# A SURVEY ON PLAGIARISM DETECTION TECHNIQUES

[1]Aditya Shankar Narayanan, [2]Anjana Vishwanath, [3]K. Senthil Kumar, [4]D. Malathi

***ABSTRACT--****Plagiarism is a major act of academic dishonesty; hence the detection of plagiarism is very essential. Therefore, Plagiarism Detection is a thriving area of research in Natural Language Processing that involves the identification of misappropriated segments of text and the retrieval of the source of the original text. This paper surveys the types of plagiarism and tasks involved in the detection of plagiarism, and analyses the existing algorithms and methods used in the Plagiarism Detection Framework. The techniques explored in this paper are: Word2vec, Monte Carlo ANN, Candidate Retrieval and Text Alignment, PV-DM and PV-DBOW, Rabin-Karp Algorithm, IR-based plagiarism detection, LSI, and Joint Word Embedding. This survey concludes that Deep Learning Based Plagiarism Detection methods show a higher accuracy than others. The survey also concludes that the existing methods (excluding LSI), lack the ability to effectively perform Cross-Language Plagiarism Detection*

***Keywords--*** *Plagiarism, Plagiarism Detection, Cross-Language Plagiarism Detection, Deep Learning Framework.*

## I. INTRODUCTION

With the increasing rate of development of technology that can be utilized for mining data from every corner of the world, data has become extremely accessible to any person with a connection to the World Wide Web. This does, however, come with the consequence of misuse of data. In the academic community, one of the most serious crimes is the act of plagiarism. Plagiarism can be defined as the “wrongful appropriation, stealing, and publication of another author’s language, thoughts, ideas, or expressions, and the representation of them as one’s original work without proper credit” [1]. This is a major issue that comes under the umbrella of Intellectual Property Rights and academic dishonesty.

Detection of plagiarism of text is an active area of research in Natural Language Processing, as opposed to plagiarism of programming language code. In this paper, we are concerned with only the former i.e., plagiarism of text. Plagiarism Detection comes under the branch of discourse-level NLP which deals with the integration of sentences and paragraphs of a text into a discourse and concentrates on the analysis of the flow of semantics throughout a discourse. Plagiarism detection systems are used to identify those texts or sections of a given text that are deemed “suspicious” of being reproduced word-for-word or paraphrased from an original text document

[1] *SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India – 603203, aditya28mar@gmail.com*

[2] *SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India – 603203, anjanavishi@gmail.com*

[3] *SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India – 603203, senthilkumar.k@ktr.srmuniv.ac.in*

[4] *SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India – 603203, malathi.d@ktr.srmuniv.ac.in*

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020**

without properly citing the reference document and giving the due credit to the original owner and author of the text.

In this paper, we present an analysis of the types of plagiarism [2], the basic architecture of plagiarism detection framework, the types of actions involved in the detection framework, and the various existing algorithms or methodologies utilized in the process of plagiarism detection. This paper surveys the following techniques: WORD2VEC [3], Monte Carlo based Artificial Neural Network (MCANN) [4], Candidate Retrieval and Text Alignment [5], Paragraph Vector Distributed Model (PV-DM) and Distributed Bag of Words (PV-DBOW) [6], Rabin-Karp Algorithm (K-gram method) and Winnowing Algorithm [7], IR-based method [8], Latent Semantic Indexing (LSI) [9], and Joint Word Embedding [10].

The paper is organized into the following sections: Section II explains the types of plagiarism (plagiarism taxonomy). Section III presents the overview of a plagiarism detection system and gives the black-box representation of the same. Section IV elaborates on the tasks involved in the detection of plagiarism (type of plagiarism detection frameworks). Section V discusses the analysis of existing plagiarism detection techniques/algorithms as mentioned in the previous paragraph, and section VI concludes this paper.

