Proposing High-Smart Approach for Content Authentication and Tampering Detection of Arabic Text Transmitted via Internet*

Fahd N. AL-WESABI††, Member

SUMMARY The security and reliability of Arabic text exchanged via the Internet have become a challenging area for the research community. Arabic text is very sensitive to modify by malicious attacks and easy to make changes on diacritics i.e. Fat-ha, Kasra and Damma, which are represent the syntax of Arabic language and can make the meaning is differing. In this paper, a Hybrid of Natural Language Processing and Zero-Watermarking Approach (HNLPZWA) has been proposed for the content authentication and tampering detection of Arabic text. The HNLPZWA approach embeds and detects the watermark logically without altering the original text document to embed a watermark key. Fifth level order of word mechanism based on hidden Markov model is integrated with digital zero-watermarking techniques to improve the tampering detection accuracy issues of the previous literature proposed by the researchers. Fifth-level order of Markov model is used as a natural language processing technique in order to analyze the Arabic text. Moreover, it extracts the features of inter-relationship between contexts of the text and utilizes the extracted features as watermark information and validates it later with attacked Arabic text to detect any tampering occurred on it. HNLPZWA has been implemented using PHP with VS code IDE. Tampering detection accuracy of HNLPZWA is proved with experiments using four datasets of varying lengths under multiple random locations of insertion, reorder and deletion attacks of experimental datasets. The experimental results show that HNLPZWA is more sensitive for all kinds of tampering attacks with high level accuracy of tampering detection.

key words: NLP, hidden Markov model, zero-watermarking, Arabic text analysis, content authentication, tampering detection

1. Introduction

More than 400 million people exchange information online in the Arabic language. Content authentication and integrity verification of digital Arabic text have gained great importance in communication technologies. Numerous applications such as e-commerce, e-Banking etc. impose challenges during transfer of contents via internet [1], [3]. Most of the digital media transferred over the internet is in text form and is very sensitive to transfer online in terms of contents, structure, syntax, and semantic. Malicious attackers may temper these digital contents during the transfer process and thus, the modified contents can result in wrong decision [3].

Several algorithms and techniques are available for information security such as content authentication, integrity verification, tampering detection, owner identification, access control and copyright protection.

Steganography and digital watermarking techniques are commonly used to solve these problems. Digital watermarking (DWM) is a technology in which different information such as text, binary image, video and audio can be embedded as a watermark key in digital contents [2], [3]. Based on homoglyph characters substitution for Latin symbols and whitespaces, a fine-grain text watermarking method is proposed [4].

Several traditional text watermarking methods and solutions have been proposed [5] and classified in various categories such as structure based, linguistic based, binary image based and format based [6]. Most of these solutions require some modifications or transformations on original digital text contents in order to embed the watermark information within text. Zero-watermarking is a modern technique that can be used with smart algorithms without any modification on original digital contents to embed the watermark information. Moreover, in this technique the contents of given digital context can be utilized to generate the watermark information [1], [6]–[8].

Limited researches have focused on the necessary applicable solutions for integrity verification of sensitive online digital media [9]–[11]. Authentication and tampering detection of digital text have gained great attention in the research community. Moreover, in the last decade the research in the field of text watermarking has focused in copyright protection, but there was less interest and attention on integrity verification, tampering detection and content authentication due to the natural language dependent nature of text contents [12].

Most common challenges in this area are to propose the most suitable techniques and solutions for different formats and contents specially in English and Arabic languages [13], [14]. Hence, content authentication, integrity verification and tampering detection of sensitive text constitutes a big problem in various applications and requires necessary solutions.

