Kurdistan Journal of Applied Research (KJAR) | Print-ISSN: 2411-7684 – Electronic-ISSN: 2411-7706 | kjar.spu.edu.iq
Volume 2 | Issue 3 | August 2017 | DOI: 10.24017/science.2017.3.1

Plagiarism Detection Techniques for Arabic Script Languages: A Literature Review

Ribwar Bakhtyar
Sulaimani Polytechnic University-Iraq
College of informatics
University of Human Development-Iraq
ribwar.ibrahim@spu.edu.iq

Karzan Wakil
Sulaimani Polytechnic University-Iraq
College of informatics
University of Human Development-Iraq
karzanwakil@gmail.com

Soran Saeed
Sulaimani Polytechnic University-Iraq
soran.saeed@spu.edu.iq

Abstract: ***Plagiarism is generally defined as literary theft and academic dishonesty. This considered as the serious issue in an academic documents and texts. There are numerous of plagiarism detection techniques have been developed for various natural languages, mainly English. In this paper we investigate and review the plagiarism detection techniques and algorithms which have been developed for Arabic Script Languages (ASL), and providing a literature review of the utilized methods in terms of techniques and outcomes. The result of this paper will help the researchers who are going to commence their development and extend their researches in ASL like Arabic, Persian, Urdu, and Kurdish.***

Keywords: Plagiarism Detection Techniques, Literature Review, Arabic, Kurdish, Persia, Urdu.

1. INTRODUCTION

Plagiarism in simple definition is representing other's works and thoughts as one's own original work, or using the words and ideas of someone else as own work without authorization [1]. It includes disguising the source after coping the words or ideas of others then diffusing them as one's own which know as literary theft. Many people think of plagiarism as copying another's work, or borrowing someone else's original ideas. However, terms like "copying" and "borrowing" can disguise the weightiness of the crime [2]. There are many objects which can be plagiarized, in text (illegal text reuse), music, pictures, maps, technical drawings, paintings, etc.[3]. Many language-sensitive tools for detecting plagiarism in natural language documents have been developed, particularly for English [4].

The big concern and the most significant problem for universities, academic organizations and researchers are text plagiarism [5]. Students easily can find and copy documents and journals through the internet due to existing giant search engines. Some of them are just copying and pasting others works without pointing to the owner of the documents. Several types of plagiarism exist, using the published text without mentioning the sources by copying of phrases directly from it, passages or entire document, plagiarism of ideas, sources, and authorship. There are other types of plagiarism which are more advanced than ordinary, like translating content to another language, converting the plagiarized item with same content but in different media like images, videos and texts, and using program code without permission [6]. Plagiarized document detection has a great impact in many applications, such as file management, copyright protection, and plagiarism prevention [7]. Interest in this literature is more oriented towards ASL plagiarism detection.

Arabic Script Languages (ASL) are the languages which use the Arabic alphabetic for writing such as Arabic, Kurdish, Persia, Urdu, and so on. These languages more complex in morphological compared with other languages like English[8-9].

The objective of this paper is to investigate and explore the researches and works that have been done for detecting plagiarism in ASL. Furthermore reveal the outcomes of the utilized techniques. Our main research question is what are the techniques and algorithms that have been developed in ASL? To demarcate this question we perform a deep literature review of existing works.

The paper is organized as follows: In Section 2, we present a short overview of the context in which the current study has been conducted, and we explain most recent works. Section 3 describes how this research has been done, and shows the stages and the processes of the workflow. The result and discussion are presented in Section 4. Conclusion discussed in section 5 and explains suggestions for future work.

2. RELATED WORK

There are a lot of works have been proposed for finding plagiarism detection in different languages with different techniques, in this section we explain important works a (survey and literature reviews) that have been presented recently on plagiarism detection algorithms.