## II. TYPES OF PLAGIARISM

Barring the use of cited text, it is considered highly improbable for two different authors to produce the exact same text regardless of how similar the thought process is. Hence, without the proper use of references, the reproduction of similar text is considered an act of plagiarism. This can include the word-to-word reproduction of a text, or use of synonymy and active-passive voice conversion to paraphrase the plagiarized text. Thus, plagiarism is broadly categorized into the following [2]:

*1. Literal Plagiarism:* This refers to the first kind of plagiarism mentioned i.e., the exact word-to-word reproduction of a text. This may involve the direct copy-paste action from a pre-existing work without the use of direct quotation around the borrowed text, and mentioning the citation of the work in the references. This kind of plagiarism is easily detected using simple plagiarism detection methods.

*2. Intelligent Plagiarism:* This class of plagiarism involves the intelligent manipulation and obfuscation of a text by a plagiarist, so as to make it seem like their original content.

*a. Text Manipulation:* This usually involves methods such as exploiting the linguistic features of the text such as lexical and syntactical paraphrasing (using synonyms and changing the grammatical structure of the copied text content).

*b. Translation:* Translation of copied text from one language to another by either retaining the exact same word-order, or by employing cross-lingual paraphrasing.

*c. Idea* adoption*:* This type of plagiarism is where the plagiarist might steal someone else’s idea, work, or result, without giving proper credit to the author of the original work through citations.

Another key factor to be considered in plagiarism detection is the size of the plagiarized content. Plagiarists can often steal parts of a pre-existing text such as sections or paragraphs, or even the entirety of the text. This is key in determining the severity of the academic misconduct.

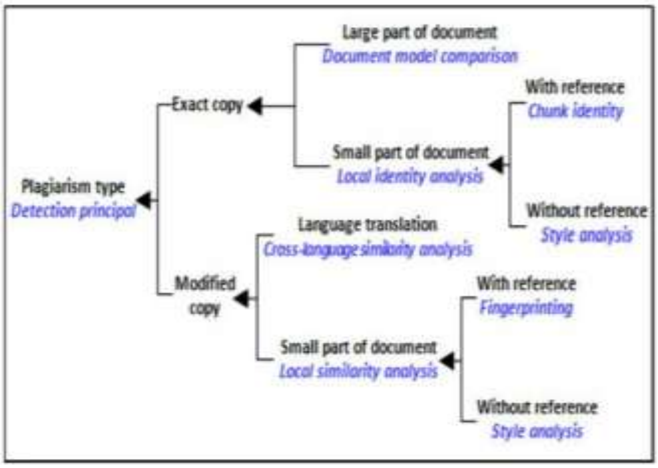


Figure 1: Types of Plagiarism [11]

## III. OVERVIEW OF A PLAGIARISM DETECTION SYSTEM

The overall design of a plagiarism detection system involves the use of a basic expert-system architecture – a knowledge base (collection of documents), and a query (the query document whose plagiarism is to be detected). The output of this system consists of the sections of the document that are suspected to have been plagiarized, and the suspected source of plagiarism that shares maximum semantic similarity. Popular plagiarism detection engines include Turnitin, WCopyFind, Docoloc, CrossCheck etc.

The black-box structure of a Plagiarism Detection system is given in fig. Consider a set of documents D which constitutes the knowledge base (training corpuses, data from web-crawlers etc.). The black box has one input $d_q$ that represents the query document. The framework returns an output as a pair of text fragments ($s_q$, $s_x$) where $s_q \in d_q$, $s_x \in d_x$, and $d_x \in D$ such that $s_q$ is the pattern of plagiarism from $s_x$ which is present in document $d_x$ in the det D.

