Integrity Verification & Temper Detection of English Documents using Hybrid Structural Component and Word Length

Fatek Saeed, Anurag Dixit

Abstract: The zero text watermarking is a practical method for protecting the copyright from being tampered and copied. This paper aim to obtain zero watermarking technique with Hybrid Structural component and word length (HSW) where contains two phases watermark embedding and extraction. From the data owner the plain text is obtained and the watermark is generated. It is based on using the text characteristics. Certifying authority is generated as a group from text and watermark and recorded for tamper detection in file by pattern matching. Index terms: tamper detection, watermarking, and structural components.

I. INTRODUCTION

The tampering is a process of copying original content or changing the font style and stealing the content from original documents. To prevent the tampering, there is the need for copyrights protection. The text tampering is a type of tampering content from textual content such as e-books, e-journals, web pages and other text documents [1].

The text media that are exchanging through the internet is easy to copy the content of the media illegally, steal the copyrighted documents and so on. It's very important to protect the digital media from unauthorized users. The methods such as cryptography and steganography are used to protect the digital content on earlier days. The watermarking technology is more stable than the cryptography and steganography. It has more efficiency to identify the copyrights of the real content owner.

The watermark in the content damaged because of the alteration of unauthorized user is said to be an attack. Some of the attacks happening in the text watermarking are classified as watermark attack, geometrical attack and system attack. The watermark attack doesn't decode the original content properly and destroys the watermark. The geometrical attack attacks the text documents that are watermarked based on the invisible watermark. The attacker needs knowledge about the technique that is used for watermarking or watermarking key to remove the watermark [2].

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Mr. Fatek Saeed, School of Computer Science & Engineering, Galgotias University, Greater Noida, India. fateksaeed@gmail.com.
Prof.(Dr.) Anurag Dixit, School of Computer Science & Engineering, Galgotias University, Greater Noida, India.

Contribution of this work is given below:

The content of the digital medium is modified with the traditional watermarking techniques by enabling watermarking approaches. But for practical scenarios it is not applicable for the plain text. The specific need for plain text is accomplished with the method of zero watermarking. This approach doesn't update the properties of source details which uses for generating watermark information. A novel method HSW is developed in this paper for tampering detection, The tampering attacks can be avoided such as insertion, deletion and reordering and the information authenticity is proved with the help of proposed algorithm.

II. RELATED WORK

Reem A. Alotaibi [3] et al. proposed two invisible watermarking methods for Arabic text content. Which works based on the pseudo-space. It makes the connected characters isolated and synchronized with word space to hide binary bits '0' or '1'. On their first method, based on the dotting feature they place the pseudo-space before and after the word space. On the second method, pseudo-space is placed and in addition it add three small spaces to maximize the capacity. These spaces indicate the bits. The absence of small space indicates '0' and presence of small space indicates the '1'. They compared their proposed watermarking methods with other two existing methods and get more capacity and accuracy than the existing methods. They also tested their work with some attacks such as text copy and pasting, text tampering, and text formatting.

Yuling Liu [4] et al. proposed zero-watermarking algorithm based on merging features of sentences for Chinese text. They separated a text into set of sentences. Then the semantic code is obtained and with the frequency of the semantic code. this method extracted tree features from the text content and selected the main sentence by using the linear weight equation. They then selected the sentence with high weight of nouns and verbs to build the watermark. The watermark will be encrypted and registered on Certificate Authority (CA). They tested their results with some existing attacks such as sentence transformation attack, synonym substitution attack, and adding & deleting attack. They earned best results on these attacks.

Milad Taleby Ahvanooey [5] et al. designed an innovative technique for web text watermarking. Their proposed algorithm aims to protect the copyrights of web content and on theirs method they used some features of Unicode standard system.

They encrypted the web text content before embedding into the web page and convert it into zero-width control characters by using binary model classification. They used HTML as a cover file to embed the invisible watermarks. These invisible watermarks embed by their method tough against basic alteration or deletion attacks. Their method got higher prevention on basic text watermarking attacks and failed when zero-width control characters removed or deleted from the host webpage. They saying that their method is language independent and easily to extend their proposed method to different languages with their same structural features.

