB) = TRUE && (ArgumentA1 = = ArgumentB1) = TRUE} or {(PredicateA = = PredicateB) = TRUE && (ArgumentA2 = = ArgumentB2) = TRUE}

or {ArgumentA1 = = ArgumentB1) = TRUE && (ArgumentA2 = = ArgumentB2) = TRUE}.

In the above example, the involved predicate and argument strings hold the compound relational condition: {(det = = det) = TRUE && (that = = that) = TRUE} as true and contribute to the count of relations exhibiting total overlap.

3. **33% matches:** Any one of the three typed-dependency portions (33% matches) corresponding to the sentence pair are found matched in the following combinations:

either {ArgumentA1 = = ArgumentB1) = TRUE or (ArgumentA2 = = ArgumentB2) = TRUE}.

In the above example, the involved predicate and argument strings hold the compound relational condition: {(like = = like) = TRUE or (that = = that) = TRUE} as true and contribute to the count of relations exhibiting total overlap.

The detailed description of these computations is provided in Table 2 with reference to tabulations on typed-dependency relation counts.

Table 2: Degree of similarity overlapping of sentence A, B Relational - Similarity Computation

100%	**67%**	**33%**	**Sim$_{rel}$ (.)**

case(others, to) <*100*>case(others, to)
cop(popular, be) <*100*>cop(popular, be)
advmod(popular, more) <*100*>advmod(popular, more)
Total No of matching in 100% category is : 3
	nsubj(considerate, she) <*67*>nsubj(considerate, She)
	cop(considerate, be) <*67*>cop(considerate, is)
	cop(considerate, be) <*67*>cop(popular, be)
	advmod(considerate, more) <*67*>advmod(considerate, enough)
	advmod(considerate, more) <*67*>advmod(popular, more)
	root(ROOT, popular)<*67*>root(ROOT, considerate)
	Total No of matching in 67% category is: : 6
		mark(considerate, If) <*33*>nsubj(considerate, She)
		mark(considerate, If) <*33*>cop(considerate, is)
		mark(considerate, If) <*33*>neg(considerate, not)
		mark(considerate, If) <*33*>advmod(considerate, enough)
		mark(considerate, If) <*33*>xcomp(considerate, popular)
		advcl(popular, considerate) <*33*>root(ROOT, considerate)
		advcl(popular, considerate) <*33*>mark(popular, to)
		advcl(popular, considerate) <*33*>nmod(popular, others)
		Total No of matching in 33% category is: : 8
			0.8781

The corresponding predicate arguments were not considered for computing one-thirds of context similarity owing to the fact that any pair of predicates may get irrelevantly matched holding multiple combinations of ‘argument1’ and ‘argument2’ strings within the same paragraph, hence, may not reflect same meaning.

The parameter settings to be applied on "Relational Similarity" (Sim_{rel}) metric which is borrowed from our previous work (Shikha et. al 2019) as shown in Table 3.

Table 3: Relational - Similarity Computation threshold (for Revised Plagiarism Taxonomy)

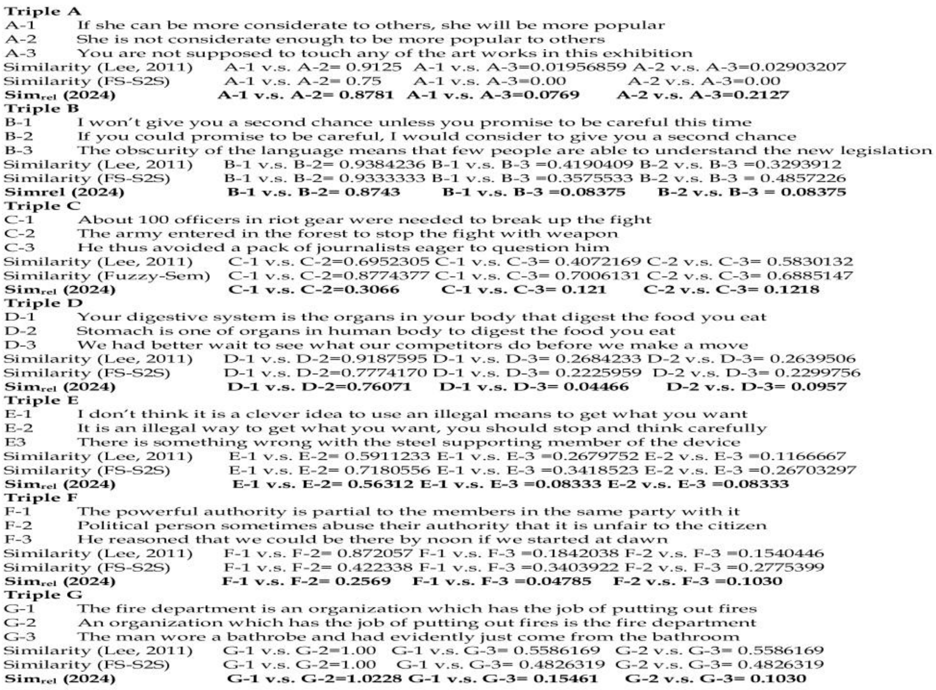
Sl. No.	Relational - Similarity (in %)	Plagiarism Taxonomy Level
1	Sim_{rel}(A,B)<= 0.3	Not Plagiarised
2	Sim_{rel}(A,B)> 0.3%&& Sim_{rel}(A,B) <=0.5	Idea Plagiarism
3	Sim_{rel}(A,B)>=0.5&& Sim_{rel}(A,B)<=0.7	Semantic Plagiarism
4	Sim_{rel}(A,B)> 0.7	Syntactic Plagiarism if J_{Sim}(A, B)>= 50%, otherwise, Semantic Plagiarism}