Maurer, et al., 2006 focused on textual plagiarism rather than plagiarism in music, paintings, pictures, maps, technical drawings, etc., firstly they discussed the complex general setting, then report on some results of plagiarism detection software. They believed that this type of papers have a value to all researchers, educators and students and should be considered as influential work that optimistically will support many still deeper analyses. Finally they claimed that the improvement of existing plagiarism techniques and algorithms are highly needed due to increasing digitizing documents day after day [10].

Five years later a group of researcher Ali, et al., (2011) argued that plagiarism is so difficult to be 100% detected by recent methods so it will continue rising and raising up. According to their survey, each method has some advantages and disadvantages. Most of them use clustering as techniques of sorting and summarization tool. They emphasized on using cluster based retrieval or clustering to achieve better results in plagiarism detection systems, and they pointed the limitation of the grammar-based method and the Semantics-based method, instead of them they advised to use semantics-based method for cluster based method as it will achieve much better results. They provide and list the advantages and disadvantages of the latest and the important effective methods used or developed in automatic plagiarism detection, according to their result. Mainly methods used in natural language text detection, index structure, and external plagiarism detection and clustering based detection [11].

Another valuable work is Osman, et al., (2012), they presented a professional study as they classified most techniques in text plagiarism into seven categories and explained the advantages and limitation each of them. Moreover they argued many important issues regarding plagiarism detection like tasks and processes of the current plagiarism detection. Finally they explained the weaknesses of some techniques which are lacking for detecting some types of plagiarized text [12].

In different work another survey in same year has been published by Bin-Habtoor, et al., (2012), they classified their survey into four categories which are plagiarism in (documents, code, techniques and algorithms). They stated that plagiarism detection for information is a big concern in universities and for teachers, policy-makers and students. Hence they proposed a system that is able to detect many plagiarism tries in deferent fields (E-Learning, E-Business, and E-Journals) and can be used to check programs, papers with images included [13].

The most recent study is Eisa, et al. (2015), they have analyzed and identified the state-of-the-art plagiarism techniques in terms of their attributes, limitations, processes and taxonomies. They revealed that the existing techniques are incapable to perform an intelligent detection efficiently for plagiarized ideas, figures, tables, formulas and scanned documents therefore they recommended that the integration of structural features and contextual information with semantic similarity methods can help to detect these types of plagiarism. They also stated that Turnitin is the most accurate in detection and steadiest tool among the existing seven tools, after analyzing their performance. Furthermore they discovered areas where further improvements are required in existing techniques and the current trends in plagiarism detection [14].

Most previous works have been focused on reviewing and analyzing the existing plagiarism detection techniques, algorithms, methods, systems, and tools meant for English language not for Arabic, Persia, Urdu or Kurdish. The absence of such survey and literature review that reflects the need of highlighting and identifying the existing plagiarism detection techniques, algorithms, methods and tools for ASL is the motivation of this work. Besides helps researchers through discovering the areas where further improvements are required in existing techniques and the current trends in plagiarism detection for ASLs.

3. METHODOLOGY

3.1 Research Question

Research Questions in plagiarism detection, acquiring a general idea of the present techniques, algorithms and tools within the scope of the ASL is the objective of this study. To clarify this aim, we demarcated three research questions:

RQ1: What are the techniques and algorithms that have been developed in ASL? Many algorithms and methods have been developed for plagiarism detections mainly in English language; we try to concentrate on the algorithms that utilized in plagiarism detection in ASL. This question is meant to observe how far these approaches provide for the overall goals at present.

RQ2: What diverse kinds of results of utilizing different algorithms for ASL plagiarism detection? This question meant providing thorough comparison between the results of the works have been done for ASL.