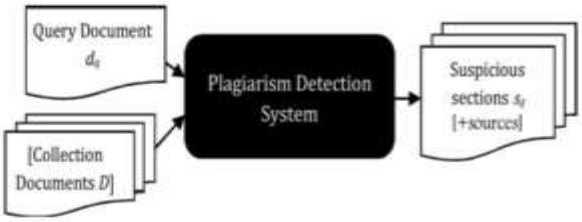

Query Document
$d_q$
[Collection Documents $D$]
Plagiarism Detection System
Suspicious sections $s_q$ [+sources]

Figure 2: Black box design of Plagiarism Detection Framework [2]

## IV. PLAGIARISM DETECTION TASKS

Plagiarism detection system broadly consists of two tasks:

***1. Extrinsic Plagiarism Detection:***

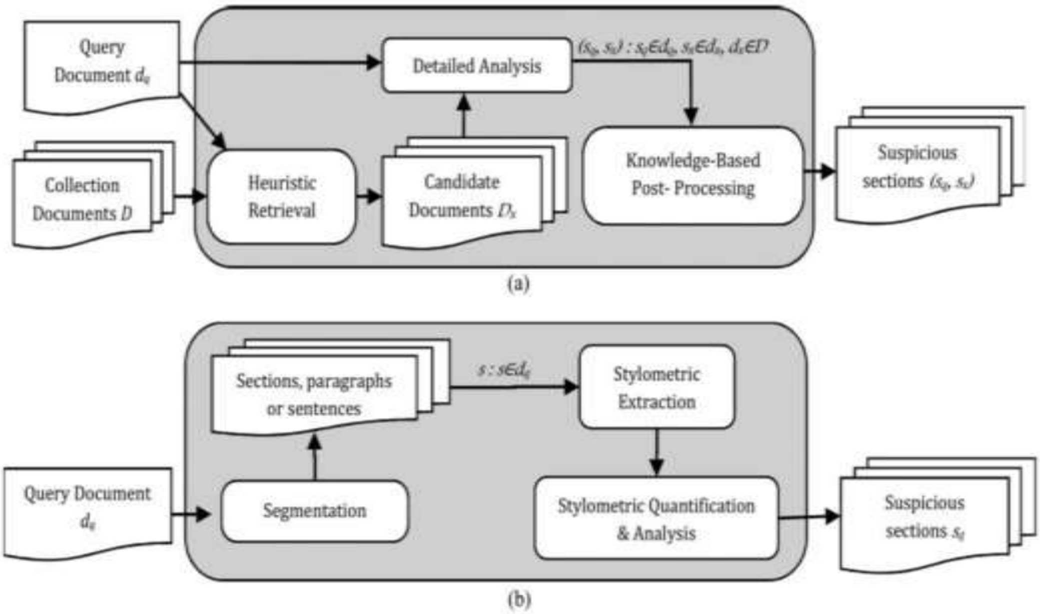
In this type, the query document is compared vis-à-vis a set of one or more source documents and the similarity is measured using various textual features. A small subset of documents from the knowledge base is selected which includes documents that are the suspected source of plagiarism and these documents are used in a pairwise feature-based exhaustive analysis with the query document to yield an output pair of text fragments that contain the

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020**

plagiarized text along with the original text. Here, the computer's ability to retrieve large text collections heuristically is used to determine the possible sources of plagiarism. This task can be done to detect almost all types of plagiarism except translation based intelligent plagiarism as shown in fig.3(a).

***2. Intrinsic Plagiarism:***

As compared to Extrinsic plagiarism detection, this task is closer to the human method of identifying plagiarism by using a stylometric system (a system that measures the variations in the writing style of the author). This includes authorship verification and authorship attribution. In simpler terms, this method checks the anomalies in the style of writing in a text and uses these anomalous fragments to check for sources of plagiarism. This task involves steps including segmentation of query document into sections, paragraphs, and sentences followed by feature extraction based on author style (stylometric extraction and quantification) which is then used to report the erroneous segments of text as shown in fig.3(b).


