An Experimental Study with Fuzzy-Wuzzy (Partial Ratio) for Identifying the Similarity between English and French Languages for Plagiarism Detection

Peluru Janardhana Rao[1], Dr. Kunjam Nageswara Rao[2], Dr. Sitaratnam Gokuruboyina[3]
Research Scholar, Department of CS&SE, Andhra University College of Engineering, Andhra University, Visakhapatnam, India[1]
Professor, Department of CS&SE, Andhra University College of Engineering, Andhra University, Visakhapatnam, India[2]
Research Scientist, Institute of Bioinformatics and Computational Biology (Recognized as SIRO), Visakhapatnam, India[3]

***Abstract*—With the rapid growth of digital libraries and language translation tools, it is easy to translate text documents from one language to other, which results in cross-language plagiarism. It is more challenging to identify plagiarism among documents in different languages. The main aim of this paper is to translate the French documents into English to detect plagiarism and to extract bilingual lexicons. The parallel corpus is used to compare multilingual text, a collection of similar sentences and sentences that complement each other. A comparative study is presented in this paper, the sentences similarity in bilingual content is found out by using the proposed Fuzzy-Wuzzy (Partial Ratio) based string similarity technique and three various techniques like Levenshtein Distance, Spacy and Fuzzy-Wuzzy (Ratio) similarity techniques in the literature. The string similarity method based on Fuzzy-Wuzzy (Partial Ratio) outperforms in terms of accuracy compared to Spacy, and Fuzzy-Wuzzy (Ratio) techniques for identifying language similarity.**

***Keywords*—Plagiarism; natural language processing; string similarity; levenshtein distance; fuzzy-wuzzy**

I. Introduction

The ability of machines to understand the human language carried by Natural Language Processing is a crucial component of Artificial Intelligence. Search engines' arrival leads to several natural language processing advancements to retrieve the text from electronic documents with string comparison. The machines can recognize and extract patterns from text data by applying several text similarity and information retrieval techniques using NLP. Their meaning identified the closeness between two text words by the NLP technique called Text Similarity.

Natural Language Processing (NLP) in Artificial Intelligence is an important field. NLP plays a vital role in the comprehension of human language by computers. NLP uses different text similarity techniques and the combination that enables machines to create and extract patterns from those text data. The proximity of two text pieces is found out using the text similarity method, which is one of the essential NLP methods. Data needs to be translated in a numerical format to carry out machine learning tasks. TF-IDF, Word2vec and Bag of Words are the different word embedding techniques used for text data encoding. The essential steps in the text-similarity are: Text planning, Feature extraction, Vector similarity and Decision function.

A. WordNet

In English, a large database of nouns, adjectives, verbs and adverbs are grouped into a collection of synonyms. WordNet that includes the link between words in over 200 languages is a lexical database. Synonyms are interlinked with lexical and semantic relations. The structure of the WordNet makes it a helpful tool for NLP. Based on their meanings, the words in WordNet are clustered as a thesaurus cursorily resembles them. Words in the network are close to each other as the definitions of words are precisely defined by WordNet. A synonym is a crucial relation in the midst of words in WordNet.

B. NLTK

The nltk.corpus package in python defines a collection of corpus reader classes that are used to access the contents of different set of corpora. A machine that can understand the meaning of a text needs analysis which is the fundamental idea of NLP.

C. Stop Words

The meaningless words that are designed to be ignored by the search engine to increase the database space or the processing time are stop words. nltk.corpus package in python has a list of stop words in 16 various languages.

1) Research statement: The content in French language can be converted into English using conversion tools which cannot be identified in plagiarism detection. The main aim of the research is to find out the copying content in a document after translating French document into English.

2) Research objectives: To find the accuracy by conducting an experimental study with Fuzzy-Wuzzy (Partial Ratio) for identifying the similarity between languages.

3) Research significance: In recent years the attention in copying the content from various sources increased while writing a new document which is an offence. There are so many plagiarism tools available in the market for checking the originality of the content. Although the tools are working efficiently for the originality checking but still there is a problem that if we copy the content from one language and using converter tools, we can convert that content into English. In such cases it is not possible for the tools to identify the

similarity between the sentences exactly leads to a less similarity index.

II. Related Work

To find and compare cross-language articles on a specific subject, a measure of similarity is required. The basis for this estimation could be bilingual dictionaries or digital techniques, for example, latent semantic indexing (LSI) [16]. To find similar Arabic/English documents in two ways, LSI is used [1]. Monolingual: the first way is to translate the English article into Arab and then map it into space in the Arabic language LSI [10]. The second method is cross-lingual. The paper then compares LSI methods on various parallel and analog English-Arabic companies with a dictionary-based approach [8-9, 11, 13-14]. The cross-language LSI framework displays the results.