Milad Talebi Ahvanooei [6] et al. described a novel approach for text watermarking in digital documents by zero-width interword distance changes. The method proposed in this paper prevents copyrights tampering of text document by using watermarking of encryption message. Theirs method is based on placing the invisible watermark content before and after the text content. The watermarking based on their method support different languages except some languages such as Hindi, Malayalam, Canada and so on. At first, the each character is converted into 16-bit binary number. To detect the initial encrypted message, they insert two control characters with hex code and without Unicode. The encrypted message contains 1 or more 8 bit binary numbers with four hidden characters. In addition, they insert two invisible characters on the beginning of the paragraph and blank lines between the paragraphs. Their method got good performance on attacks of characters change, the main letters are modified and the proposed algorithm failed when the marked watermarked content in the Stego file get attacked.

Zhangjie Fu [7] et al. presented Verifiable Text Watermarking Detection to improve the digital watermarking security. For that they used a scheme called Zero knowledge-based watermark detection (ZKWD). The ZKWD is introduced for the plain text and to prevent the ambiguity attack by homomorphic property of asymmetric encryption algorithm. They include four step processes on their method to improve the security of watermark. They are, robust feature extraction method, watermarks generation method, watermarking embedding method and ZKWD. They saying theirs method ensure the security for watermark detection process.

III. PROPOSED ALGORITHM FOR TAMPERING DETECTION

The proposed approach of tampering detection is based on zero watermarking which uses the characteristics of text contents to generate watermark rather than embedding the watermark into the text. The proposed approach is based on structural component and word length and it is applicable for all kinds of documents. It consists of two stages namely text embedding and text extraction. The watermark generation can be done with the data owner and the extraction can be accomplished with the CA. The architecture of proposed tampering detection is shown in figure 1.

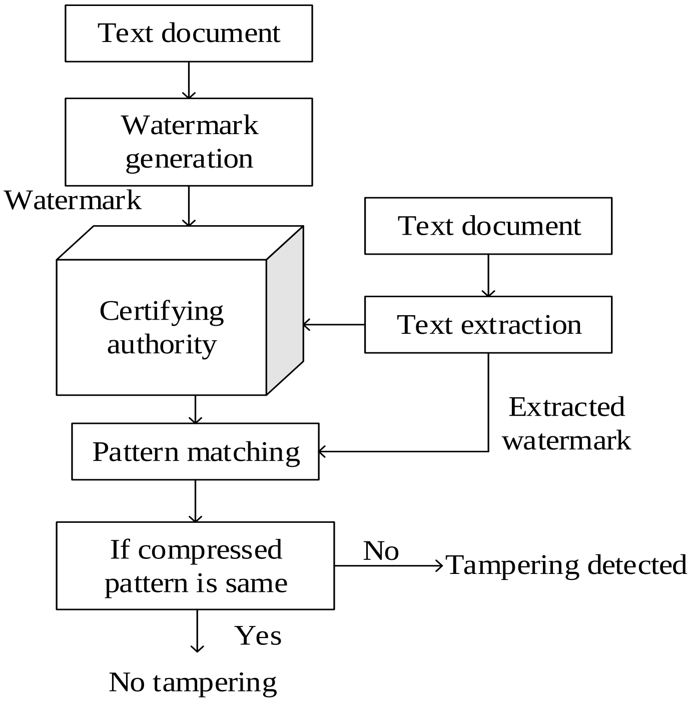


Fig 1: Overview of proposed algorithm

Figure 1 represented the overview of the proposed method.