Query Document $d_q$
Detailed Analysis
$(s_q, s_x) : s_q \in d_q, s_x \in d_x, d_x \in D$
Collection Documents $D$
Heuristic Retrieval
Candidate Documents $D_x$
Knowledge-Based Post- Processing
Suspicious sections $(s_q, s_x)$
(a)
Sections, paragraphs or sentences
$s : s \in d_q$
Stylometric Extraction
Query Document $d_q$
Segmentation
Stylometric Quantification & Analysis
Suspicious sections $s_q$
(b)

## V. PLAGIARISM DETECTION TECHNIQUES

***1. Word2vec***

Word2vec model is a technique that employs deep learning to create a one-hot vector representation of words used in natural language. This is done by using an ANN with a single hidden layer on a large corpus. Additionally, word2vec is trained using the sliding window concept wherein the words within the neighbourhood context window are considered to compute the probability of occurrence of words while the window slides over the whole corpus in a recursive manner. The model then projects the result onto n-dimensions where every word is mapped to one vector in the n-dimension space. Then the words can be compared using similarity metrics such as cosine-similarity of the corresponding vectors. This technique can identify common semantics between samples of text if the manipulations are of the form of replacement of a word or a change in the order of grammatical classes.

Figure 3: (a) White Box for Extrinsic Plagiarism Detection; (b) White Box for Intrinsic Plagiarism Detection

Moreover, the cosine similarity between vectors represents semantic context-based similarity due to the dependence on probability of a word within a given context. [3]

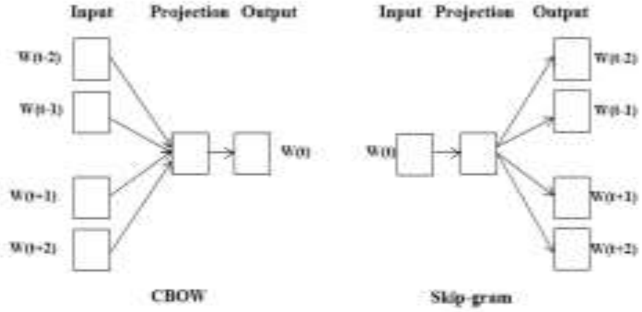


Figure 4: Word2Vec Models - Continuous Bag-of-Words (CBOW), Skip-gram [3]

***2. Monte Carlo ANN***

The Monte Carlo Method uses random sampling to solve complex problems numerically. This is a randomized algorithm which can be used to update the weights of a neural network during training for which some samples are drawn from the cosine and Jaccard similarity between vectors. Bachchan et. al. aims to develop and compare the performance of two plagiarism detection frameworks - Monte Carlo based Artificial NN, and Back Propagation. [4]

***3.*** Candidate ***Retrieval and Text Alignment***

This methodology represents records as vectors (Doc2vec) utilizing a CNN. The records in a corpus are represented as vectors and document in consideration is recovered using algorithms such as k-means clustering. The features of different n-grams are extricated at the sentence level by the CNN and the characterization of sentences are done using a Support Vector Machine (SVM). The results in Lazemi et. al. demonstrated the efficiency and success of the proposed technique. [5]

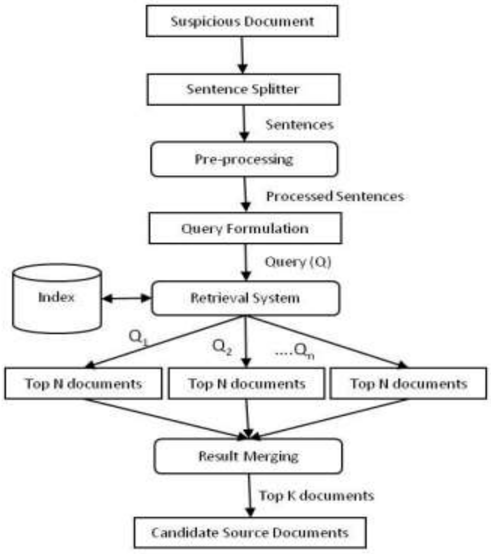


Figure 5: Flowchart for Candidate Retrieval [5]