String similarity search is used in many real-life applications, like data cleaning, spell checking, fuzzy keyword search or DNA sequence comparison. Given a set of large string and a query string, the problem of string similarity search is to discover all strings in the string set that are identical to the query string efficiently [12]. Similarity is defined by using similarities measures like edit distance or Hamming distance. State Set Index (SSI) is presented as an effective solution to this work's search problem. SSI is interpreted as a finite automaton of non-determinism. SSI introduces a modern state labeling method that makes the index extremely space efficient. In addition, space usage by SSI can be traded against search time. On various sets of individual names with up to 170 million strings from a social network, they measured SSI and compared it to other state-of-the-art approaches. They show that SSI is substantially faster in most cases than other methods and needs less index space.

To retrieve math formulae from the text, three separate assessments were analyzed, in specific, Sequence Matcher, string matching algorithms, Levenshtein, and Fuzzy-Wuzzy. There are four types of Fuzzy-Wuzzy, two versions of which are found to be useful for the retrieval of Math formulae. The retrieval time of partial ratio-based Fuzzy-Wuzzy is less than the ratio-based Fuzzy-Wuzzy [2-7, 15]. They found Fuzzy-Wuzzy outperforms than Levenshtein distance and Sequence matcher techniques in terms of retrieval time and accuracy through their observations.

III. Proposed Technique

In the proposed technique, French and English documents are first loaded and then stop words and special characters are removed from them. The preprocessed documents are then converted into a list of words. French synonyms are found for each word in the French list of words using NLP and then find the corresponding English synonyms for each French word. English synonyms are found for each word in the English list of words using NLP. Now prepare the final lists list1 and list2 which have all the English synonyms of French list and English synonyms of English list. String similarity between both these final lists is computed using Spacy, Levenshtein Distance, Fuzzy-Wuzzy (Ratio) and Fuzzy-Wuzzy (Partial Ratio) techniques by taking a Threshold as shown in Fig. 1. Compare List1 with List2 and remove all the words that have a match in List2. For the leftover words check semantic closeness and if it satisfies the threshold remove the words. Likewise perform for all words until we get final leftover list.

The similarity between documents is identified with the formula:

$$100 - \left(\frac{finalLeftOverEnglishListLength}{OriginalEnglishLength}\right) * 100$$

The presented techniques are discussed below.

A. Spacy

The Spacy technique predicts how two objects are identical by comparing the objects. For flagging replicas, the similarity prediction is helpful. Context-Sensitive tensors and Word vectors are the two approaches to find the relation between the terms assisted by Spacy. Two values 0 and 1 are used to define the spectrum of similarities. The value 1 means that the two sentences are identical, and the value 0 means that the sentences are not similar. In certain instances, they still have a high similarity meaning, even though they have no standard terms. One of the essential steps in NLP is text pre-processing to remove high similarity between unmatched sentences.

B. Levenshtein Distance

The fields in which Levenshtein distance is used are computer science, computational linguistics, bioinformatics, molecular biology, and DNA analysis. The similitude between objective string and source string is evaluated by using Levenshtein Distance. In everyday life, the Levenshtein distance is commonly used. In speech recognition and plagiarism detection, Levenshtein distance is primarily used.

C. Fuzzy-Wuzzy

Fuzzy string matching often described as precise string matching to find a string that almost matches a particular pattern. Applications of Fuzzy String Matching include spell checking, detection of text reuse, spam filtering, and matching DNA sequences in the bioinformatics domain. The string similarity is checked by the Fuzzy-Wuzzy library between two terms or phrases and gives a value between 0 and 1. If the ratio is nearer to 1, the terms are well-matched. If the ratio is closer to 0, the terms are unrelated to each other. The two common fuzzy matches supported by Fuzzy-Wuzzy are suitable for finding the languages close to each other. Pure Levenshtein Distance-based matching is used in Fuzzy-Wuzzy (Ratio) technique and in Fuzzy-Wuzzy (Partial Ratio) technique matching is done based on best substrings.