3.2 Research Strategy

For any work or study there should be a plan to obtain the its goal so the strategy of this work consists of two main phases as shown in the Figure 1, the first phase is search process which includes, search string, literature resources, and utilizing Endnote to preparing the extracted papers to the second phase. Primary study selection as second phase includes four stages. More details about research strategy explained as below:

- **Search Process:** The search process conduct by the search string the resulting search terms were: (approach OR method OR methodology OR technique) AND (plagiarism detection OR plagiarism software OR plagiarism tool) AND (Arabic OR Persia OR Urdu OR Kurdish)

The recourse of the string is literature resources. Five electronic database resources were used to extract data for synchronization in this research: IEEE Xplore, ACM Digital Library, Science Direct, Springer and Google Scholar. Title and index terms were used to conduct searches for published journal papers, conference proceedings, workshops, symposiums, book chapters and IEEE Bulletins. A survey of extracted papers entails a comprehensive search of all relevant sources about a subject of discussion. For managing and referencing purposes all found papers imported to Endnote software.

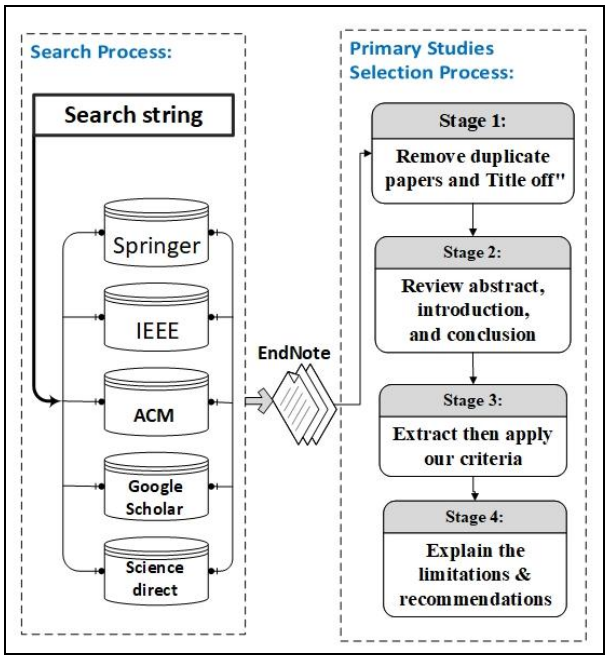

Figure 1 Research Strategy

- **Primary studies selection process:** This phase starts directly after importing all extracted papers through Endnote software; the first stage of this phase is removing the duplicated papers and performing title off from extracted papers which has been prepared in the previous phase from literature resources. Preparing those

papers for reviewing the abstract, introduction and conclusion are performed in the second stage in order to segregate the topic related papers which are contains the required and enough info that coverage the focused research area. The number of papers that cover our research question is 28 papers.

The third stage is the most time consuming process because we reviewing the 28 papers in deep, from collected papers in the previous stages so that we can perform a second extraction then apply our criteria to achieve first research question. Furthermore; the all works were gathered classified and categorized based on their utilized algorithms and their results. As a final process of the final stage all works are explained and performing the evaluation in terms of their results.

4. RESULT AND DISCUSSION

In this section we explain the results of the survey on plagiarism detection for ASLs that has been reviewed. We organize a table to explain a thorough survey of state-of-the-art plagiarism detection techniques and to better understanding we produce some charts based on our literature review statistics. Most techniques detect plagiarism by using certain text features along with fingerprint matching techniques and most of the them used some algorithms in the pre-processing stage of the system like normalization, tokenization, stemming and part of speech (POS) tagging, stop-word removal, sentence segmentation, synonymy recognition, number replacement, lemmatization. It is obvious that all utilized techniques are showed in the table 1 has its own impact on developing plagiarism detection for ASL. Most of the studies and developments are stretched in literal type of plagiarism while the minor works dealt with intelligent type. A few numbers of study produced an implemented tool or software meanwhile the others proposed a development in a particular algorithm or technique, the summery of each study that have been reviewed are explained in table 1.