***4. Paragraph Vectors (PV-DBOW and PV-DM)***

Doc2Vec, also called as Paragraph Vector, is an unsupervised learning algorithm, similar to word2vec which employs two different models; PV-DM model similar to CBOW which aims in learning to predict the word by the context and PV -DBOW model is similar to skip-gram model of the word2vec. PV-DBOW model ignores the order of words in the context. In PV-DM model, apart from the CBOW model, document vector uses the context word to predict the target word. PV-DM method outperforms PVDBOW significantly with an overall accuracy rate of 0.69 for classifying 20 authors of the Hürriyet newspaper in a study of Turkish documents. [6]

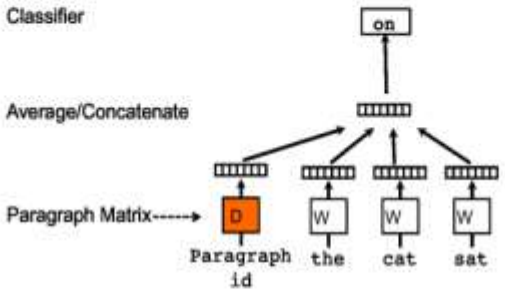


Figure 6: Paragraph Vector-Distributed Model (PV-DM) [6]

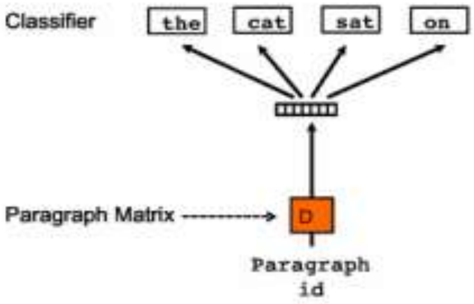


Figure 7: Paragraph Vector - Distributed Bag-of-Words (PV-DBOW) [6]

***5. K-gram method***

A study aimed at checking the similarity of documents based on the percentage of word-resemblance and comparing the system result with the human result employed the use of the K-gram method, also known as the Rabin-Karp method is one of the algorithms used to detect the similarity levels of two strings. This study proved that Winnowing algorithm with K-gram method gives a relatively good accuracy of similarity values and performs better than human method. Winnowing algorithm with K-gram method has good performance characteristics (runtime, computational complexity etc.) but has poor accuracy comparable to that of human methods. [7]

**Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020**

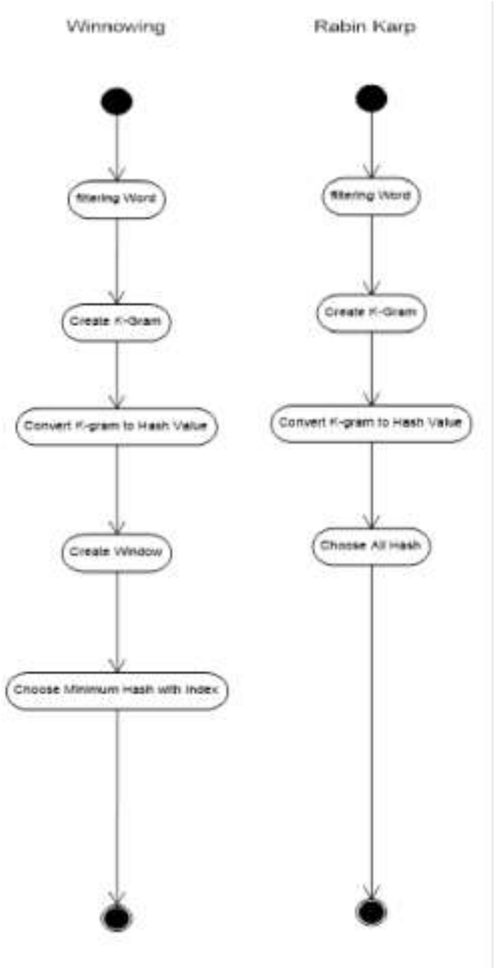


Figure 8: Winnowing and Rabin-Karp Algorithm [7]