3.4 Performance Comparisons with corpus of Moderate-length Sentences

Having obtained favourable results in performance comparison experiments over moderate-length sentence corpus, the innovatively designed "Relational Similarity" metric was further experimented over sentence corpus of moderate lengths. (Lee, 2011) discussed that the performance of state-of-the-art approaches at that time could not be relied upon using experimental corpora having very moderate length-length sentences. The research group worked over carefully articulated set of seven moderate-length sentences as their experimental corpus each set comprising three sentences (triplets) that were checked for text overlap with one another.

The "Relational Similarity" measures along with labelling of plagiarism categories were tabulated for each of the triplets from moderate-length sentence corpus in Table 1. The sentential triplets in table 1 were compared in combinations of sentence-pairs in the same manner, as the performance evaluations were done for moderate length-sentence pairs. Further, for calculation of Relational Similarity following dataset is taken as mentioned in (Lee, 2011).

Datasets: Our previous assessment was on extremely short length sentence data sets. To exhibit that this approach (i.e. Relational Similarity) works well for medium-length sentences an experiment is conducted to detect plagiarism on medium-length sentences which are taken from data sets composed by (Lee, 2011). Lee's experiment was planned oven seven medium-length sentence sets and each one of them was further divided in 3 sentences as shown in Figure 1.



Triple A
A-1 If she can be more considerate to others, she will be more popular
A-2 She is not considerate enough to be more popular to others
A-3 You are not supposed to touch any of the art works in this exhibition
Similarity (Lee, 2011) A-1 v.s. A-2= 0.9125 A-1 v.s. A-3=0.01956859 A-2 v.s. A-3=0.02903207
Similarity (FS-S2S) A-1 v.s. A-2= 0.75 A-1 v.s. A-3=0.00 A-2 v.s. A-3=0.00
Sim_{rel} (2024) A-1 v.s. A-2= 0.8781 A-1 v.s. A-3=0.0769 A-2 v.s. A-3=0.2127

Triple B
B-1 I won’t give you a second chance unless you promise to be careful this time
B-2 If you could promise to be careful, I would consider to give you a second chance
B-3 The obscurity of the language means that few people are able to understand the new legislation
Similarity (Lee, 2011) B-1 v.s. B-2= 0.9384236 B-1 v.s. B-3 =0.4190409 B-2 v.s. B-3 =0.3293912
Similarity (FS-S2S) B-1 v.s. B-2= 0.9333333 B-1 v.s. B-3 =0.3575533 B-2 v.s. B-3 = 0.4857226
Simrel (2024) B-1 v.s. B-2= 0.8743 B-1 v.s. B-3 =0.08375 B-2 v.s. B-3 = 0.08375

Triple C
C-1 About 100 officers in riot gear were needed to break up the fight
C-2 The army entered in the forest to stop the fight with weapon
C-3 He thus avoided a pack of journalists eager to question him
Similarity (Lee, 2011) C-1 v.s. C-2=0.6952305 C-1 v.s. C-3= 0.4072169 C-2 v.s. C-3= 0.5830132
Similarity (Fuzzy-Sem) C-1 v.s. C-2=0.8774377 C-1 v.s. C-3= 0.7006131 C-2 v.s. C-3= 0.6885147
Sim_{rel} (2024) C-1 v.s. C-2=0.3066 C-1 v.s. C-3= 0.121 C-2 v.s. C-3= 0.1218