Digital Holy Qur’an in Arabic, eChecks, online exams and marks are some cases of such sensitive digital text contents. Various features of Arabic alphabets like diacritics, extended letters, and other Arabic symbols make it easy to change the main meaning of text contents by making simple modifications such as changing diacritics arrange-
This paper presents an intelligent hybrid approach for content authentication and tampering detection of Arabic text, called HNLPZWA. The proposed technique combines Markov Model as NLP and zero watermarking. The fifth level order of word mechanism of Markov model has been used for text analysis in order to extract the interrelationships between contents of the given Arabic text which consequently generates the watermark key. The generated watermark is logically embedded in the original Arabic context without modifications of the original text. After transmission of the text, the embedded watermark is used to detect any tampering occurred on received Arabic text and ensures the authenticity of the transmitted text.

The major objective of HNLPZWA approach is to achieve high accuracy of content authentication and sensitive detection of tampering attack in Arabic text that is transferred via Internet. The main contributions of HNLPZWA are as follows:

- Combined zero text watermarking and NLP approach has been developed for content authentication and tampering detection of digital Arabic contents have been ignored by researchers in the literature for the main reasons that Arabic text is natural language dependent and the complexity of hiding the watermark information which there is no locations to hide it within text as pixels in case of image, waves in audio and frames in video.
- Unlike the previous work, in which the watermarking is performed by effecting text, content, and size, HNLPZWA approach embeds the watermarking logically without any effect on the text, content, and size.
- In HNLPZWA approach, watermarking does not need any external information because the watermark key is produced as a result of text analysis and extracting the relationship between the content itself and then making it as a watermark.
- HNLPZWA approach is highly sensitive to any simple modification on the text and the meaning in the Arabic text, which is known as complex text, including the Arabic symbols which can change the meaning of the Arabic word. The three contributions mentioned above are found somehow only in images but not in text. This is the vital point concerning to the contribution of this paper.
- HNLPZWA is compared to other baseline approaches, performed implementation of zero programs, and an extensive experiments using various scenarios of Arabic datasets under main text attacks and volumes.
- By studying and analyzing the results, it was observed that HNLPZWA outperforms the baseline approaches in terms of tampering detection accuracy.

The rest of the paper has four more sections. Section 2 explains the existing works done so far. Section 3 presents HNLPZWA, Section 4 describes the implementation, experiment setup, results and discussion, and finally, Sect. 5 concludes the article.

2. Literature Review

The text watermarking methods and solutions proposed so far in the literature are based on a number of features and watermark embedding modes. In this paper, survey of the most common classifications of watermarking schemes has been performed, which involve linguistic based, structural based, and zero-watermark based solutions [3], [6], [12].

2.1 Linguistic-Based Techniques

Linguistic text watermarking approaches are natural language dependent. Mechanism of these techniques to embed the watermark relies on modifications applied to semantic and the syntactic nature of plain text [3].

A text watermarking method based on open word space [16] has been proposed to improve the capacity and imperceptibility of Arabic text. In this approach, each word space is utilized in order to hide the binary bit 0 or 1 through which physical modification on original text is done.

A text steganography technique [17] has been proposed to hide the information in Arabic language. Mechanism of this technique is considering the existence of Harakat (dialectics, i.e. Fat-ha, Kasra and Damma) in Arabic language and reversing the Fatha for message hiding.

A Kashida-marksinvisible watermarking technique [18] has been proposed which is based on frequency recurrence features of the characters for document protection and authentication. This technique relies on predefined watermark key whereby a Kashida is placed for a bit 1 and omitted for a bit 0. Text steganography methods [19], [20] have been proposed for writing digital contents of Arabic and English languages using extensions Kashida based on the “moon” and “sun” characters and using English alphabet Unicode. In this method, Kashida characters used alongside Arabic characters to determine which specific characters is holding the hidden secret bits.

2.2 Structural-Based Techniques

Structural text watermarking approaches are based on content dependent structure in which modifications on feature or structure of plain text are done for watermark embedding process [21]–[24].