### 6. IR Based Model

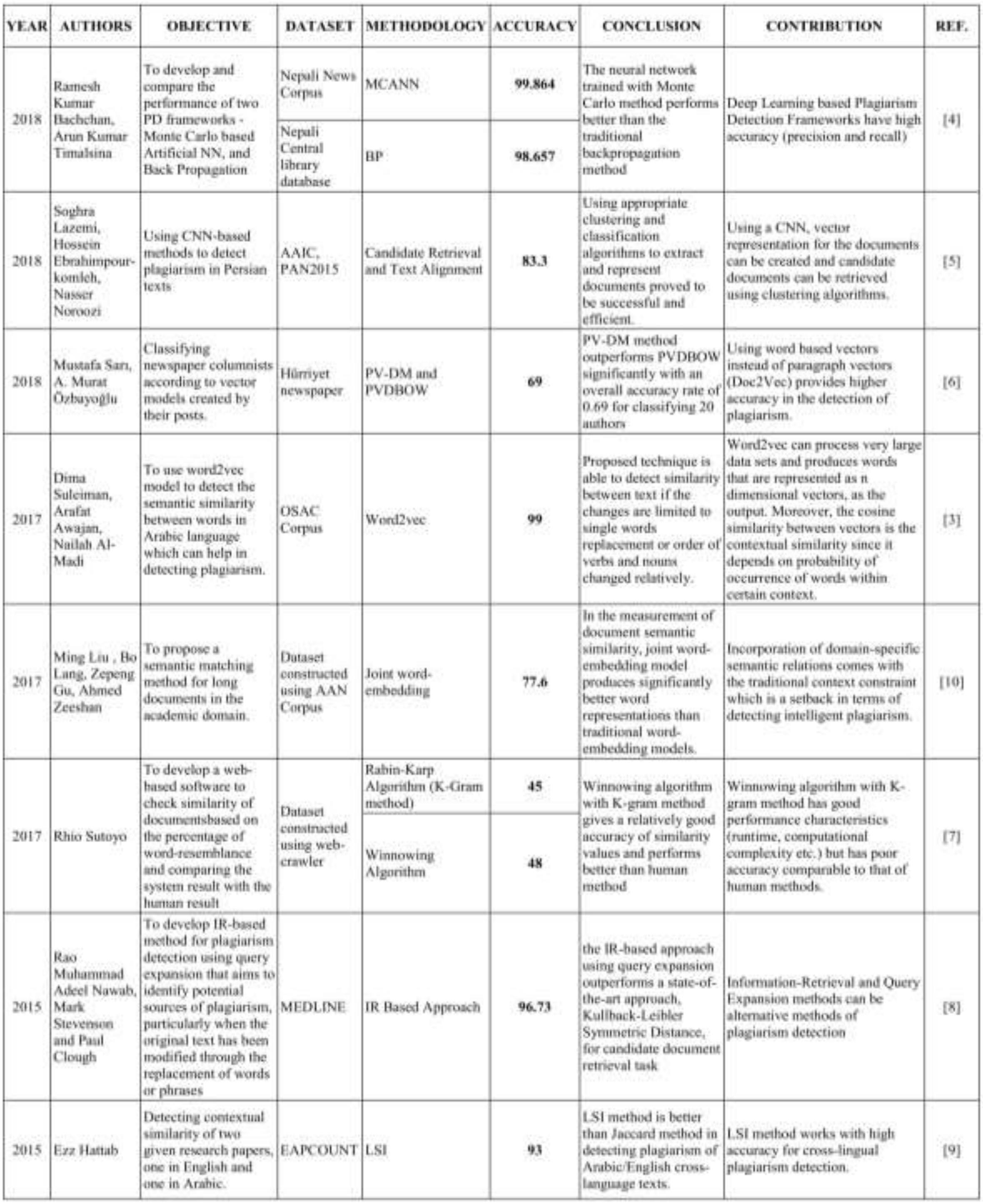
This method is developed for plagiarism detection using query expansion that aims to identify potential sources of plagiarism by Information Retrieval and Query Expansion, particularly when the words/phrases of the original text have been replaced. This approach outperforms the state-of-the-art approach, Kullback-Leibler Symmetric Distance, conventionally used in the task of candidate document retrieval. It can be concluded that Information-Retrieval and Query Expansion methods are viable alternative methods of plagiarism detection. [8]