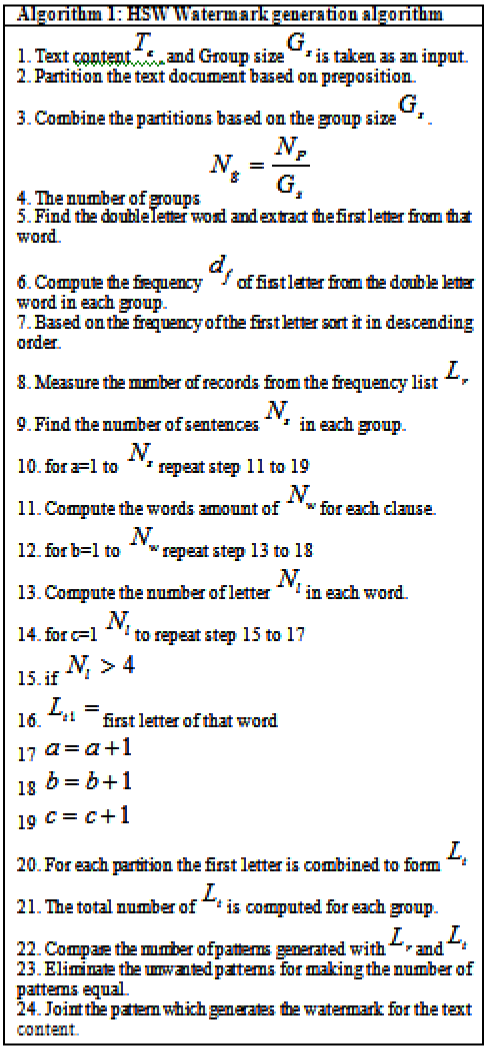
Retrieval Number: A1999109119/2019©BEIESP
DOI: 10.35940/ijeat.A1999.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The proposed algorithm takes text document as an input for tampering detection and the watermark is generated for the text document based on HSW approach. The extracted pattern of watermark is registered with the certifying authority (CA). The attacker may modify the content of the document. In the tampering detection process, the extraction algorithm is applied to extract the watermark and the pattern is matched with the registered pattern of CA. Based on the level of pattern matching with MD5 compression is set to take decision about tampering [8].

A. Watermark generation and embedding

The proposed watermarking generation is based on word length and structural components of the text document. Initially, the text partition is formed by considering preposition as a separator. Then the groups are formed by combining the partitions based on the size of the group. In order to form partition and group, the group size and preposition is taken as an input.



```
Algorithm 1: HSW Watermark generation algorithm
1. Text content $T_c$ and Group size $G_s$ is taken as an input.
2. Partition the text document based on preposition.
3. Combine the partitions based on the group size $G_s$.
4. The number of groups $N_g = \frac{N_p}{G_s}$
5. Find the double letter word and extract the first letter from that word.
6. Compute the frequency $d_f$ of first letter from the double letter word in each group.
7. Based on the frequency of the first letter sort it in descending order.
8. Measure the mmber of records from the frequency list $L_r$
9. Find the number of sentences $N_s$ in each group.
10. for a=1 to $N_s$ repeat step 11 to 19
11. Compute the words amount of $N_w$ for each clause.
12. for b=1 to $N_w$ repeat step 13 to 18
13. Compute the number of letter $N_l$ in each word.
14. for c=1 $N_l$ to repeat step 15 to 17
15. if $N_l > 4$
16. $L_{t1} =$ first letter of that word
17 $a = a + 1$
18 $b = b + 1$
19 $c = c + 1$
20. For each partition the first letter is combined to form $L_t$
21. The total number of $L_t$ is computed for each group.
22. Compare the number of patterns generated with $L_r$ and $L_t$
23. Eliminate the unwanted patterns for making the number of patterns equal.
24. Joint the pattern which generates the watermark for the text content.
```

Fig 2: HSW based watermark generation

The example of proposed watermark generation process is shown in figure 2. The frequency of first letter form the double letter word is identified to form the high frequency list. Then the letters are sorted in descending order which places the letter with highest value first.

Watermark embedding process:

The watermark pattern generated is registered with CA which is considered as a trusted third party of digital community. Embedding process produces the water marked digital image. This method is changed depending on the image is being processed. Embedded watermark represents the amount of text element characters. Figure 3 represented the process flow of an embedding process.