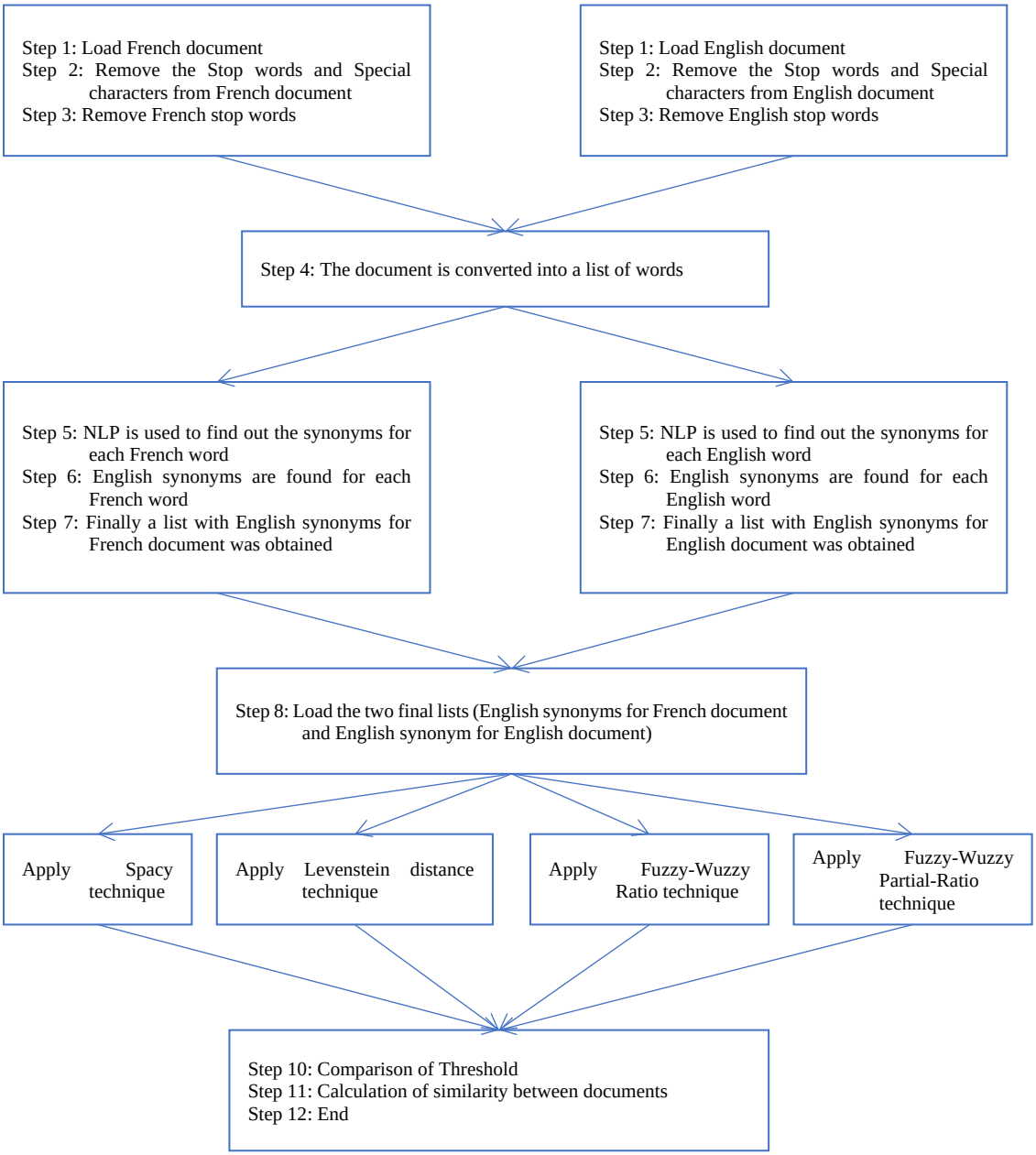

Fig. 1. Flow Diagram of the Presented String Similarity Methods.

IV. Results

The comparison between the proposed and the literature methods is presented in terms of the metric accuracy. The similarity between the languages English and French is calculated by using the accuracy and based on that the presented techniques are compared. The proposed method's competence is presented with three kinds of mappings like one-to-one, one-to-many, and many-to-many between French documents and English documents.

From the results, the results obtainable with proposed Levenshtein Distance is much appropriate for string similarity. The tests carried out on nearly 200documents; out of the proposed methods Fuzzy-Wuzzy (Partial Ratio) approach listed accuracy values in the range of 99 to 100 percent.

Table I and II represent accuracy with the Spacy, Levenshtein distance, Fuzzy-Wuzzy (Ratio) and Fuzzy-Wuzzy (Partial Ratio) techniques on 16 samples which are one-to-one mappings of French and English documents. In one-to-one mapping, there are four different ways in which the documents are being compared. 1) French Document + English document 2) French document rewrite + English document 3) French document + English document rewrite 4) French document rewrite + English document rewrite. It is observed from the tables that the accuracy ranges from 90.19 to 95.12 for Spacy technique, ranges from 94.63 to 100 for Levenshtein distance, ranges from 82.35 to 97.05 for Fuzzy-Wuzzy (Ratio) technique and ranges from 99.47 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

TABLE I. MEASURING ACCURACY IN BETWEEN ENGLISH - FRENCH ONE-TO-ONE DOCUMENTS USING THE PRESENTED TECHNIQUES

Samples	Datasets	Spacy	Levenshtein Distance	Fuzzy- Wuzzy (Ratio)	Fuzzy- Wuzzy (Partial Ratio)
Sample 1	French document + English document	90.76	98.23	94.7	100
	English document + French document re-write	90.87	98.23	94.7	100
	English document re-write + French document	92.88	97.32	88.23	100
	English document re-write+ French document re-write	92.99	97.32	93	100
Sample 2	French document + English document	91.5	98.13	95.79	100
	English document + French document re-write	91.2	98.59	95.32	100
	English document re-write + French document	94.01	98.69	91.17	100
	English document re-write+ French document re-write	93.78	99.13	95.65	100
Sample 3	French document + English document	91.72	99.31	96.9	100
	English document + French document re-write	91.48	99.31	96.56	100
	English document re-write + French document	93.39	98.12	97.05	100
	English document re-write+ French document re-write	93.22	98.12	96.25	100
Sample 4	French document + English document	92.86	96.76	93.23	100
	English document + French document re-write	93.07	96.47	93.23	100
	English document re-write + French document	94.69	98.13	91.17	100
	English document re-write+ French document re-write	94.95	97.86	95.2	100
Sample 5	French document + English document	92.56	96.24	92.22	100
	English document + French document re-write	92.3	94.63	89.81	100
	English document re-write + French document	93.99	97.28	85.29	100
	English document re-write+ French document re-write	93.78	96.79	93.82	100
Sample 6	French document + English document	92.44	97.16	92.3	99.59
	English document + French document re-write	92.94	97.16	91.49	99.59
	English document re-write + French document	93.93	97.31	85.29	99.61
	English document re-write+ French document re-write	94.51	97.31	91.95	99.61