Table 1: Extracted Papers Based on the Criteria

Ref.	Type	Source or target	Year	Language	Techniques	Result
[15]	intelligent	Document	2009	Arabic	Fuzzy technique in information retrieval	Stated that Fuzzy technique is better than Boolean IR , in plagiarism detection
[16]	Literal	Text	2009	Urdu	based on interpolation of n-gram probabilities techniques.	It proposed method based on interpolation of n-gram probabilities of author attribution to Urdu that outstanding bigram and trigram results for Urdu poetry.
[17]	literal	E-learning	2009	Arabic	Syntax Similarity based detection	For the first time created APD tool for Arabic in e-learning.
[18]	Literal	Text	2010	Arabic	fingerprint matching	Improved fingerprint matching technique through adding four key features of the text.
[19]	Literal	Text	2011	Urdu	n-gram model for word retrieval	Tri-gram model is better than both bi-gram and four-gram models for Urdu text plagiarism.
[20]	Literal	Document	2011	Arabic	Fingerprinting	APlag, a plagiarism detection tool for Arabic language.
[21]	Literal	Text	2012	Arabic	Stylysis tool.	Discover the effect of some well-known language-independent stylistic features on Arabic text to improve plagiarism detection.
[22]	Literal	Text & document	2012	Arabic	winnowing n-gram fingerprinting	It proposed mono-lingual system (Iqtabs 1.0) for plagiarism detection that precedes multi-lingual
[4]	Literal	Document	2012	Arabic	Fingerprinting and Similarity metric	Improved Aplag
[23]	Intelligent	Text	2013	Arabic	Examined the existing literal	It presented a new taxonomy of plagiarism that highlights differences between literal and intelligent

					systems.	plagiarism. They emphasized that existing systems for intelligent plagiarism detection are failed.
[24]	literal	Authorship	2013	Arabic	Word N-Grams.	Stated that good attribution performances with an optimal score of 80% of good authorship attribution
[25]	Literal	Authorship	2014	Arabic	MBNB technique Naı¨ ve Bayes classifiers	Attribute the author of a text with an accuracy of 97.43%.
[26]	literal	Authorship	2014	Arabic	Two popular classifiers: FT and SVM.	Stated that the FT method has better performance as accuracy of 82% was achieved.
[27]	Literal	Text	2014	Persia	NLP techniques and N-gram	It proposed AMZPPD: implemented system using python,
[28]	Literal	Text	2015	Persia	an artificial obfuscation strategy	The first plagiarism detection corpus in Persian language
[29]	literal	document	2015	Arabic	Similarity technique in information retrieval	A web-based plagiarism detection framework for Arabic documents.
[30]	Literal	Document and Text	2015	Persia	Fuzzy approach	It presented fuzzy method PFPD to distinguish paraphrased cases.
[31]	Literal	Text	2015	Persia	a mixed fuzzy inference system method	It proposed method gained accuracy rate of 78% which increases precision and recall measure. It also proposed to overcome the ambiguity and consider structural features and author style of writing in the similarity measurement in Persian texts
[32]	Literal	Text and Document	2016	Persia	Obfuscation strategies to provide corpus	An intrinsic and extrinsic plagiarism detection corpus that consists of thousand Persian academic texts Mahak Samim.
[33]	Literal	Text	2016	Persia	N-gram.	It proposed a method, word by word and sentence by sentence. This method has resulted in %90.6 plagiarism detection; % 85.8 recalls, %95.9precision on the PAN2016 provided data sets.
[34]	Literal	Document and Text	2016	Persia	Fingerprinting the text in the tri-grams words.	It stated that the lack of accuracy in the language-free tools in relation to the language-sensitive methods, it showed the effect of data-mining algorithm prevent excess comparison.
[35]	Intelligent	Text	2016	Persia	N-gram	Their approach was unable to achieve the expected result based on modification to the approach used for PAN between 2011 and 2014.
[36]	Literal	Document and text	2016	Persia	vector space model	It proposed a method which consists of three building named seeding, match merging and extraction filtering. It can remove the common words in the sentences which are likely to be a source of plagiarism.
[37]	Literal	Text	2016	Persian	An extrinsic SVM-based	the functionality and performance of SVM method to detect plagiarism in Persian texts was evaluated. A new approach called “Index Word Replacement” was suggested to detect semantic similarities.
[38]	Literal	text	2016	Persia	sentence-level algorithm based on tf-idf features	It proposed an algorithm designed for *near-copy* and *paraphrasing* types of plagiarism.
[39]	Literal	Document and Text	2016	Persia	Bi-gram and a graph structure based method.	They stated that graph based approach achieve better results in plagiarism detection.
[40]	Intelligent	Text	2017	Persia	Similarity Techniques in Information retrieval,	It proposed a method for a cross-lingual plagiarism detection based on a semantic approach. they revealed that the highly accurate translation has a big impact on intelligent plagiarism detection. Compared its method with/out employing Google translation. 98.82% when employing highly accurate translation tools, 56.9%. Without accurate translation. It also showed that monolingual methods.
[41]	literal	*document*	2017	Arabic	word stemming, Fingerprinting.	A web-based plagiarism detection framework for Arabic documents.