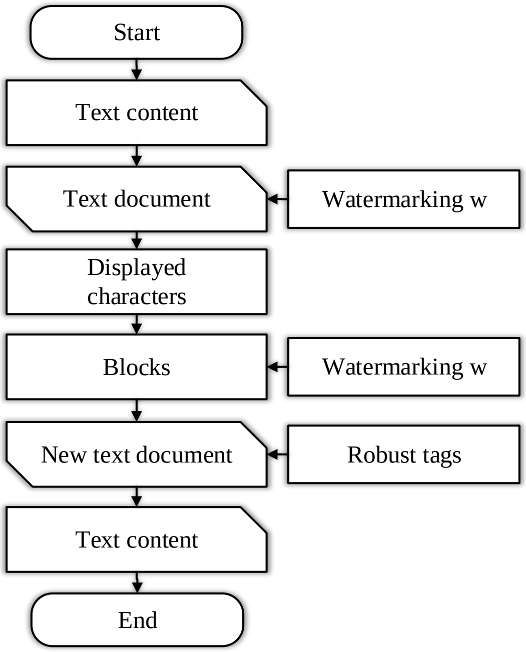
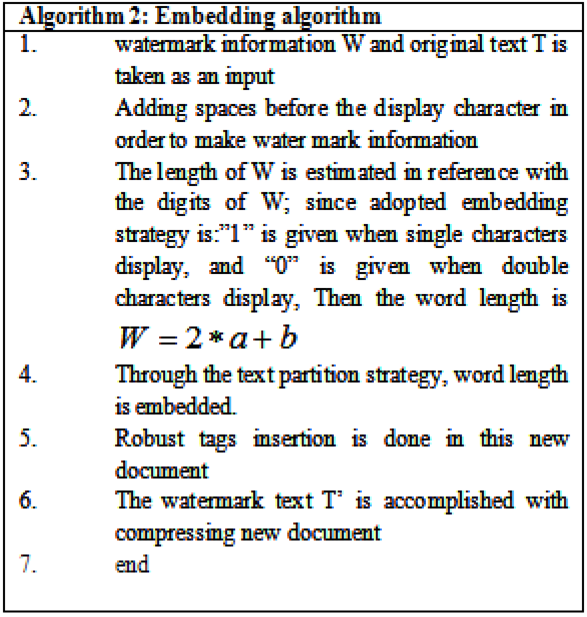


Fig 3: Process of embedding strategy

Tamper localization is achieved by representing “1” for the single display characters and “0” for the double display characters. This embedding technique inserts a watermark into a cover document. Algorithmic representation of embedding process is given in the below figure:



```
Algorithm 2: Embedding algorithm
1.    watermark information W and original text T is taken as an input
2.    Adding spaces before the display character in order to make water mark information
3.    The length of W is estimated in reference with the digits of W; since adopted embedding strategy is:"1" is given when single characters display, and "0" is given when double characters display, Then the word length is
      $W = 2 * a + b$
4.    Through the text partition strategy, word length is embedded.
5.    Robust tags insertion is done in this new document
6.    The watermark text T' is accomplished with compressing new document
7.    end
```

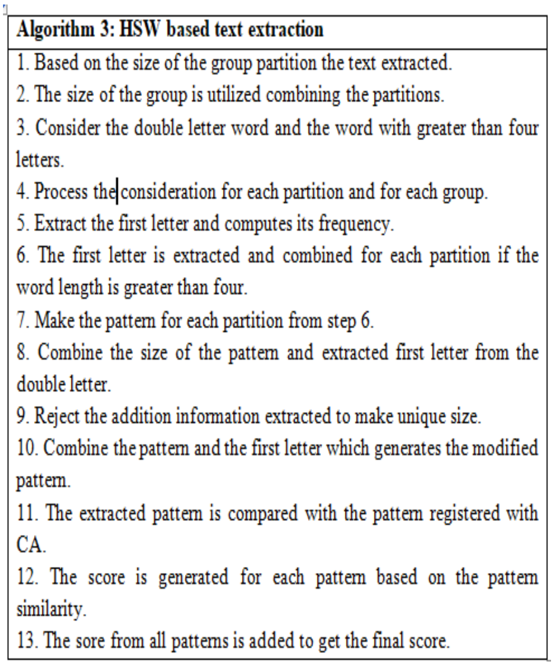
Fig 4: Embedding algorithm

When conflict arises, the extracted watermark pattern is analyzed with the text content to evaluate its authenticity.

The CA executed the detection algorithm and it provides response to the data owner who registered the watermark pattern. The text content may be attacked with several possible ways. The proposed algorithm detects tampering for insertion, deletion and reordering attacks.