Triple D
D-1 Your digestive system is the organs in your body that digest the food you eat
D-2 Stomach is one of organs in human body to digest the food you eat
D-3 We had better wait to see what our competitors do before we make a move
Similarity (Lee, 2011) D-1 v.s. D-2=0.9187595 D-1 v.s. D-3= 0.2684233 D-2 v.s. D-3= 0.2639506
Similarity (FS-S2S) D-1 v.s. D-2=0.7774170 D-1 v.s. D-3= 0.2225959 D-2 v.s. D-3= 0.2299756
Sim_{rel} (2024) D-1 v.s. D-2=0.76071 D-1 v.s. D-3= 0.04466 D-2 v.s. D-3= 0.0957

Triple E
E-1 I don’t think it is a clever idea to use an illegal means to get what you want
E-2 It is an illegal way to get what you want, you should stop and think carefully
E3 There is something wrong with the steel supporting member of the device
Similarity (Lee, 2011) E-1 v.s. E-2= 0.5911233 E-1 v.s. E-3 =0.2679752 E-2 v.s. E-3 =0.1166667
Similarity (FS-S2S) E-1 v.s. E-2= 0.7180556 E-1 v.s. E-3 =0.3418523 E-2 v.s. E-3 =0.26703297
Sim_{rel} (2024) E-1 v.s. E-2= 0.56312 E-1 v.s. E-3 =0.08333 E-2 v.s. E-3 =0.08333

Triple F
F-1 The powerful authority is partial to the members in the same party with it
F-2 Political person sometimes abuse their authority that it is unfair to the citizen
F-3 He reasoned that we could be there by noon if we started at dawn
Similarity (Lee, 2011) F-1 v.s. F-2= 0.872057 F-1 v.s. F-3 =0.1842038 F-2 v.s. F-3 =0.1540446
Similarity (FS-S2S) F-1 v.s. F-2= 0.422338 F-1 v.s. F-3 =0.3403922 F-2 v.s. F-3 =0.2775399
Sim_{rel} (2024) F-1 v.s. F-2= 0.2569 F-1 v.s. F-3 =0.04785 F-2 v.s. F-3 =0.1030

Triple G
G-1 The fire department is an organization which has the job of putting out fires
G-2 An organization which has the job of putting out fires is the fire department
G-3 The man wore a bathrobe and had evidently just come from the bathroom
Similarity (Lee, 2011) G-1 v.s. G-2=1.00 G-1 v.s. G-3= 0.5586169 G-2 v.s. G-3= 0.5586169
Similarity (FS-S2S) G-1 v.s. G-2=1.00 G-1 v.s. G-3= 0.4826319 G-2 v.s. G-3= 0.4826319
Sim_{rel} (2024) G-1 v.s. G-2=1.0228 G-1 v.s. G-3= 0.15461 G-2 v.s. G-3= 0.1030

It was observed in each triplet, two sentences (numbered as Sentence 1 and Sentence 2 components) were articulated in such a way that they are bound to be inferred as plagiarised, while the third sentence (numbered as Sentence 3 component) in each triplet was articulated in totally different context. Hence, the similarity

calculations under each triplet cause (Sentence 1, Sentence 2) sentential pairs to be declared as plagiarised pair of sentences.
The comparison chart which show similarity index for 1 and 2 sentence pair for each triplet is shown in Figure 3.

