A Plagiarism Detection Approach Based on SVM for Persian Texts

Fezeh Esteki
Department of Computer Engineering,
Najafabad Branch,
Islamic Azad University,
Najafabad, Iran,
fe.esteki@Gmail.com

Faramarz Safi Esfahani
Department of Computer Engineering,
Najafabad Branch,
Islamic Azad University,
Najafabad, Iran
fsafi@iaun.ac.ir

ABSTRACT
Plagiarism is defined as an unauthorized act of using or adapting others' works and ideas without referring to them. Numerous methods have been proposed to detect plagiarism in different languages; however, not a lot has been accomplished in Persian. The present study has utilized statistical and semantic features to determine the functionality of Support Vector Machines (SVMs) in detecting acts of plagiarism in Persian. To increase accuracy, a stemmer was designed to stem Persian words. The statistical and semantic features were used to train and apply the SVM. The statistical features used are Jaccard coefficient, Dice coefficient, Levenshtein distance, and Longest Common Subsequence. To detect semantic similarities, a new method called "Index Words Replacement" was proposed. The proposed framework was tested on PAN data set. The results show the precision of 0.93337, recall of 0.70124 and Plagdet of 0.80083.

CCS Concepts
• Text mining→ Paraphrase detection→ Plagiarism detection→ Support Vector Machine

Keywords
Plagiarism detection; Text similarity; Paraphrase detection; SVM; Support vector machine; Persian texts; Farsi texts;

1. INTRODUCTION
Vast presence of e-texts on the web and their availability has caused a remarkable increase in plagiarism [15]. Plagiarism is divided into different categories including code stealing, paraphrasing, summarizing, translating and copying [3]. Changes made to a text can be lexical changes at which words are replaced with their synonyms, or structural changes at which sentence structure undergo some changes to make the piece untraceable [8]. To detect plagiarism in texts which have been changed thoroughly (have not been exactly copied), some more complicated methods are needed [13].

Plagiarism detection methods are divided into two main categories including Intrinsic and Extrinsic. Through extrinsic methods, a suspicious document is compared with some reference documents and similar parts are labelled as plagiarized. In intrinsic methods, no reference collection is used; parts whose writing style is different from the rest of the text are defined as plagiarized. The majority of researches conducted in plagiarism field have utilized extrinsic methods [23].

Numerous methods have been suggested to detect plagiarism among which statistical-based Techniques such as N-gram and Longest Common Subsequence [9], syntax-based techniques such as tree edit distance [6], semantic-based techniques such as WordNet Bi-Gram [22], etc. are noteworthy. One of the best ways to integrate a fine collection out of the mentioned methods is to use Machine Learning Algorithms.

There are not many methods to detect plagiarism in Persian texts. Among them we can point to [1] which used a combinational fuzzy approach and [18] which utilized Vector Space Model method. Determining the level of similarity in above-mentioned methods necessitates setting some threshold limits based on training data. This may lead to wrong results on other sets of data. To avoid this, machine learning algorithms can be utilized [7]. Among these algorithms, SVM has shown better performance over other algorithms [14]. It is notable that there is not any research which has applied SVM for plagiarism detection in Persian texts.

In the current study, the functionality of a Support Vector Machine (SVM) to detect plagiarism in Persian texts through extrinsic method has been studied. The SVM algorithm has shown superior applicability over other algorithms at solving 2-class classification problems with multiple features and data. Statistical techniques have been used to train the machine and apply it to detect plagiarism since they operate faster and show better performance in NLP applications. Jaccard coefficient, Dice coefficient, Levenshtein distance, and Longest Common Subsequence are the statistical techniques which will be briefly discussed in the following sections. The proposed framework works as follows: first, each of the mentioned techniques is applied to sentences in both suspicious and original documents. Outputs (mostly digits) are then compared and make an attribute vector to train the SVM. Using a set of attribute vectors, the SVM is trained and then applied to detect plagiarism. In addition to the mentioned methods, a new procedure called "Index Words Replacement" which uses a semantic Database named FarsNet [24] is proposed to detect semantic similarities. The proposed method chooses a word from a collection of synonyms to mark it as the index. It then detects words related to the index word in both original and suspicious text and replace them with the index. Using the mentioned techniques, level of similarity is detected at the final step.