To better understanding and make our literature review more clear, we generate a bar chart of publications per year as shown in Figure 2. Figure 2 shows 28 papers per year between 2009 and 2017; the result of a bar chart is the publication of continual plagiarism detection growth. In 2009, only three papers were found, but in 2016 there

were 8 papers, with most publications between 2014 and 2016. However, the result for 2017 was such because probably, our search in March 2017 found some unpublished papers. Hence these results state that this area is a new and active area, which means that in the last decade the researchers have focused on this area in publications especially in the last three years.

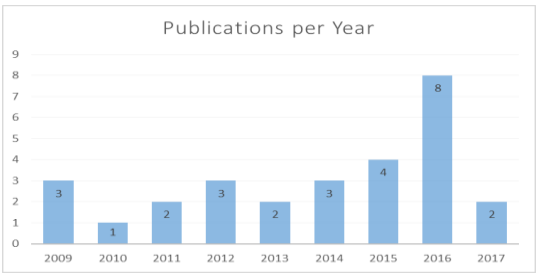

Figure 2 Publications per Year

We generate a pie chart to display a distribution for all works that have been found for ASL and present in figure 3, Figure 3 shows the distribution 28 papers of Plagiarism detection for ASL(Arabic, Persia and Urdu), Arabic portion is (47%), followed by Persia (46%) and Urdu (7%).

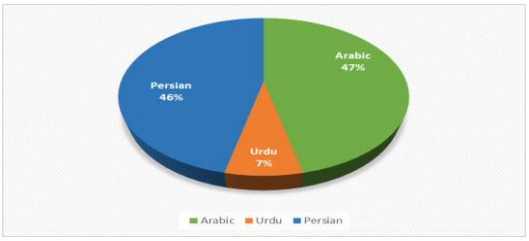

Figure 3 Distributions of Languages

We make another distribution for found papers for plagiarism detection based on plagiarism taxonomy[42], most of the papers that are related to the literal type of plagiarism detection focused on document and text, text, authorship and rarely e-learning, meanwhile there are only four papers that dealt with text and document in a smart way for detecting plagiarism which this refers to the intelligent type, as shown in figure 4, figure 4 shows distribution of plagiarism detection types

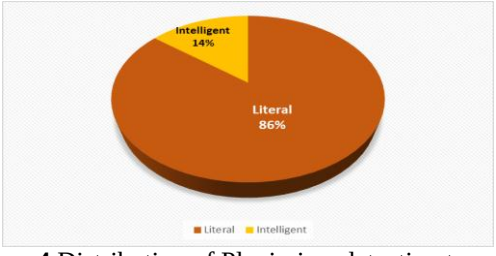

Figure 4 Distribution of Plagiarism detection types for ASLs

The results of our literature review will be affected by various factors, for instance, the researchers who conducted the study, the databases, and the search string developed, as well as the time restrictions chosen.