TABLE II. MEASURING ACCURACY IN BETWEEN ENGLISH-FRENCH ONE-TO-ONE DOCUMENTS USING THE PRESENTED TECHNIQUES

Samples	Datasets	Spacy	Levenshtein distance	Fuzzy- Wuzzy (Ratio)	Fuzzy-Wuzzy (Partial Ratio)
Sample 7	French document + English document	91.73	94.8	89.96	100
	English document + French document re-write	91.45	95.15	88.92	100
	English document re-write + French document	93.35	96.01	85.29	100
	English document re-write+ French document re-write	93.15	96.67	92.3	100
Sample8	French document + English document	90.73	97.55	93.7	100
	English document + French document re-write	90.76	97.9	92.65	100
	English document re-write + French document	93.3	98.71	94.11	100
	English document re-write+ French document re-write	93.43	99.03	93.89	100
Sample 9	French document + English document	90.31	96.48	93.54	99.65
	English document + French document re-write	90.46	96.77	92.66	99.65
	English document re-write + French document	93.02	97.13	91.17	99.67
	English document re-write+ French document re-write	93.25	97.65	94.27	99.67
Sample 10	French document + English document	91.51	98.64	94.57	100
	English document + French document re-write	91.53	98.3	94.57	99.7
	English document re-write + French document	93.54	98.77	85.29	100
	English document re-write+ French document re-write	93.49	98.77	95.71	99.47
Sample 11	French document + English document	91.55	97.4	92.5	100
	English document + French document re-write	91.23	96.54	91.93	100
	English document re-write + French document	93.05	97.297	94.11	100

	English document re-write+ French document re-write	92.91	97.29	92.43	100
Sample 12	French document + English document	90.19	96.64	91.76	100
	English document + French document re-write	90.58	96.64	90.54	100
	English document re-write + French document	91.9	97.19	94.11	100
	English document re-write+ French document re-write	92.44	97.47	93.27	100
Sample 13	French document + English document	92.78	98.04	91.53	100
	English document + French document re-write	92.36	98.04	91.2	100
	English document re-write + French document	93.88	96.95	88.23	100
	English document re-write+ French document re-write	93.55	97.25	91.76	100
Sample 14	French document + English document	91.35	96.55	91.95	100
	English document + French document re-write	91.43	96.26	91.37	100
	English document re-write + French document	92.9	96.91	94.11	100
	English document re-write+ French document re-write	93.01	96.63	92.15	100
Sample 15	French document + English document	93.01	100	96.42	100
	English document + French document re-write	93.29	100	93.57	100
	English document re-write + French document	93.84	99.37	82.35	100
	English document re-write+ French document re-write	94.14	98.75	93.75	99.71
Sample 16	French document + English document	93.67	96.27	92.02	100
	English document + French document re-write	93.39	95.74	89.89	100
	English document re-write + French document	95.12	97.56	85.29	100
	English document re-write+ French document re-write	94.87	98.04	93.65	100

TABLE III. MEASURING ACCURACY IN BETWEEN ENGLISH-FRENCH ONE-TO-MANY DOCUMENTS USING THE PRESENTED TECHNIQUES

No of French Samples	**No of English Samples**	**Datasets**	**Spacy**	**Levenshtein Distance**	**Fuzzy- Wuzzy (Ratio)**	**Fuzzy- Wuzzy (Partial Ratio)**
1	1	English document + French document	90.67	98.23	94.71	99.41
1	2		91.13	97.66	89.71	100
1	3		92.28	98.62	92.43	99.65
1	4		92.99	96.76	88.52	100
1	5		92.46	93.02	82.30	99.19
1	1	French document re write + English document	90.87	98.23	94.70	99.41
1	2		91.21	96.72	87.85	100
1	3		92.37	98.62	93.47	99.65
1	4		93.05	96.17	89.70	100
1	5		92.60	93.29	80.96	99.19
1	1	French document + English document re write	92.88	97.32	92.51	99.46
1	2		93.73	98.26	90.43	100
1	3		94.07	97.81	93.75	99.68
1	4		95.00	98.13	89.06	99.73
1	5		94.19	94.81	86.41	99.50
1	1	French document re write + English document re write	92.99	97.32	92.51	100
1	2		93.77	97.39	90.43	99.56
1	3		94.16	97.81	93.75	99.68
1	4		95.09	97.6	89.06	100
1	5		94.31	94.81	86.41	99.75
2	1	English document + French document	90.74	97.64	91.76	100
2	2		91.50	98.13	95.79	100
2	3		92.30	98.28	92.09	99.65
2	4		93.17	96.17	89.70	100
2	5		92.45	93.03	84.18	99.73