### 7. Latent Semantic Indexing

Latent semantic indexing (LSI) is the process of correlating semantically related terms in a collection of text using singular value decomposition (SVD) of word-count-per-paragraph matrix to identify patterns in the relationships between the terms and semantic classes contained in raw text. [12] Hattab et. al. indicates that LSI performs better than Jaccard method with a higher accuracy in detecting plagiarism of Arabic/English cross-language texts. [9]

### 8. Joint Word-Embedding

The joint word-embedding model incorporates domain-specific semantic relations into the word-embedding training procedure. The objective of the model is to maximize the probability of both the context constraint and the specific semantic relation constraint. In conclusion, to measure semantic similarity in documents, joint word-embedding model produces significantly better word representations than traditional word-embedding models. [10]



| YEAR | AUTHORS | OBJECTIVE | DATASET | METHODOLOGY | ACCURACY | CONCLUSION | CONTRIBUTION | REF. |
|---|---|---|---|---|---|---|---|---|
| 2018 | Ramesh Kumar Bachchan, Arun Kumar Timalsina | To develop and compare the performance of two PD frameworks - Monte Carlo based Artificial NN, and Back Propagation | Nepali News Corpus | MCANN | **99.864** | The neural network trained with Monte Carlo method performs better than the traditional backpropagation method | Deep Learning based Plagiarism Detection Frameworks have high accuracy (precision and recall) | [4] |
| | | | Nepali Central library database | BP | **98.657** | | | |
| 2018 | Soghra Lazemi, Hossein Ebrahimpour-komleh, Nasser Noroozi | Using CNN-based methods to detect plagiarism in Persian texts | AAIC, PAN2015 | Candidate Retrieval and Text Alignment | **83.3** | Using appropriate clustering and classification algorithms to extract and represent documents proved to be successful and efficient. | Using a CNN, vector representation for the documents can be created and candidate documents can be retrieved using clustering algorithms. | [5] |
| 2018 | Mustafa Sarı, A. Murat Özbayoğlu | Classifying newspaper columnists according to vector models created by their posts. | Hürriyet newspaper | PV-DM and PVDBOW | **69** | PV-DM method outperforms PVDBOW significantly with an overall accuracy rate of 0.69 for classifying 20 authors | Using word based vectors instead of paragraph vectors (Doc2Vec) provides higher accuracy in the detection of plagiarism. | [6] |
| 2017 | Dima Suleiman, Arafat Awajan, Nailah Al-Madi | To use word2vec model to detect the semantic similarity between words in Arabic language which can help in detecting plagiarism. | OSAC Corpus | Word2vec | **99** | Proposed technique is able to detect similarity between text if the changes are limited to single words replacement or order of verbs and nouns changed relatively. | Word2vec can process very large data sets and produces words that are represented as n dimensional vectors, as the output. Moreover, the cosine similarity between vectors is the contextual similarity since it depends on probability of occurrence of words within certain context. | [3] |
| 2017 | Ming Liu , Bo Lang, Zepeng Gu, Ahmed Zeeshan | To propose a semantic matching method for long documents in the academic domain. | Dataset constructed using AAN Corpus | Joint word-embedding | **77.6** | In the measurement of document semantic similarity, joint word-embedding model produces significantly better word representations than traditional word-embedding models. | Incorporation of domain-specific semantic relations comes with the traditional context constraint which is a setback in terms of detecting intelligent plagiarism. | [10] |
| 2017 | Rhio Sutoyo | To develop a web-based software to check similarity of documentsbased on the percentage of word-resemblance and comparing the system result with the human result | Dataset constructed using web-crawler | Rabin-Karp Algorithm (K-Gram method) | **45** | Winnowing algorithm with K-gram method gives a relatively good accuracy of similarity values and performs better than human method | Winnowing algorithm with K-gram method has good performance characteristics (runtime, computational complexity etc.) but has poor accuracy comparable to that of human methods. | [7] |
| | | | | Winnowing Algorithm | **48** | | | |
| 2015 | Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough | To develop IR-based method for plagiarism detection using query expansion that aims to identify potential sources of plagiarism, particularly when the original text has been modified through the replacement of words or phrases | MEDLINE | IR Based Approach | **96.73** | the IR-based approach using query expansion outperforms a state-of-the-art approach, Kullback-Leibler Symmetric Distance, for candidate document retrieval task | Information-Retrieval and Query Expansion methods can be alternative methods of plagiarism detection | [8] |
| 2015 | Ezz Hattab | Detecting contextual similarity of two given research papers, one in English and one in Arabic. | EAPCOUNT | LSI | **93** | LSI method is better than Jaccard method in detecting plagiarism of Arabic/English cross-language texts. | LSI method works with high accuracy for cross-lingual plagiarism detection. | [9] |