2. RELATED WORK
There are a few methods for detecting similarities and plagiarism in monolingual Persian texts. Some of these methods can be found in [1] and [18].

In [1], a combinational fuzzy approach was used to detect plagiarism. This approach utilized a dataset for e-learning domain. Using the dataset jargon, the approach broke each sentence into

general and knowledge domain. Then it calculated Skip-gram, N-gram and Number of words for each part. At the end, it identified the level of similarity based on fuzzy rules.

Mahdavi et al. [18] used vector space model. This model transferred all texts in the database and the source text to attribute vectors. It then evaluated cosine similarity among semi-similar texts and finally detected similarity level by measuring the overlapping area of Tri-gram.

The SVM has not been used for detecting text similarities in Persian texts yet. In this research, we used SVM to detect plagiarism in monolingual Persian texts.

What follows reviews some researches which have adopted SVM to detect similarities between English texts.

In [2], a method which used SVM and Naïve Bayes techniques to detect plagiarism was employed. The study used fingerprint similarity, latent semantic analysis, word pair and word similarity attributes to train the SVM. The word similarity attribute was applied to detect similar words and the word pair was employed to detect non-similar words with similar meanings. The Results showed that plagiarism detection by SVM has more accuracy compared to that by Naïve Bayes.

Plagiarism detection by counting similar tokens and taking advantage of SVM method was the approach proposed in [25]. At the first step of this approach, words were labelled and named as Token. At the next level, sentence tokens were compared with each other; meanwhile, the tokens were being replaced with their similar words and comparison process was repeated over and over again to increase the accuracy of semantic similarity detection process. This approach used The Latent Semantic Analysis (LSA), Translation Edit Rate–Plus (TER-P), Dictionary-based Similarity, Maximum Similarity and BLEU attributes to train and test the SVM.

Similarity detection through the use of SVM by utilizing statistical and semantic attributes was the main goal of the proposed systems in [7] and [14]. In these systems, statistical similarity was detected using Skip-gram and LCS attributes and semantic similarity was determined using Noun/Verb Similarity Measure, Lin Similarity Measure, Cardinal Number Attribute and Proper Name Attribute. The semantic features were extracted using a Tree Tagging tool. This tool labels words based on the semantic relationships existed in WordNet database. The Cardinal Number Attribute was used to detect similarities between numbers and the Name Attribute Proper was applied to do so between names.

[5] proposed a system to detect similarities between sentences using SVM and statistical attributes. The system divided the sentences of both suspicious and similar documents into segments and then calculated the statistical attributes for each segment. This research proposed a new feature called EDU ((Elementary Discourse Unit) Similarity, and utilized BLEU, NIST, TER, TERP, METEOR and BADGER attributes parallel to EDU Similarity. This features usually used in automatic translation systems.

3. METHODOLOGY

The present study has proposed a new method based on SVM to detect plagiarism in Persian texts. Statistical attributes were used to train and apply the SVM. A brand-new approach called "Index Word Replacement" which will be discussed below was proposed to determine semantic similarities. The suggested algorithm was implemented in java language. The schema of the proposed approach is presented in figure 1. Different phases of the proposed plagiarism detection process will be discussed below.