Text watermarking algorithm [21] has been proposed to protect text contents from malicious attacks based on Unicode extended characters. This algorithm encompasses three core processes, which are watermark generation, embedding and extraction. Watermark embedding depends on construction of predefined encoding tables while scrambling techniques are used in generation and extraction to protect the watermark key. The replacement attack approach [22] has been proposed which is based on maintaining words locations in text document. This approach relies on ex-
exploit transitions of words in text document. Text based watermarking approaches [23], [24] have been proposed for Chinese text document authentication based on merging properties of sentences. Mechanism of these approaches is as follows: firstly, a Chinese text is divided into sets of sentences, and then a semantic code is obtained for each word. Entropy of a sentence is calculated by semantic codes frequency. However, sentence relevance is calculated by similarity of semantic between words through the tree structure of words in Tongyici Cilin. Finally, weight of each sentence is obtained by utilizing entropy, sentence relevance, sentence length, and a weighting function. Watermark is constructed by using the nouns and verbs of the high weight sentences, which is encrypted and registered with a trusted third party Certificate Authority (CA).

2.3 Zero Watermark-Based Techniques

Zero watermark-based techniques are dependent on text features. Several methods and solutions of Zero-Text based watermarking have been proposed as in the studies [12], [15], [25]–[29].

AnITH algorithm [12] has been proposed for validation of accuracy and reliability of digital texts shared on social media. This algorithm conceals an invisible watermark within a digital text and can be extracted later to check the accuracy and reliability of text contents. A zero watermark-based method [15] has been proposed to ensure the data integrity over the internet of things in which watermark is embedded in plain text before its transmission. Generated watermark key is based on some features of contents such as data size, data appearance frequency, and time of data capturing. These two measures are calculated to determine the unvoiced frames for watermark embedding. The mechanism of this method depends on embedding the identity of an individual without inform any distortion in medical speech signals.

A zero-watermarking approach [26] has been presented to address the problems associated with English text documents protection, such as copyright protection and content verification. A zero watermarking approaches [27], [28] have been proposed which are based on Markov model for content authentication of English text. In this approach, the probability properties of the English text are used to extract the secure watermark information and stored to check the authenticity of attacked text document. These approaches offer protection against common text attacks with distortion rate of a watermark if it is greater than one for all identified attacks. Traditional watermark method [29] has been proposed for copyright protection of English text based on appearance frequency of non-vowel ASCII letters and words.

3. The Proposed Approach

This paper proposes a smart hybrid approach by integrating NLP and zero-watermark techniques so there is no need of extra information and modifications on the original Arabic text to embed as watermark key. Fifth-level order of word mechanism of Markov model is used as a NLP technique to analyze the contents of Arabic text and extract the interrelationships features of these text contents. The following sets of assumptions are created by HNLPZWA approach:

- Watermark key information will be generated as a feature of the given Arabic text and embedded without any required modifications in the original contexts and data size.
- Attacks may be as insertion, deletion or reorder.
- Attacks volumes may be very low, low, mid or high.
- Tampering may occur randomly in different locations of given Arabic text.
- Arabic text may be very small, small, medium or large in size.
- Matching or distortion rate of tampering may be very low, low, medium or high.

The following subsections explain in detail two main processes that should perform in HNLPZWA the first one called text analysis and watermark generation process, and second one called watermark extraction and detection process.

3.1 Text Analysis and Watermark Generation Process

The three main sub-algorithms included in this process are pre-processing algorithm, text analysis algorithm, and watermark generation algorithm as illustrated in Fig. 1.

3.1.1 Pre-Processing Algorithm

The preprocessing of the original Arabic text is one of the
key steps in both the watermark generation and extraction processes to remove extra spaces and new lines, and it will directly influence the tampering detection accuracy. The original Arabic text (OAT) is required as input for Pre-processing process.

3.1.2 Text Analysis Algorithm

This algorithm include two sub processes are building Markov matrix, and text analysis processes.