2	1	French document re write + English document	90.24	97.64	91.76	100
2	2		91.20	98.59	95.32	100
2	3		91.96	98.62	93.47	99.65
2	4		93.05	96.17	90.0	100
2	5		92.30	92.76	84.45	99.73
2	1	French document + English document re write	92.84	97.32	88.77	100
2	2		94.01	98.69	95.21	100
2	3		94.07	97.5	93.75	99.68
2	4		95.11	97.6	90.93	100
2	5		94.18	94.56	86.41	99.75
2	1	French document re write + English document re write	92.44	97.86	89.93	100
2	2		93.78	99.13	95.62	99.56
2	3		93.73	97.81	95.625	99.68
2	4		94.97	97.6	91.46	100
2	5		94.02	95.06	86.91	99.75
3	1	English document + French document	90.67	98.23	94.71	100
3	2		91.13	97.66	89.71	100
3	3		92.28	98.62	92.43	100
3	4		92.99	96.76	88.52	100
3	5		92.46	93.02	82.30	99.73
3	1	French document re write + English document	90.87	98.23	94.70	100
3	2		91.21	96.72	87.85	100
3	3		92.37	98.62	93.47	100
3	4		93.05	96.17	89.70	100
3	5		92.60	93.29	80.96	99.73
3	1	French document + English document re write	92.88	97.32	92.51	100
3	2		93.73	98.26	90.43	100
3	3		94.07	97.81	93.75	100
3	4		95.00	98.13	89.06	100
3	5		94.19	94.81	86.41	99.75
3	1	French document re write + English document re write	92.99	97.32	92.51	100
3	2		93.77	97.39	90.43	100
3	3		94.16	97.81	93.75	100
3	4		95.09	97.6	89.06	100
3	5		94.31	94.81	86.41	100

TABLE IV. MEASURING ACCURACY IN BETWEEN ENGLISH-FRENCH ONE-TO-MANY DOCUMENTS USING THE PRESENTED TECHNIQUES

No of French Samples	**No of English Samples**	**Datasets**	**Spacy**	**Levenshtein Distance**	**Fuzzy- Wuzzy (Ratio)**	**Fuzzy- Wuzzy (Partial Ratio)**
4	1	English document + French document	90.74	97.64	91.76	97.64
4	2		91.50	98.13	95.79	98.59
4	3		92.30	98.28	92.09	98.96
4	4		93.17	96.17	89.70	99.70
4	5		92.45	93.03	84.18	99.19
4	1	French document re write + English document	90.24	97.64	91.76	97.64
4	2		91.20	98.59	95.32	98.59
4	3		91.96	98.62	93.47	99.31
4	4		93.05	96.17	90.0	99.70
4	5		92.30	92.76	84.45	99.19
4	1		92.84	97.32	88.77	97.86

4	2	French document + English document re write	94.01	98.69	95.21	98.69
4	3		94.07	97.5	93.75	99.37
4	4		95.11	97.6	90.93	99.46
4	5		94.18	94.56	86.41	99.01
4	1	French document re write + English document re write	92.44	97.86	89.93	97.86
4	2		93.78	99.13	95.62	98.69
4	3		93.73	97.81	95.625	99.375
4	4		94.97	97.6	91.46	99.46
4	5		94.02	95.06	86.91	99.01

Table III and IV represent accuracy with the Spacy, Levenshtein distance, Fuzzy-Wuzzy (Ratio) and Fuzzy-Wuzzy (Partial Ratio) techniques for one-to-many mappings of French and English documents. It is observed from the tables that the accuracy ranges from 90.24 to 95.11 for Spacy technique, ranges from 92.76 to 99.13 for Levenshtein distance, ranges from 80.96 to 95.79 for Fuzzy-Wuzzy (Ratio) technique and ranges from 97.64 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

TABLE V. COMPARISON OF ACCURACY BETWEEN PRESENTED METHODS FOR ENGLISH-FRENCH MANY-TO-MANY DOCUMENTS WITH FRENCH [1 6] AND DIFFERENT ENGLISH PAIRS

French pair Samples	English pair Samples	Spacy	Levenshtein Distance	Fuzzy-Wuzzy (Ratio)	Fuzzy-Wuzzy (Partial Ratio)
[1 6]	[1 8]	91.4	97.53	92.80	100
[1 6]	[3 10]	90.4	97.50	89.88	100
[1 6]	[2 6]	92.89	95.32	93.41	99.72
[1 6]	[4 7]	92.46	96.39	93.45	100
[1 6]	[3 7]	92.21	97.10	93.01	100
[1 6]	[1 8]	89.54	97.18	90.45	100
[1 6]	[5 9]	90.4	97.50	89.88	100
[1 6]	[4 6]	92.59	95.46	90.78	99.81
[1 6]	[3 9]	91.89	94.02	92.44	100