B. Watermark extraction and tampering detection

The extraction algorithm extracts the watermark from the text content. The issue regarding copy right protection can be resolved with CA which keeps the extraction algorithm. Text extraction process is described in figure 5.



```
Algorithm 3: HSW based text extraction
1. Based on the size of the group partition the text extracted.
2. The size of the group is utilized combining the partitions.
3. Consider the double letter word and the word with greater than four letters.
4. Process the consideration for each partition and for each group.
5. Extract the first letter and computes its frequency.
6. The first letter is extracted and combined for each partition if the word length is greater than four.
7. Make the pattern for each partition from step 6.
8. Combine the size of the pattern and extracted first letter from the double letter.
9. Reject the addition information extracted to make unique size.
10. Combine the pattern and the first letter which generates the modified pattern.
11. The extracted pattern is compared with the pattern registered with CA.
12. The score is generated for each pattern based on the pattern similarity.
13. The sore from all patterns is added to get the final score.
```

Fig 5: HSW based text extraction

Watermark detection process

After removing the attacked pattern, the detection is achieved in three steps: Initial matching is done on the MD5 [10] of the attacked document and original document. When these two patterns with compressed patterns are same, then the document is represented the non-tampering pattern. When the initial matching is not done, then the word is known as attacked pattern.

Additional matching is done when associating the components with each state of the entire pattern. In order to calculate Pattern Matching Rate, transition and state of matching rate is determined. This procedure is defined through the following mathematical equations.

$$R_{PM_T} = \left| \frac{WM_O - (WM_O - WM_A)}{WM_O} \right| \qquad (1)$$

$$R_{PM_S} = \left| \frac{\sum_{j=1}^{n} \left(R_{PM_T} \right)}{Total\ state\ pattern\,count} \right| \qquad (2)$$

Here WM_O represented the water mark pattern of original document WM_A represented the attacked document. Wherever R_{PM_T} and R_{PM_S} means the pattern marking rate transition and state which is given in equation (1) and (2). From the Markov matrix, weight of every state is calculated and then the pattern matching rate is calculated for the attacked and the original document. The watermark distortion rate (R_{WD}),is calculated in equation (4), refers to tampering occurred with this document.

$$R_{PM} = \frac{\sum_{j=1}^{n} \left| \frac{\left(R_{PM_S}\, transition\ frequency \right)}{Total\ state\ pattern\,count} \right|}{n} \qquad (3)$$

$$R_{WD} = 1 - R_{PM} \qquad (4)$$

When there is no attack detected means it detects the generated watermark and this is referred as authentic text without tampering. In our algorithm original text files authenticates and used the text characteristics for attacks tampering detects.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed technique of tampering detection is simulated in MATLAB by varying the number of documents. The absolute score between original watermark pattern and extracted pattern is calculated to evaluate the degree of tampering. The work is evaluated for all possible attacks such as insertion, reordering and deletion. The detection accuracy is considered for the performance evaluation of the proposed method. The performance of the proposed method is compared with the existing approach in [9].

A. Attack measurements

Here we evaluated the deletion, insertion and reordering attacks.

- Deletion Attack

In order to select effective characters list (ECL), an effective rate is calculated for the different values such as 70%, 80%, and 90%. Here watermark distortion rate and pattern matching rate is evaluated and is given in table (1). It shows the detected text is small, medium or big. If the alteration level (DR) is larger than Zero referred the tampering data. Moreover, for all the tested cases the accuracy has been improved when the Effectiveness Ratio increased.

Table 1: Deletion attack measurement based on RPM and RWD

Category	Word count	Deletion volume	Effective ratio for characters list					
			70%		80%		90%	
			R_{PM}	R_{WD}	R_{PM}	R_{WD}	R_{PM}	R_{WD}
VLST	4647	10%	0.5933	0.4067	0.6334	0.3666	0.8002	0.1998
		20%	0.4281	0.5719	0.5441	0.4559	0.5772	0.4228
		50%	0.2115	0.7885	0.2703	0.7297	0.2998	0.7002
LST	2018	10%	0.772	0.2228	0.8381	0.1619	0.8449	0.1551
		20%	0.5618	0.4382	0.5977	0.4023	0.6125	0.3875
		50%	0.2632	0.7368	0.2689	0.7311	0.3755	0.6245