- Building a Markov matrix is the starting point of Arabic text analysis and watermark generation process using Markov model. A Markov matrix that represents the possible states and transitions available in a given text is constructed without reputations. In this approach, each unique alphanumeric within a given Arabic text represents a present state, and each unique alphanumeric a transition in Markov matrix. During the building process of Markov matrix, the proposed algorithm initializes all transition values by zero to use these cells later to keep track of the number of times that the \(i\)th unique combined of five of words is followed by the \(j\)th word within the given Arabic text.

Pre-processing and building Markov matrix algorithm executes as presented below in Algorithm 1.

**Algorithm 1.** Pre-processing and building Markov algorithm of HNLPZWA

\[
\text{PROCEDURE Pre_BMM (OAT)}
\]

1. Input: original Arabic text (OAT)
2. Output: Markov matrix with zeros initial value
3. BEGIN
4. // perform pre-processing process
5. for each word in OAT
6. OAT \(\rightarrow\) removes ("space" or "newLine")
7. // Build list of non values text words
8. \(W_5 \text{mm}[ps][ns] = []\)
9. for each word in OAT
10. if word not in \(W_5 \text{mm}[ps][ns]\)
11. \(W_5 \text{mm}[ps][ns] \leftarrow W_5 \text{mm}[ps][ns] \cup \{\text{word}\}\)
12. for ps \(= 1\) to \(W_5 \text{mm}[ps][ns]\) length - 5
13. for ns \(= 1\) to \(W_5 \text{mm}[ps][ns]\) length
14. \(W_5 \text{mm}[ps][ns] = 0\)
15. return \(W_5 \text{mm}[ps][ns]\)

where,

- OAT: is an original Arabic text, \(OAT_{ps}\): is a pre-processed Arabic text, \(W_5 \text{mm}[ps][ns]\): states and transitions matrix with zeros values for all cells, \(ps\): refers to current state, \(ns\): refers to next state.

According to the above, a method is presented to construct two-dimensional matrix of Markov states and transitions named \(W_5 \text{mm}[ps][ns]\), which represents the backbone of Markov model for Arabic text analysis. The length of \(W_5 \text{mm}[ps][ns]\) matrix of HNLPZWA is dynamic in which depends on the context of given Arabic text, and it is equal to the number of the five continues words without repetition.

- Text analysis process: after the Markov matrix was constructed, NLP and text analysis process should be performed to found interrelationships between contexts of the given Arabic text. In this algorithm, the number of appearances of possible next states transitions for each current state of the package of unique five words will calculated and constructed as transition probabilities by Eq. (1) below.

\[
W_5 \text{mm}[ps][ns] = \sum_{i,j=1}^{n} \text{Total number of transitions (i, j)}
\]

where,

- \(n\): is total number of states.
- \(i\): is \(i\)th current state of the package of unique five words.
- \(j\): is \(j\)th next state transition.

The following example of Arabic text sample describes the mechanism of the transition process of present state to other next states.

فيتفرج الحفلات البيني السريع فوق التعب البيني الطبي، للوصول إلى التعب البيني البسيط.

When using the fifth level order of word mechanism of Hidden Markov model, every unique five of words is a present state. Text analysis is processed as the text is read to obtain the interrelationships between the present state and the next states. Figure 2 below illustrates the available transitions of the above sample of Arabic text.

As a result of analysing the given Arabic sentence based on the fifth level order of word mechanism of Markov model, all present states and their possible transitions are represented as illustrated in Fig. 3.

Author assume here that “فيتفرج الحفلات البيني السريع فوق التعب البيني البسيط” is a present state combined of five words, and the available next
Algorithm 2. Text analysis algorithm of HNLPZWA

where,

- pw: previous word, nw: current word,
  W5\_imm[ps][ns] refers to initial Markov matrix with zero values.