Table V represents accuracy with the presented techniques for many-to-many mappings of [1, 6] French documents and different English documents. It is observed from the table that the accuracy ranges from 89.54 to 92.89 for Spacy technique, ranges from 94.02 to 97.53 for Levenshtein Distance, ranges from 89.88 to 93.45 for Fuzzy-Wuzzy (Ratio) technique and ranges from 99.72 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

Table VI represents accuracy with the presented techniques for many-to-many mappings of [3, 10] French documents and different English documents. It is observed from the table that the accuracy ranges from 89.46 to 98.65 for Spacy technique, ranges from 95.71 to 97.86 for Levenshtein Distance, ranges from 90.45 to 100 for Fuzzy-Wuzzy (Ratio) technique and ranges from 99.25 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

Table VII represents accuracy with the presented techniques for many-to-many mappings of [5, 7] French documents and different English documents. It is observed from the table that the accuracy ranges from 89.56 to 94.76 for Spacy technique, ranges from 96.45 to 98.82 for Levenshtein distance, ranges from 90.66 to 93.47 for Fuzzy-Wuzzy (Ratio) technique and ranges from 99.82 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

Table VIII represents the accuracy with the presented techniques for many-to-many mappings of [4, 6] French documents and different English documents. It is observed from the table that the accuracy ranges from 88.94 to 95.05 for Spacy technique, ranges from 95.11 to 98.24 for Levenshtein distance, ranges from 89.20 to 93.56 for Fuzzy-Wuzzy (Ratio) technique and ranges from 99.12 to 100 for Fuzzy-Wuzzy (Partial Ratio) technique.

TABLE VI. COMPARISON OF ACCURACY BETWEEN PRESENTED METHODS FOR ENGLISH-FRENCH MANY-TO-MANY DOCUMENTS WITH FRENCH [3 10] AND DIFFERENT ENGLISH PAIRS

French pair Samples	English pair Samples	Spacy	Levenshtein Distance	Fuzzy-Wuzzy (Ratio)	Fuzzy-Wuzzy (Partial Ratio)
[3 10]	[1 9]	92.71	97.72	93.16	100
[3 10]	[3 6]	90.45	97.63	91.21	99.81
[3 10]	[2 6]	92.55	96.16	92.95	99.77
[3 10]	[4 7]	89.46	97.45	90.45	99.25
[3 10]	[2 8]	92.47	95.86	92.45	99.67
[3 10]	[2 9]	92.71	95.74	93.68	99.64
[3 10]	[5 10]	93.14	95.71	93.66	100
[3 10]	[2 7]	92.48	95.97	92.31	99.64
[3 10]	[2 10]	98.65	97.86	95.66	100

TABLE VII. COMPARISON OF ACCURACY BETWEEN PRESENTED METHODS FOR ENGLISH-FRENCH MANY-TO-MANY DOCUMENTS WITH FRENCH [5 7] AND DIFFERENT ENGLISH PAIRS

French pair Samples	English pair Samples	Spacy	Levenshtein Distance	Fuzzy-Wuzzy (Ratio)	Fuzzy-Wuzzy (Partial Ratio)
[5 7]	[3 9]	92.64	98.60	93.29	99.82
[5 7]	[2 9]	93.11	98.82	93.47	100
[5 7]	[2 8]	92.87	98.66	93.22	100
[5 7]	[1 6]	91.89	97.44	92.59	100
[5 7]	[2 6]	91.94	97.86	92.45	100
[5 7]	[5 7]	90.78	98.45	91.25	100
[5 7]	[5 10]	94.21	96.45	92.78	100
[5 7]	[1 7]	89.56	97.62	90.66	100
[5 7]	[5 9]	94.76	98.55	91.89	99.83

TABLE VIII. COMPARISON OF ACCURACY BETWEEN PRESENTED METHODS FOR ENGLISH-FRENCH MANY-TO-MANY DOCUMENTS WITH FRENCH [4 6] AND DIFFERENT ENGLISH PAIRS

French pair Samples	English pair Samples	Spacy	Levenshtein Distance	Fuzzy-Wuzzy (Ratio)	Fuzzy-Wuzzy (Partial Ratio)
[4 6]	[3 6]	92.89	95.11	93.42	99.24
[4 6]	[3 8]	92.79	96.23	93.29	99.56
[4 6]	[1 7]	93.06	98.12	91.68	100
[4 6]	[4 6]	95.05	97.46	91.02	99.63
[4 6]	[2 8]	92.88	95.54	93.22	99.12
[4 6]	[1 8]	91.64	98.24	91.87	100
[4 6]	[5 10]	90.21	96.77	90.90	99.41
[4 6]	[5 6]	88.94	96.12	89.20	99.39
[4 6]	[2 6]	93.45	96.51	93.56	99.57