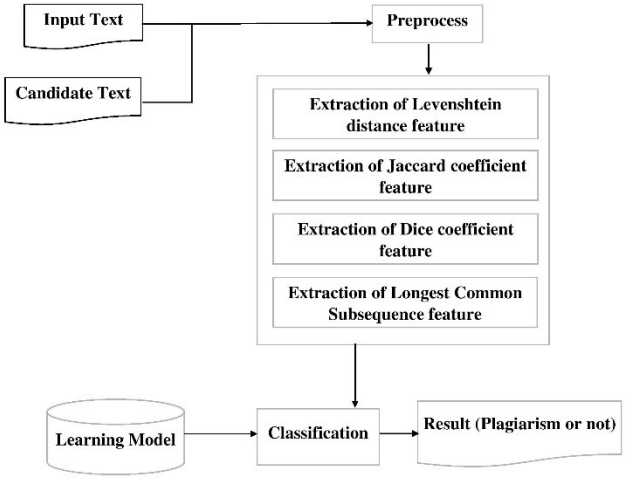

Figure 1. The schema of proposed approach in this study.

3.1 Preprocessing Phase

At this phase, original and suspicious texts were preprocessed. This phase includes normalization, stemming and eliminating stop words. The normalization process standardized the space between words, their forms and punctuation marks. Then, words were stemmed and their prefixes and suffixes were eliminated.

Since there was not a perfect stemmer implemented in Java programming language for Persian and the available stemmers did not eliminate some affixes and also the verbs were not stemmed by existing stemmers, in this study, a stemmer which stems the words based on part of speech categories was designed and used.

After stemming, stop words were eliminated by using the available stop word list. Stop words are frequently used words in a text. The majority of these words do not have an independent meaning and are used to establish relationships between main words.

3.2 Suggested Approach: Index Word Replacement

A great deal of methods proposed to detect plagiarism in English texts use **WordNet** lexicon database. A similar database has been created for Persian which is called **FarsNet** [24]. This database is very limited in terms of semantic relationships (hypernyms and hyponyms), so instead of extracting semantic features from the sentences by using this database, we suggested a new approach that uses synonyms to detect semantic similarity between sentences. The proposed approach operates as follows: an index word is chosen out of each collection of synonyms. Then, if there is a match for each of the index words in original and suspicious texts, it will be replaced by its related index word. For example, from the lexical set of "beautiful", "good looking" and "pretty", the word "beautiful" is selected as the index and it will be substituted with its similar words- "good looking" or "pretty"– if a text includes them.

3.3 Extracting Features from a Text

Jaccard similarity, Dice similarity, Levenshtein distance and Longest Common Subsequence are the statistical attributes which have been used in the present study. They are calculated as follows:

3.3.1 Levenshtein distance

Levenshtein distance between two strings is the minimum number of editing operators required to change one string into another. By editing operators, one means the operations of insertion, deletion and substitution of words. [16] It is calculated by equation 1.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \quad \textit{Equation 1}$$

3.3.2 Jaccard coefficient

For text document, the Jaccard coefficient is the number of common words to the number of words which are not common between two texts. [11]. The formula presented in equation 2.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \textit{Equation 2.}$$

3.3.3 Dice coefficient

The Dice measure is very similar to the Jaccard measure [17]. It is computed by equation 3.

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \quad \textit{Equation 3.}$$

3.3.4 Longest Common Subsequence

It measures the longest sequence of words with the same order in two string [22]. This attribute is calculated using Equation 4.

$$\mathrm{Sim}(A, B) = \frac{LCS(A,B)}{Min(|A| \; a \; |B|)} \quad \textit{Equation 4.}$$

3.4 SVM Training Process

At this phase, the non-similar sentence pairs were extracted from the xml files in "non-plagiarism" part of dataset and the similar sentence pairs were derived from the xml files in other parts of dataset. It is notable that in the utilized training dataset, instead of paragraphs, the sentences are tagged in xml files. After extracting the sentence pairs, the similar index words, explained in section3.2, were substituted in sentences. Then, the statistical techniques, mentioned in section3.3, were applied to sentences and the attributes for each pair were extracted. Using the numeric outputs, some training vectors were created to train the machine. Attribute vectors of similar sentences were labelled as "PLAGIARISM" class and those of non-similar sentences were labelled as "NONPLAIGIARISM" class.

3.5 Plagiarism Detection Using SVM

The statistical attributes of sentences were extracted from both original and suspicious texts and classified by SVM. If a pair is labelled as "PLAGIARISM", it means that the suspicious one has been stolen; otherwise, the sentence is considered as original.