3.1.3 Watermark Generation Algorithm

After Arabic text analysis has been performed and probability features were extracted, watermark key is generated as watermark patterns by finding all non-zeros values in Markov matrix. All of this non-zeros values will be concatenated sequentially to generate the original watermark pattern W5\_WM0, as given in equation no 2 and illustrated in Fig. 5.

\[
W5\_WM0 = W5\_imm[ps][ns], \text{ for } i, j = \text{non-zeros values resulted in } w5\_imm
\]  

(2)

Watermark embedding will have performed logically in this approach without need to made any modification within the original text. The generated W5\_WM0 is stored in watermark database beside basic information of Arabic text. The generated watermark sequential patterns are then digested by using MD5 Hash algorithm to find a secure watermark form and improve the watermark capacity, and they are denoted as W5\_DM0, notational as given in equation no 3 and illustrated in Fig. 6.

\[
W5\_DM0 = MD5(W5\_WM0)
\]  

(3)

Algorithm of watermark generation based on fifth level order of word mechanism of Markov model is presented formally and executed as illustrated below in Algorithm 3.

3.2 Watermark Extraction and Detection Algorithms

Before the detection of pre-processed attacked Arabic text (AAT), attacked watermark patterns (W5\_WM0) should be generated, and matching and distortion rate of watermark patterns should be calculated by HNLPZWA for detecting any tampering occurred in the given contents. Two core algorithms are involved in this process, which are watermark extraction and watermark detection. However, W5\_WM0 will be extracted from the received AAT and matched with W5\_WM0 by detection algorithm. AAT should be provided as the input for the proposed watermark extraction algorithm. The same process of watermark generation algorithm should have been performed to obtain the watermark pattern for (AAT) as illustrated in Fig. 7.

3.2.1 Watermark Extraction Algorithm

AAT is the main input required to run this algorithm. However, the output of this algorithm is W5\_WM0. The watermark extraction algorithm is presented formally and executed as illustrated in Algorithm 4.

Algorithm 4. Watermark extraction algorithm of HNLPZWA

where,
3.2.2 Watermark Detection Algorithm

$W_{5,WMA}$ and $W_{5,WMO}$ are the main inputs required to run this algorithm, while the output of this algorithm is the notification Arabic text document, which can be authentic or tampered. Detection process of extracted watermark is achieved in two main steps:

- **Primary matching** is achieved for $W_{5,WMA}$ and $W_{5,WMO}$. If these two patterns appear identical, then an alert will appear as “Arabic text is an authentic and no tampering occurred”. Otherwise, the notification will be “Arabic text is tampered and not authentic”, and then it continues to the next step.

- **Secondary matching** is achieved by matching the transition of each state in the whole generated pattern. This means $W_{5,WMA}$ of each state will be compared to equivalent transition of $W_{5,WMO}$ as given by Eqs. (4) and (5) below.

$$W_{5,PMR_T}(i, j) = \frac{|W_{5,WMO}[i][j] - (W_{5,WMO}[i][j] - W_{5,WMA}[i][j])|}{W_{5,WMO}[i][j]} \quad \text{for all } i, j \text{ states and transitions}$$

where,

- $W_{5,PMR_T}$: represents pattern matching rate value in transition level, $(0 < W_{5,PMR_T} \leq 1)$
- $i, j$: refers to indexes of states and transitions respectively, $i = 0 .. \text{total number of non zeros states}$

of each combined of unique five words, and $j = 0 .. \text{total number of non zeros transitions in the given Arabic text}$.

- $W_{5,WMO}$: refers to original watermark value in transition level.

- $W_{5,WMA}$: refers to attacked watermark value in transition level.

$$W_{5,PMR_S}(i) = \frac{\sum_{j=1}^{n-5}(W_{5,PMR_T}(i, j))}{\text{Total StatePatternCount}(i)} \quad \text{for all } i$$

(5)

where,

- $n$: is a total number of non zeros transitions of every state represented in matrix of Markov model.
- $i$: is a total number of non zeros patterns of every state of pair words represented in matrix of Markov model.

- $W_{5,PMR_S}$: refers to value of pattern matching rate in state level, $(0 < W_{5,PMR_S} \leq 100)$.