Table 1. Summary of the literature survey

# VI. CONCLUSION

This paper surveyed the different classes of plagiarism, and the existing methods used to detect plagiarism. The existing tools for plagiarism detection focus mainly on literal plagiarism and use extrinsic techniques to detect the same. Furthermore, the existing techniques often focus only on monolingual plagiarism detection. However, Hattab et. al. presents an efficient method for Cross-Language Plagiarism Detection called Latent Semantic Indexing (LSI) that shows a high accuracy and superior performance over other methods. This paper also draws a conclusion that Deep Learning based frameworks for Plagiarism Detection have shown greater accuracy than their counterparts. As there is an observed increase in intelligent plagiarism, it is essential to use intrinsic methods to detect plagiarism that focus on style-extraction. Neural networks are most suitable for this, and so further research in deep-learning frameworks for plagiarism detection may increase the potential to develop an efficient system to detect all types of plagiarism – literal and intelligent, monolingual and multilingual.

# REFERENCES

1. https://en.wikipedia.org/wiki/Plagiarism
2. S. M. Alzahrani, N. Salim and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 133-149, March 2012.
3. D. Suleiman, A. Awajan and N. Al-Madi, "Deep Learning Based Technique for Plagiarism Detection in Arabic Texts," 2017 International Conference on New Trends in Computing Sciences (ICTCS), Amman, 2017, pp. 216-222.
4. R. K. Bachchan and A. K. Timalsina, "Plagiarism Detection Framework Using Monte Carlo Based Artificial Neural Network for Nepali Language," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, 2018, pp. 122-127
5. S. Lazemi, H. Ebrahimpour-Komleh and N. Noroozi, "Persian Plagirisim Detection Using CNN s," 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, 2018, pp. 171-175.
6. M. Sarı and A. M. Özbayoğlu, "Classification of Turkish Documents Using Paragraph Vector," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-5.
7. R. Sutoyo et al., "Detecting documents plagiarism using winnowing algorithm and k-gram method," 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Phuket, 2017, pp. 67-72.
8. R. M. A. Nawab, M. Stevenson and P. Clough, "An IR-Based Approach Utilizing Query Expansion for Plagiarism Detection in MEDLINE," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 4, pp. 796-804, 1 July-Aug. 2017.
9. E. Hattab, "Cross-Language Plagiarism Detection Method: Arabic vs. English," 2015 International Conference on Developments of E-Systems Engineering (DeSE), Duai, 2015, pp. 141-144.

10. M. Liu, B. Lang, Z. Gu and A. Zeeshan, "Measuring similarity of academic articles with semantic profile and joint word embedding," in Tsinghua Science and Technology, vol. 22, no. 6, pp. 619-632, December 2017.
11. Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence Lecture Notes in Bioinformatics), vol. 5253 LNAI, pp. 83–92, 2008.
12. 12. http://en.wikipedia.org/Latent_semantic_analysis