TABLE IX. COMPARISON OF ACCURACY BETWEEN PRESENTED METHODS

Technique	One-One Mapping	One-Many Mapping	Many –Many Mapping
Spacy Similarity	90.19-95.12	90.24-95.11	88.94-98.65
Levenshtein Distance	94.63-100	92.76-99.13	95.11-98.82
Fuzzy-Wuzzy (Ratio)	82.35-97.05	80.96-95.79	89.20-95.66
Fuzzy-Wuzzy (Partial-Ratio)	99.47-100	97.64-100	99.12-100

Table IX represents the overall accuracy of the presented techniques with three kinds of mappings like one-one, one-many and many-many between French documents and English documents. Out of all the presented techniques Fuzzy-Wuzzy (Partial Ratio) technique outperformed all the remaining techniques with accuracy ranging from 99.12 to 100.

Table X presents the time taken to find the similarity between languages like English and French. The time calculation was accomplished on various documents of different sizes from 3KB to 50 KB with all the techniques discussed in this article. From the table, it is clear that the Spacy method identifies the similarity between English and French languages in less time than the other techniques in the literature.

Fig. 2 illustrates the data from Table I with eight samples of French and English documents with one-to-one mapping. The graph shows that the string similarity measure values of Spacy and Fuzzy-Wuzzy (Ratio) techniques are less than the Levenshtein Distance and Fuzzy-Wuzzy (Partial Ratio) techniques. It also shows that Fuzzy-Wuzzy (Partial Ratio) technique outperforms the remaining presented techniques.

Fig. 3 illustrates the data from Table III which contains one-to-many mapping of French and English documents. It shows that the accuracy of Fuzzy-Wuzzy (Partial Ratio) technique is more than the accuracy of remaining presented techniques.

TABLE X. TIME REQUIRED TO FIND THE SIMILARITY BETWEEN ENGLISH AND FRENCH DOCUMENTS

Sample Size	Spacy	Levenshtein Distance	Fuzzy-Wuzzy(Ratio)	Fuzzy-Wuzzy (PartialRatio)
3 KB	5	8	6	5
6 KB	8	14	12	9
9 KB	10	19	14	14
12 KB	13	23	19	18
15 KB	16	28	23	22
18 KB	19	36	27	26
24 KB	26	47	38	37
27 KB	30	52	43	43
30 KB	33	58	47	46
36 KB	38	70	56	55
39 KB	41	75	61	61
42 KB	43	79	67	65
45 KB	46	84	72	71
50 KB	50	92	86	84

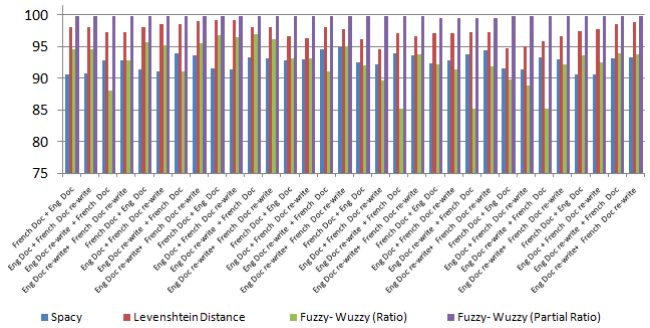

Fig. 2. Measuring Accuracy with Fuzzy-Wuzzy, Spacy Similarity and Levenshtein Distance in between French-English One-one Mapping.

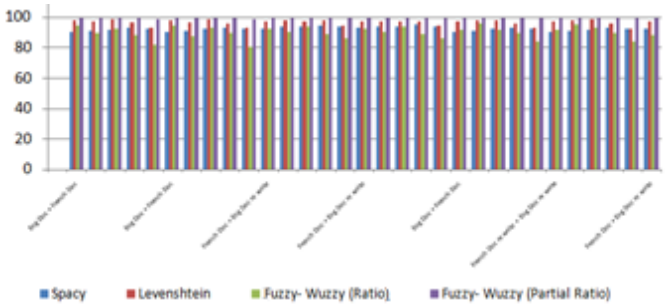

Fig. 3. Measuring Accuracy with Fuzzy-Wuzzy, Spacy Similarity and Levenshtein Distance in between French-English one-many Mapping.

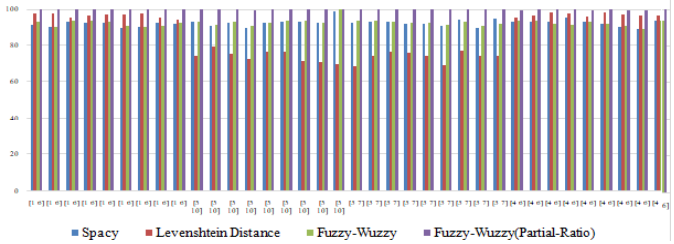

Fig. 4. Measuring Accuracy with Fuzzy-Wuzzy, Spacy Similarity and Levenshtein Distance in between French-English many-many Mapping.

Fig. 4 illustrates the data from Table V to VIII which contains many-to-many mappings of [1, 6], [3, 10], [5, 7] and [4, 6] French Documents and different English documents. It shows that the accuracy of Fuzzy-Wuzzy (Partial Ratio) technique is more than the accuracy of remaining presented techniques.