3.5.1 Support Vector Machine (SVM)

Support vector machines which use statistical learning techniques were proposed by Vapnik in 1998. These algorithms utilize the strategy of maximizing the distance between a hyperplane and training samples to choose a proper hyperplane in order to classify data correctly. When there is noise in training data, using a liner classifier (hyperplane) for classifying training data is impossible, so primary samples are mapped into a higher space in non-linear form. In the new and larger space, data will be linearly classified by a kernel function using the proper hyperplane without raising the complexity of computations. In fact, kernel function uses the similarity between data in the original space to find similarities between vectors in a larger space. Kernel function can be selected form polynomial functions, RBF function, hyperbolic tangent or other proper functions. [12]

The present study has used a Support Vector Machine with RBF (Radial Base Function) kernel.

3.6 Evaluation

3.6.1 The utilized data set for training the SVM

The dataset used for training the SVM is the one used in [21]. The dataset consists of 3 different categories of text as follows: 1) TMC which contains news texts, 2) IRANDOC which includes texts related to sciences and technology and 3) Selected texts from Prozheh.com which consists of students' researches. Paraphrased texts which are based on this collection have been generated manually and mechanically and are divided into four categories: "Non-plagiarism" category that includes non-similar texts. "Synonyms replacement" category that contains texts whose words have been replaced with their synonyms. "Change Structured" category that includes texts whose sentence structures have been changed, and "Combined category" consists of texts with changed structures and word-replaced-with-synonym vocabulary.

3.6.2 The utilized data set for testing and evaluating the proposed system

Evaluation of the proposed system was conducted on a dataset introduced for Persian by [4] in PAN 2016 competition. This dataset contains a collection of Source and Suspicious document pairs which have been divided into four categories with regard to different levels of obfuscation. "No-obfuscation" category includes copied and clear stolen texts. "Random-obfuscation" group contains documents which have been changed artificially and by a machine. "Obfuscation-simulated" group keeps documents which have been changed manually and by man and bear high levels of obfuscation.

3.6.3 Evaluation results

Parameters used in PAN2016 include Recall, Precision and Plagdet which have been proposed by Potthast and colleagues [19], [20]. Also, the evaluation platform is designed by [10]. Evaluation results can be found in figure 2.

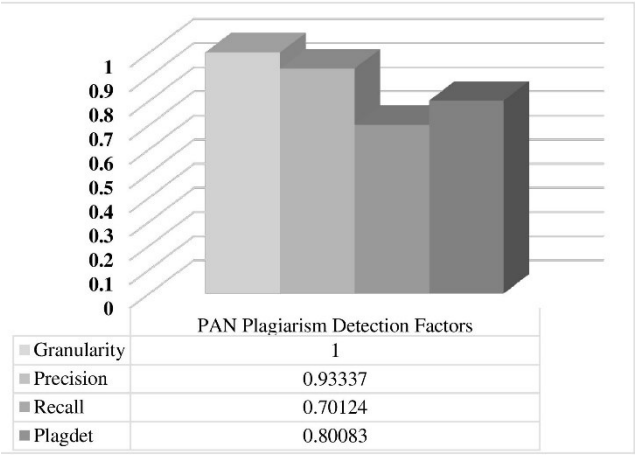

Figure 2. The evaluation results

4. CONCLUSION

An extrinsic SVM-based method was proposed to detect plagiarism in Persian texts in the current study. Also, the functionality and performance of SVM method to detect plagiarism in Persian texts was evaluated. To train the SVM, a combination of statistical attributes were used. A new approach called "Index Word Replacement" was suggested to detect semantic similarities. According to the results, it can be concluded that statistical methods operate effectively at similarity detection processes. As further suggestions, testing different statistical or semantic and syntactic attributes to train the machine and evaluating the following improvements to the system can be another field of further research.