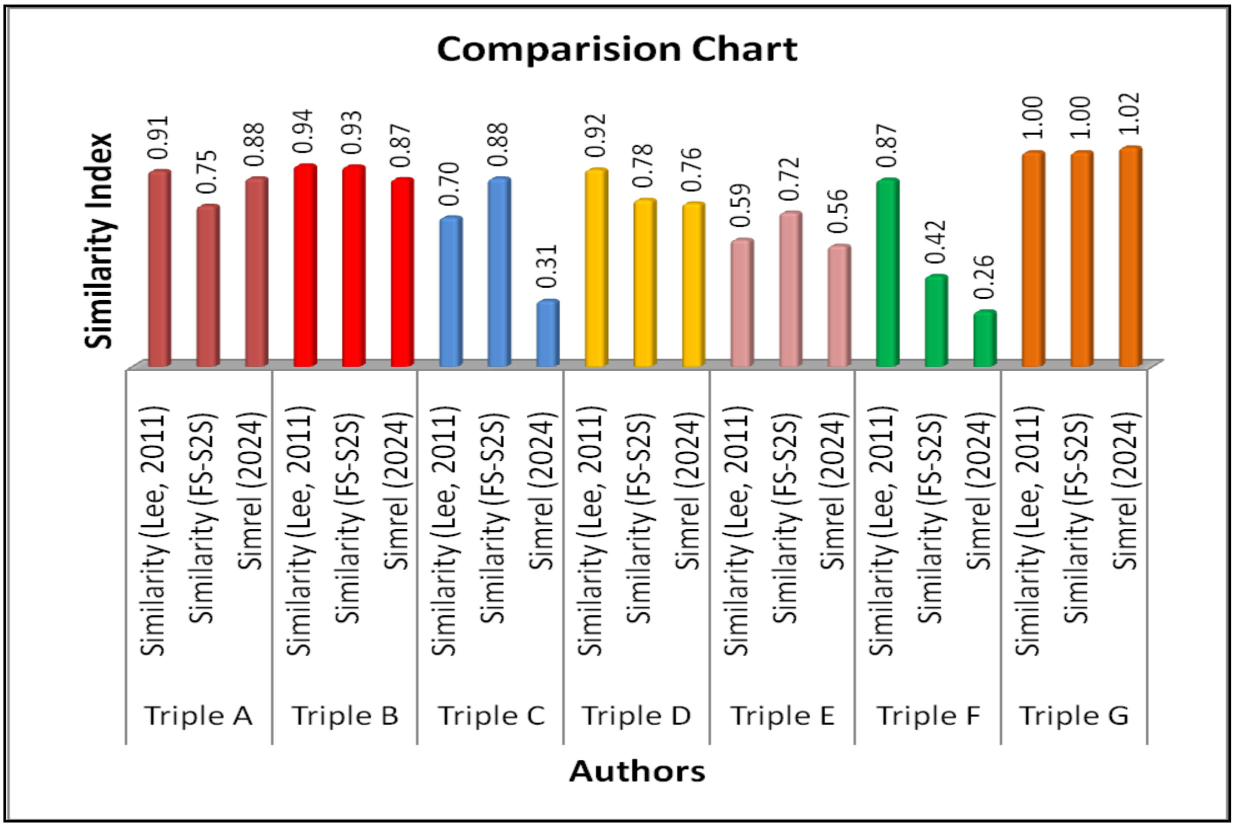


Figure 3: Performance Comparisons: Sentence-level plagiarism detection in moderate-length sentences

The similarity type is already categorised in Table3.Further, Summarised output of Sim$_{rel}$, Sim and similarity type for Triple sentence is shown in Table4.

Table 4: Shows Summarised output of Sim$_{rel}$, Sim and similarity type for Triple sentence

Sr. No	Sentence 1	Sentence 2	Sim$_{rel}$ (A,B)	Sim(A, B)	Similarity Type	Explanation
1	A-1 If she can be more considerate to others, she will be more popular	A-2 She is not considerate enough to be more popular to others	0.8781	0.166	Semantic Paraphrased	Although words are almost same but A-1 is conditional sentence and A-2,so there are semantically similar
2	B-1 I won't give you a second chance unless you promise to be careful this time	B-2 If you could promise to be careful, I would consider to give you a second chance	0.8743	0.25	Semantic Paraphrased	Although words are almost same but B-1 is conditional sentence and B-2,so there are semantically similar
3	C-1 About 100 officers in riot gear were needed to break up the fight	C-2 The army entered in the forest to stop the fight with weapon	0.3066	Nil	Ideally Paraphrased	There is no any direct similarity on both sentences so there are ideally similar
4	D-1:Your digestive system is the organs in your body that digest the food you eat	D-2: Stomach is one of organs in human body to digest the food you eat	0.7607	0.55	Syntatic Paraphrased	Both sentences pointing out same thing.
5	E-1 I don't	E-2 It is an illegal	0.563	Nil	Semantic	Although words are almost

	think it is a clever idea to use an illegal means to get what you want	way to get what you want, you should stop and think carefully	1		Paraphrased	same but B-2 is derived from B-1 ,so they are semantically similar
6	F-1 The powerful authority is partial to the members in the same party with it	F-2 Political person sometimes abuse their authority that it is unfair to the citizen	0.2569	Nil	No similarity	No direct relation ship in both sentnces. Here our results is very good than the others.
7	G-1 The fire department is an organization which has the job of putting out fires	G-2 An organization which has the job of putting out fires is the fire department	1.0228	1.00	Syntactic Similar	Both sentences are pointing out same thing and complete related ot each oher.