MST	**467**	**10%**	**0.8091**	**0.1909**	**0.8113**	**0.1887**	**0.8535**	**0.1465**
		20%	**0.5806**	**0.4194**	**0.6273**	**0.3727**	**0.6640**	**0.3360**
		50%	**0.2992**	**0.7008**	**0.3186**	**0.6814**	**0.3255**	**0.6745**
SST	**179**	**10%**	**0.8136**	**0.1864**	**0.8301**	**0.1699**	**0.8792**	**0.1208**
		20%	**0.6288**	**0.3712**	**0.6531**	**0.3469**	**0.7371**	**0.2629**
		50%	**0.3177**	**0.6823**	**0.2986**	**0.7014**	**0.3510**	**0.6490**

- Insertion Attack

Table 2 signifies insertion attack with the values of matching rate and distortion rate with the same input as the previous experiment.

Table 2: Insertion attack measurement based on RPM and RWD

Category	**Word count**	**Insertion volume**	**Effective ratio for characters list**					
			70%		80%		90%	
			R_{PM}	**R_{WD}**	**R_{PM}**	**R_{WD}**	**R_{PM}**	**R_{WD}**
VLST	4647	10%	0.3183	0.6817	0.5227	0.4773	0.5448	0.4552
		20%	0.2100	0.7900	0.2392	0.7608	0.3106	0.6894
		50%	0.3715	0.6285	0.3938	0.6062	0.4118	0.5882
LST	2018	10%	0.6117	0.3883	0.6333	0.3667	0.6693	0.3307
		20%	0.5255	0.4745	0.5382	0.4618	0.5571	0.4429
		50%	0.2210	0.7790	0.2863	0.7137	0.3030	0.6970
MST	467	10%	0.7677	0.2323	0.7726	0.2274	0.8081	0.1919
		20%	0.6117	0.3883	0.6552	0.3448	0.7115	0.2885
		50%	0.3007	0.6993	0.3775	0.6225	0.4123	0.5877
SST	179	10%	0.7919	0.2081	0.8546	0.1454	0.8766	0.1234
		20%	0.6421	0.3579	0.6770	0.3230	0.7628	0.2372
		50%	0.3409	0.6591	0.3551	0.6449	0.3831	0.6169

The reordering results shows that the accuracy of proposed tamper detection. Related to deletion attack, the tamper detection's precision becomes more adequate through increasing the effective rate value. Thus, the procedure is more or less effective for Very Large Size Text and Large categories.

- Reordering Attack

We have achieved three levels of arbitrarily reordering attack on every sample document of our dataset.

Table 3: Reordering attack measurement based on RPM and RWD

Category	**Word count**	**Reordering volume**	**Effective ratio for characters list**					
			70%		80%		90%	
			R_{PM}	**R_{WD}**	**R_{PM}**	**R_{WD}**	**R_{PM}**	**R_{WD}**
VLST	4647	10%	0.3183	0.6817	0.5227	0.4773	0.5448	0.4552
		20%	0.4552	0.7900	0.2392	0.7608	0.3106	0.6894
		50%	0.2554	0.6285	0.3938	0.6062	0.4118	0.5882
LST	2018	10%	0.3889	0.5448	0.5387	0.4613	0.5701	0.4299
		20%	0.5255	0.7446	0.2735	0.7265	0.3431	0.6569
		50%	0.2210	0.6111	0.4007	0.5993	0.4130	0.5870
MST	467	10%	0.7932	0.2068	0.8117	0.1883	0.8443	0.1557
		20%	0.6421	0.3579	0.6729	0.3271	0.7073	0.2927
		50%	0.3498	0.6502	0.3892	0.6108	0.4293	0.5707
SST	179	10%	0.8447	0.1553	0.8559	0.1441	0.8802	0.1198
		20%	0.6931	0.3069	0.7223	0.2777	0.7781	0.2219
		50%	0.3807	0.6193	0.4112	0.5888	0.4375	0.5625

Table 3 represented the reordering attack based on pattern matching and watermark distortion ratio. We have smeared our current work with different ER values. The distortion and matching rates of extracted pattern related to the watermark of this simulation are given in Table (3). It can be obviously perceived that the matching and distortion values for reordering attack are even more correct than both attacks deletion and insertion. Where significance of DR represents text has not secured and it has been tampered. It is evident that the watermark instability shows that text has been attacked.