5. REFERENCES

[1] Ahangarbahan, H. and Montazer, G.A. 2015. A Mixed Fuzzy Similarity Approach to Detect Plagiarism in Persian Texts. In *International Work-Conference on Artificial Neural Networks*. Springer.

[2] Alfikri, Z.F. and Purwarianti, A. 2014. Detailed Analysis of Extrinsic Plagiarism Detection System Using Machine Learning Approach (Naive Bayes and SVM). *Indonesian Journal of Electrical Engineering and Computer Science,* 2014. **12**(11): p. 7884-7894.

[3] Alzahrani, S.M., Salim, N. and Abraham, A. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 2012. 42(2): p. 133-149.

[4] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., and Potthast, M., 2016. Algorithms and Corpora for Persian Plagiarism Detection: Overview of PAN at FIRE 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[5] Bach, N.X., Le Minh, N. and Shimazu, A. 2014. Exploiting discourse information to identify paraphrases. *Expert Systems with Applications,* 2014. 41(6): p. 2832-2841.

[6] Bille, P. 2005. A survey on tree edit distance and related problems. *Theoretical computer science,* 2005. 337(1): p. 217-239.

[7] A. Chitra and C. Kumar. 2010. Paraphrase identification using machine learning techniques. In *Proceedings of the 12th international conference on Networking, VLSI and signal processing*, 2010: p. 245-249.

[8] Chong, M.Y.M. 2013. A study on plagiarism detection and plagiarism direction identification using natural language processing techniques.

[9] Dagan, I. and Glickman, O. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining,* 2004: p. 26-29.

[10] Gollub, T., Stein, B. and Burrows, S. 2012. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, August 2012. ACM. ISBN 978-1-4503-1472-5. p. 1125-1126.

[11] Huang, A. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*.

[12] Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer.

[13] Keck, C. 2006. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing,* 2006. **15**(4): p. 261-278.

[14] Kozareva, Z. and Montoyo, A. 2006. Paraphrase identification on the basis of supervised machine learning techniques. In *Advances in natural language processing*. 2006, Springer. p. 524-533.

[15] Leung, C.-H. and Chan, Y.-Y. 2007. A natural language processing approach to automatic plagiarism detection. In *Proceedings of the 8th ACM SIGITE conference on Information technology education.* 2007. ACM.

[16] Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady,* 1966, p. 707.

[17] Ljubešić, N., Boras, D., Bakarić, N., Njavro, J. 2008. Comparing measures of semantic similarity. In *Information Technology Interfaces, 2008. ITI. 30th International Conference on*. IEEE.

[18] Mahdavi, P., Siadati, Z. and Yaghmaee, F. 2014. Automatic external Persian plagiarism detection using vector space model. In *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE.

[19] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P. 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters*. 2010. Association for Computational Linguistics.

[20] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E. and Stein, B. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification,

and Author Profiling. In *Evangelos Kanoulas et al, editors, Information Access Evaluation meets Multilinguality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, Berlin Heidelberg New York, September2014. Springer. ISBN 978-3-319-11381-4. p. 268-299.

[21] Rakian, S., Esfahani, F.S. and Rastegari, H. 2015. A Persian Fuzzy Plagiarism Detection Approach. *Journal of Information Systems and Telecommunication (JIST),* 2015. **3**(3): p. 182-190.

[22] Roostaee, M., Fakhrahmad, S.M., Sadreddini M.H. and Khalili, A. 2014. Efficient calculation of sentence semantic similarity: a proposed scheme based on machine learning approaches and NLP techniques. *Scientific Journal of Review,* 2014. 3(3): p. 94-106.

[23] Seaward, L. and S. Matwin. 2009. Intrinsic plagiarism detection using complexity analysis. In *Proc. SEPLN*.

[24] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M. and Assi, S.M. 2010. Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference, Mumbai, India*.

[25] Vigneshvaran, P., Jayabalan, E. and Kathiravan, A.V. 2014. An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine. In *Intelligent Computing Applications (ICICA), 2014 International Conference.* IEEE.