The Table 4 is only for combination 1 Vs 2 for all triples. Remaining combinations like 1 Vs 3 and 2 Vs 3 is in already in the category of "No Similarity", because their percentage of similarity is less than 30%. In order to verify the authenticity of these "relational similarity" measures, the third experimental setup was articulated, where high-obfuscation detection experiments were repeated on these eight pairs of moderate length sentences, but with a modification that, this time, typed-dependencies include noun-phrase arguments and recognize phrase-equivalence owing to synonymy observed from WordNet look-up dictionary. Moreover, the improved "relational similarity" measures (I-Simrel) obtained through this supporting experimental setup was found quite appealing to judgments made by nominated team of human assessors.

On the basic of above experiment conclusion is drawn which is stated in section4.

4. CONCLUSION

The NLP experiments performed did not cause any loss of textual information. This is with reference to, not using third party tools like segmentation, stop word removal, stemmed words, delimiters, numerals and punctuation symbols, as an embedded layer to the above mentioned 4-layered text pre-processing task. The text (relational) similarity computation was well-implemented on grammatical structural representations of text paragraphs and sentences. The advantages of using the "Relational Similarity" metric (Sim_{rel}(.)) was that it did not require any text pre-processing overheads, both in terms of space and time, hence, contributing to enormous reduction in plagiarism detection costs and construction of plagiarism reports.

REFERENCES

1. Abdullah, M.,S., Mazin., S., & Makttof., M., A. (2020). Modifying Jaccard Coefficient for Texts Similarity. Opcion. 32(19). 2899-2921.
2. Alzahrani, S. M., & Salim, N. 2010. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. In Lab Report for PAN at CLEF, 22-23 September. 1176, Padua.
3. Alzahrani, S. M., Salim, N. Abraham, A., & Palade, V. 2011. iPlag: intelligent plagiarism reasoner in scientific publications. Proceeding in World Congress on Information and Communication Technologies WICT, IEEE, Los Alamitos, CA, 11-14.
4. Alzahrani, S. M., Salim, N., & Palade, V. 2015. Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. In Journal of King Saud University Computer and Information Sciences:27(3), 248–268.
5. Alzahrani, S. M., Salmon, M., & Abraham A. 2012. An Understanding plagiarism linguistic patterns, textual features, and detection methods. In IEEE transactions on systems, man, and cybernetics part c: application and reviews: 42(2), 133-149.
6. Chong M. 2013. A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques. Ph.D. Thesis, University of Wolverhampton, UK .
7. Hurtik, Petr & Števuliáková, Petra. (2015). FTIP: A tool for an image plagiarism detection. 42-47. 10.1109/SOCPAR.2015.7492780.
8. Lee, M.C. 2011. A novel sentence similarity measure for semantic based expert systems. Expert Systems with Applications 38(5), 6392-6399.
9. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., & Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics, in IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138–1150.
10. Mohammed, M. T., Kadhim N. J., & Ibrahim A. A. 2019. Improved VSM Based Candidate Retrieval Model for Detecting External Textual Plagiarism. Iraqi Journal of Science, 60(10), 2257-2268.

11. Pandey, S. & Rawal, A.2019. Mapping Ethical Writing Guidelines to Plagiarism Detection Heuristics.Journal of Advanced Research in Dynamical and Control Systems, 4(11) , 1172-1177.
12. Pandey, S. & Rawal, A.2018. An NLP Based Plagiarism Detection Approach for Short Sentences. International Journal of Recent Technology and Engineering. 7(4), pp 215-219.
13. Pandey, S., & Pandey, V. Recommendations for Ensuring Equitable and Comparable Paragraphs in a Research Paper,2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1-3, https://ieeexplore.ieee.org/document/10568720
14. Schuster, S. & Manning, C. D. (2016). Enhanced English Universal Dependencies:
15. An Improved Representation for Natural Language Understanding Tasks. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2371–2378.
16. Vani, K., Gupta, D. 2016. A Study on Extrinsic Text Plagiarism Detection Techniques and Tools. Journal of engineering science and technology review. 9. 150-164. 10.25103/jestr.094.23.