5. CONCLUSION AND FUTURE WORK

In this paper we have analyzed and presented a thorough review of state-of-the-art plagiarism detection techniques that have been proposed for ASLs. To the best of our knowledge this is the first study that survey on plagiarism detection for ASLs. We revealed that the most techniques detect plagiarism by using certain text features along with n-gram with fingerprint matching techniques. Furthermore some techniques like normalization, tokenization, stemming, etc. are used in pre-processing stage to make the system more efficient.

In ASL most of plagiarism detections mainly belonging to literal type meanwhile there are a few works that dealt with intelligent type. We recommend researchers that there are many areas in ASLs especially in Arabic and Persia that can be develop with intelligent type. There is no plagiarism detection for Kurdish language so far, neither literal nor intelligent. For that matter, we think it will be a hot topic for the next few years.

6. REFERENCES

[1] plagiarism.com, "glatt plagiarism services," 2017.

[2] UKessays, "A Survey Of Plagiarism Detection Methods Information Technology Essay," 2015.

[3] H. A. Maurer, *et al.*, "Plagiarism-a survey," *J. UCS*, vol. 12, pp. 1050-1084, 2006.

[4] M. E. B. Menai, "Detection of plagiarism in Arabic documents," *International journal of information technology and computer science (IJITCS)*, vol. 4, p. 80, 2012.

[5] Plagiarism.org, "What is Plagiarism," 2015.

[6] A. Jadalla and A. Elnagar, "A plagiarism detection system for Arabic text-based documents," *Intelligence and Security Informatics*, pp. 145-153, 2012.

[7] C. Lyon, *et al.*, "Plagiarism is easy, but also easy to detect," *Plagiary*, 2006.

[8] M. Mirdehghan, "Persian, Urdu, and Pashto: A comparative orthographic analysis," *Writing Systems Research*, vol. 2, pp. 9-23, 2010.

[9] E. Britannica, "Encyclopædia Britannica Online. Encyclopædia Britannica, 2011," *Web. Feb*, vol. 10, 2011.

[10] H. A. Maurer, *et al.*, "Plagiarism-a survey," 2006.

[11] A. M. E. T. Ali, *et al.*, "Survey of plagiarism detection methods," in *Modelling Symposium (AMS), 2011 Fifth Asia*, 2011, pp. 39-42.

[12] A. H. Osman, *et al.*, "Survey of text plagiarism detection," *Computer Engineering and Applications Journal (ComEngApp)*, vol. 1, pp. 37-45, 2012.

[13] A. Bin-Habtoor and M. Zaher, "A survey on plagiarism detection systems," *International Journal of Computer Theory and Engineering*, vol. 4, p. 185, 2012.

[14] T. A. E. Eisa, *et al.*, "Existing plagiarism detection techniques: A systematic mapping of the scholarly literature," *Online Information Review*, vol. 39, pp. 383-400, 2015.