V. Conclusion

In this paper, the cross-language plagiarism detection between French and English documents is discussed. Some string similarity techniques such as Fuzzy-Wuzzy (Ratio), Fuzzy-Wuzzy (Partial Ratio), Spacy similarity, and Levenshtein distance are used to retrieve the similarity of sentences and words in multilingual content. Accuracy is the criterion used in comparing the output of thepresented techniques. More methods need to be identified to find a similarity between languages with improved precision. The Fuzzy-Wuzzy (Partial Ratio) accuracy is more significant than Fuzzy-Wuzzy (Ratio), Levenshtein distance, and Spacy similarity, but time required to find the similarity is substantial with Spacy compared to other techniques.

References

[1] D. LANGLOIS, M.Saad, K.SMAILIA, “Alignment of comparable documents: Comparison of similarity measures on French–English–Arabic data”, Volume 24, Issue 5September 2018 , pp. 677-694.

[2] M. Sree Ram Kiran Nag, G. Srinivas, K. Venkata Rao, Sairam Vakkalanka,.Nagendram, “Comparative and experimental study in identifying the similarity between languages for plagiarism detection and efficient language translation”, Materials Today, Elseveir, PP no 1-8, 2021.

[3] G.AppaRao, K.VenkataRao, P.V.G.D.Prasad Reddy and T.Lava Kumar, "An Efficient Procedure for Characteristic mining of Mathematical Formulas from Document", International Journal of Engineering Science and Technology (IJEST), Mar 2018, Vol. 10 No.03, pp. 152-157.

[4] G.AppaRao,G.Srinivas, K.VenkataRao, P.V.G.D.Prasad Reddy, "Characteristic mining of Mathematical Formulas from Document - A Comparative Study on Sequence Matcher and Levenshtein Distance procedure", International Journal of Computer Sciences and Engineering, Apr 2018, Volume-6, Issue-4, pp 400-403.

[5] G.AppaRao, G.Srinivas, K.VenkataRao, P.V.G.D.Prasad Reddy, "APartial Ratio and ratio Based Fuzzy-Wuzzy Procedure for Characteristic Mining of Mathematical Formulas from Documents", IJSC- ICTACT Journal on Soft Computing, July 2018, Vol 8, Issue 4, pp. 1728-1732.

[6] K.N.Brahmaji Rao, G.Srinivas, P.V.G.D.Prasad Reddy.," An Experimental Study with Tensor Flow Characteristic mining of Mathematical Formulae from a Document", EAI Endorsed Transactions on Scalable Information Systems,03 2019 - 06 2019 | Volume 6 | Issue 21 | e6.

[7] K.N.Brahmaji Rao, G.Srinivas, P.V.G.D Prasad Reddy, T.surendra, "A Heuristic Ranking of Different Characteristic Mining Based Mathematical Formulae Retrieval Models", Volume-9 Issue-1, October 2019.

[8] Hieber.F., and Riezler.S“Bag-of-Words Forced Decoding for Cross-Lingual Information Retrieval”. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Association for Computational Linguistics, pp. 1172–118.

[9] Morin.E., Hazem.A., Boudin.F., and Clouet.E.L. 2015, “Lina: Identifying Comparable Documents from Wikipedia”. In Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC@ACL/IJCNLP 2015), Beijing, China, Association for Computational Linguistics, pp. 88–91.

[10] Motaz Saad, David Langlois, Kamel Smaïli,“Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities”, Procedia - Social and Behavioral Sciences 95 (2013) 40 – 47.

[11] Saad.M., Langlois.D., and Smaïli, K. 2014. Cross-lingual semantic similarity measure for comparable articles. In Proceedings of the Advances in Natural Language Processing – 9th International Conference on NLP (PolTAL 2014), Warsaw, Poland, Springer International Publishing, pp. 1-12.

[12] Dandy Fenz, Dustin Lange, Astrid Rheinländer, Felix Naumann, Ulf Leser. "Chapter 18 Efficient Similarity Search in Very Large String Sets" , Springer Science and Business Media LLC, 2012.

[13] Vulić.I., and Moens.M-F. 2014,“Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data”. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Association for Computational Linguistics (ACL), pp. 349–62.

[14] Vulić.I., and Moens.M-F. 2015, “Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings”. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, New York, NY, USA, Association for Computing Machinery, pp. 363–72.

[15] K.N.Brahmaji Rao, G.Srinivas, P.V.G.D Prasad Reddy, B.Tarakeswara Rao "Non-negative Matrix Factorization Procedure for Characteristic Mining of Mathematical Formulae from Documents", Communication Software and Networks, pp: 539-551, Vol-134, Oct 2020.

[16] Nag, M. S. R. K., Srinivas, G., Rao, K. V., Vakkalanka, S., & Nagendram, S. (2021). Comparative and experimental study in identifying the similarity between languages for plagiarism detection and efficient language translation. Materials Today: Proceedings.