B. Accuracy measurements

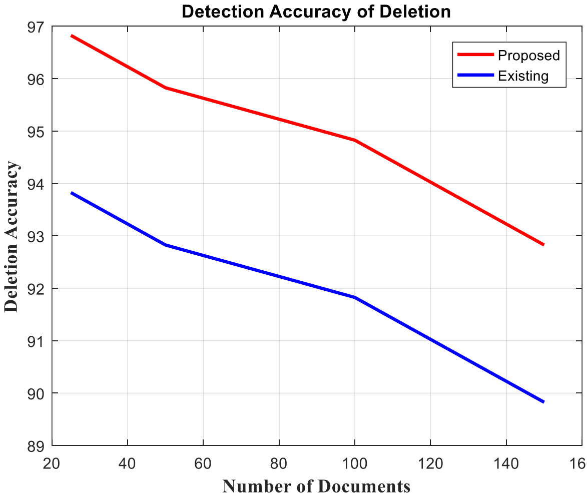


Fig 6: Detection accuracy comparison by varying the number of documents

Figure 6 shows the detection accuracy of detection attack by varying the number of documents. Numbers of documents are varying from 20 to 160 to evaluate the proposed detection performance. The deletion accuracy is evaluated for the number of documents 20, 40, 60, 80, 100, 120, 140 and 160. The detection accuracy for the deletion task is above 97 for the numbers of documents are 20. For the existing approaches, this detection can be lower than the proposed one. By increasing the number of documents the detection accuracy is reduced. When the numbers of documents are 150, the detection accuracy for the deletion task is between 92 and 93. For the existing detection technique, these values are reduced below 90.The detection accuracy for the insertion task is shown in figure 4. The accuracy of detection is higher than the existing based on the performance evaluation.

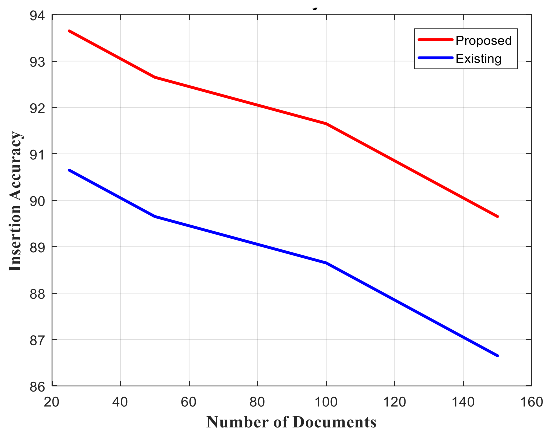


Fig 7: Insertion accuracy comparison by varying the number of documents

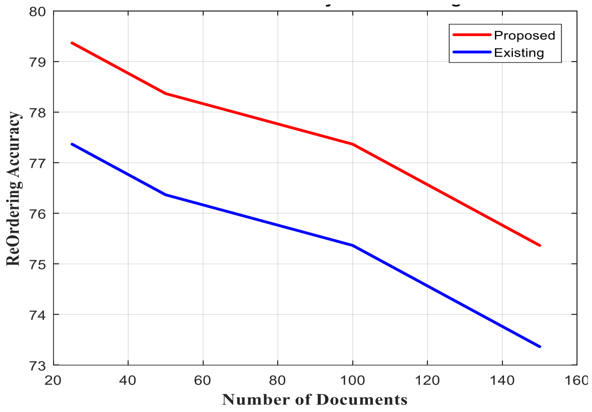


Fig 8: Reordering accuracy comparison by varying the number of documents

The insertion accuracy and reordering accuracy for the task is shown in figure 7 and 8. When the numbers of documents are 20, the detection accuracy is 90 and 80. By increasing the number of document to 40, 60, 80, 100, 120, 140 and 160 the accuracy for the proposed approach is reduced to 89.5, 88.7, 88.3, 88, 87, 86.3 and 86. By increasing the number of document to 40, 60, 80, 100, 120, 140 and 160 the accuracy of existing approach is reduced to 88, 87.5, 86.7, 86.3, 86, 85.2, 84.3 and 84.