[15] S. M. Alzahrani and N. Salim, "On the use of fuzzy information retrieval for gauging similarity of arabic documents," in *Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the*, 2009, pp. 539-544.
[16] A. A. Raza, *et al.*, "N-Gram Based Authorship Attribution in Urdu Poetry," in *Proceedings of the Conference on Language & Technology*, 2009, pp. 88-93.
[17] S. M. Alzahrani, *et al.*, "Work in progress: Developing Arabic plagiarism detection tool for e-learning systems," in *Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of*, 2009, pp. 105-109.
[18] C. K. Kent and N. Salim, "Features based text similarity detection," *arXiv preprint arXiv:1001.3487*, 2010.
[19] M. A. Khan, *et al.*, "Copy detection in Urdu language documents using n-grams model," in *Computer Networks and Information Technology (ICCNIT), 2011 International Conference on*, 2011, pp. 263-266.
[20] M. E. B. Menai and M. Bagais, "APlag: A plagiarism checker for Arabic texts," in *Computer Science & Education (ICCSE), 2011 6th International Conference on*, 2011, pp. 1379-1383.
[21] I. Bensalem, *et al.*, "Intrinsic plagiarism detection in Arabic text: Preliminary experiments," in *II Spanish Conference on Information Retrieval (CERI'12)*, 2012.
[22] A. Jadalla and A. Elnagar, "A fingerprinting-based plagiarism detection system for Arabic text-based documents," in *Computing Technology and Information Management (ICCM), 2012 8th International Conference on*, 2012, pp. 477-482.
[23] L. Ramya and R. Venkatalakshmi, "Intelligent plagiarism detection," *International Journal of Research in Engineering & Advanced Technology (IJREAT)*, vol. 1, pp. 171-174, 2013.
[24] S. Ouamour and H. Sayoud, "Authorship attribution of short historical arabic texts based on lexical features," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*, 2013, pp. 144-147.
[25] A. S. Altheneyan and M. E. B. Menai, "Naïve Bayes classifiers for authorship attribution of Arabic texts," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, pp. 473-484, 2014.
[26] A. F. Otoom, *et al.*, "Towards author identification of Arabic text articles," in *Information and Communication Systems (ICICS), 2014 5th International Conference on*, 2014, pp. 1-4.
[27] M. Mahmoodi and M. M. Varnamkhasti, "Design a Persian Automated Plagiarism Detector (AMZPPD)," *arXiv preprint arXiv:1403.1618*, 2014.
[28] K. Khoshnavataher, *et al.*, "Developing monolingual Persian corpus for extrinsic plagiarism detection using artificial obfuscation," *Notebook for PAN at CLEF*, 2015.
[29] I. H. Khan, *et al.*, "AFramework FOR PLAGIARISM DETECTION IN ARABIC DOCUMENTS."
[30] S. Rakian, *et al.*, "A Persian Fuzzy Plagiarism Detection Approach," *Journal of Information Systems and Telecommunication (JIST)*, vol. 3, pp. 182-190, 2015.
[31] H. Ahangarbahan and G. A. Montazer, "A Fuzzy Approach for Ambiguity Reduction in Text Similarity Estimation (Case Study: Persian Web Contents)," *Information Systems & Telecommunication*, p. 216, 2015.
[32] M. R. Sharifabadi and S. A. Eftekhari, "Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems," in *FIRE (Working Notes)*, 2016, pp. 190-192.
[33] E. Gharavi, *et al.*, "A Deep Learning Approach to Persian Plagiarism Detection," in *FIRE (Working Notes)*, 2016, pp. 154-159.
[34] S. Rafieian, "Plagiarism checker for Persian (PCP) texts using hash-based tree representative fingerprinting," *Journal of AI and Data Mining*, vol. 4, pp. 125-133, 2016.
[35] L. Gillam and A. Vartapetiance, "From English to Persian: Conversion of Text Alignment for Plagiarism Detection," *PAN@ FIRE2016 Shared Task on Persian Plagiarism Detection and Text Alignment Corpus Construction. Notebook Papers of FIRE 2016*, 2016.
[36] N. Ehsan and A. Shakery, "A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection," in *FIRE (Working Notes)*, 2016, pp. 145-148.
[37] F. Esteki and F. S. Esfahani, "A Plagiarism Detection Approach Based on SVM for Persian Texts," in *FIRE (Working Notes)*, 2016, pp. 149-153.
[38] M. Mansoorizadeh, *et al.*, "Persian Plagiarism Detection Using Sentence Correlations," in *FIRE (Working Notes)*, 2016, pp. 163-166.
[39] M. Momtaz, *et al.*, "Graph-based Approach to Text Alignment for Plagiarism Detection in Persian Documents," in *FIRE (Working Notes)*, 2016, pp. 176-179.
[40] F. Safi-Esfahani, *et al.*, "English-Persian Plagiarism Detection based on a Semantic Approach," *Journal of AI and Data Mining*, vol. 5, pp. 275-284, 2017.
[41] Y. A. Abdelrahman, *et al.*, "A Method For Arabic Documents Plagiarism Detection," *International Journal of Computer Science and Information Security*, vol. 15, p. 79, 2017.
[42] S. M. Alzahrani, *et al.*, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, pp. 133-149, 2012.