V. CONCLUSION

Firstly, the plain text is achieved from the writer details and the script Meta data is analyzed. Secondly, watermark is generated with our novel algorithm. The text partition is grouping depends on the group size. Next, generate the MOL depends on groups of the letter or word, this list is used for the building of watermark key based on characteristics of plain text, where this key combined with CA, original plain text, and author details. The HSW detect algorithm used for pattern creation. Each extracted pattern is compared with the pattern registered with the certifying authority and in which distortion rate is calculated for detecting the tampering attacks.

REFERENCES

1. Leena Goyal, Manoj Raman, Prateek Diwan, Mukesh Vijay, Leena Goyal, Manoj Raman, Prateek Diwan, and Mukesh Vijay. "A robust method for integrity protection of digital data in text document watermarking." Int. J. Sci. Res. Dev 1, no. 6 (2014): 14-18.
2. Maria Chroni, and Stavros D. Nikolopoulos. "Watermarking PDF Documents using Various Representations of Self-inverting Permutations." arXiv preprint arXiv:1501.02686(2015).
3. Reem A. Alotaibi, and Lamiaa A. Elrefaei. "Improved capacity Arabic text watermarking methods based on open word space." Journal of King Saud University-Computer and Information Sciences (2017).
4. Yuling Liu, Yinghong Zhu, and Guojiang Xin. "A zero-watermarking algorithm based on merging features of sentences for Chinese text." Journal of the Chinese Institute of Engineers 38, no. 3 (2015): 391-398.
5. Milad Taleby Ahvanooey, Hassan Dana Mazraeh, and Seyed Hashem Tabasi. "An innovative technique for web text watermarking (AITW)." Information Security Journal: A Global Perspective 25, no. 4-6 (2016): 191-196.
6. Milad Talebi Ahvanooei, Seyed Hashem Tabasi, and Sajad Rahmani. "A novel approach for text watermarking in digital documents by zero-width interword distance changes." DAV International Journal of Science 4, no. 3 (2015): 550-558.
7. Zhangjie Fu, Xingming Sun, Jiangang Shu, Lu Zhou, and Jin Wang. "Verifiable Text Watermarking Detection to Improve Security." International Journal of Security and Its Applications8, no. 5 (2014): 1-10.
8. Fatek Saeed and Anurag Dixit, “Hybrid HSW Based Zero Watermarking for Tampering Detection of Text Contents”, Lecture Notes on Data Engineering and Communications Technologies Vol 31, Springer, Cham Switzerland AG 2020.
9. Xin G, Qi X, Ding C. An Improved Tamper Detection and Location Scheme for DOCX Format Documents. International Conference on Cloud Computing and Security 2018 Jun 8 (pp. 242-251). Springer, Cham.
10. den Boer, B. and Bosselaers, A., 1993, May. Collisions for the compression function of MD5. In Workshop on the Theory and Application of Cryptographic Techniques (pp. 293-304). Springer, Berlin, Heidelberg.

AUTHORS PROFILE

Mr. Fatek Saeed, did Master Degree from Arab Academy ,Jordon, Sanaa branch, he is doing Ph.D. in Computer Apps His interests includes E-Learning , Cloud Computing ,AI ,Data Mining, Information System Management, Information Security, Accreditation & quality Assurance, Higher Education Management, institutions evaluation and monitoring. He has so many international Certificates and he has travelled to Syria, Jordon, Philippines, UAE, KSA, India and Oman.

Dr. Anurag Dixit, did Ph.D. in Computer Science From Jawaharlal Nehru University (JNU), New Delhi in Jan 2010. He has teaching and research experience of more than 20 years. Presently he is Associated with Galgotias University as Professor and Head (MCA) Since April 2013. Have experience as the Director, Dean and Professor in reputed University VIT ,Vellore ,MITS, Sikar , Regional Engineering College ..

His interests include AI , Machine Learning , Cloud Computing Security, Software Engineering . Based on his research work he had so far published more than 70 technical publications in reputed International/National Journals and Conferences and applied few Patents. He has travelled widely to countries like USA, UK, France, Italy , Russia , Tanzania , Thailand, UAE etc.