After the pattern matching rate of every state has been produced, I have to find the weight of every state stored in Markov matrix as presented in Eq. (6).

$$W_{5,Sw} = \frac{W_{5,PMR_S}(i) \cdot \text{Transitions frequency}(i)}{\text{total number of transitions}}$$

(6)

where,

- $W_{5,PMR_S}$: is the total pattern matching rate of $i^{th}$ state.
- $i$: is a number of all states in the given Arabic text.

The final $W_{5,PMR}$ of AATp and AAT are calculated by Eq. (7).

$$W_{5,PMR} = \frac{\sum_{i=1}^{n-5}W_{5,PMR_S}(i)}{N}$$

(7)

where,

- $N$: is a total number of non-zeros values in $W_{5,mm}$ matrix.

The rate of watermark distortion represents the amount of tampering attacks occurring on the contents of the attacked Arabic context, which is denoted by $W_{5,WDR}$ and calculated by Eq. (8).

$$W_{5,WDR} = 1 - W_{5,PMR} \times 100$$

(8)

The watermark detection algorithm is presented formally and executed as illustrated in Algorithm 5.
where,
- W5_SW: refers to weight value of states correctly matched.
- W5_WDR: refers to value of watermark distortion rate $(0 < W5_WDR \leq 100)$.

The results of watermark extraction and detection process are illustrated in Fig. 8.

### 4. Implementation, Simulation and Comparison

To evaluate the performance of HNLPZWA, several scenarios of implementation and simulation are performed. This section depicts an implementation and experimental environment, experiment parameters, experimental scenarios of standard datasets, discussion and analysis.

#### 4.1 Implementation Environment and Setup

The proposed technique HNLPZWA, is implemented in PHP programming language using VS Code IDE on the system having the following features. (CPU: Intel Core i7-4650U/2.3 GHz, RAM: 8.0 GB, Windows 10 – 64 bit).

#### 4.2 Simulation and Experimental Parameters

All aforementioned experimental and simulation parameters of HNLPZWA and their associated values are given in Table 1.

### 4.3 Performance Metrics

HNLPZWA performance refers to tampering detection accuracy which is evaluated using the following metrics:

- Accuracy evaluation of tampering detection (W5_PMR and W5_WDR) under very low volume (5%), low volume (10%), mid volume (20%) and high volume (50%) of all addressed attacks with all scenarios of Arabic dataset sizes.
- Desired tampering detection accuracy values close to zero.
- Watermark tampering detection accuracy comparison of HNLPZWA with ZWAFWMMM and HHZWMA and results evaluation of dataset size effect, attacks type effect, and attacks volumes effect against tampering detection accuracy.

#### 4.4 Baseline Approaches

The tampering detection accuracy of HNLPZWA is compared with HHZWMA (Hybrid of Hidden Markov model and Zero Watermarking Approach) and ZWAFWMMM (Zero-Watermarking Approach based on Fourth level order of Arabic Word Mechanism of Markov Model) [30]. Comparison is performed under all performance metrics to find which approach gives the best accuracy of tampering detection. HHZWMA approach has been proposed for content authentication of Arabic text using soft computing tools and text watermarking. Several experiments scenarios have been done under all types of attacks and their volumes using HHZWMA approach as show in Table 2 and graphically illustrated in Fig. 9.

### 4.5 Simulation and Experiment Results of HNLPZWA

In this subsection, performance evaluation of HNLPZWA approach in terms of tampering detection accuracy will be presented. The character set covers all Arabic characters,
AL-WESABI: PROPOSING HIGH-SMART APPROACH FOR CONTENT AUTHENTICATION AND TAMPERING DETECTION OF ARABIC TEXT TRANSMITTED

4.6 Comparison Results

This subsection presents the tampering detection accuracy comparison of HNLPZWA, ZWAFWMMM and HHZWMA approaches and study their effect under core affected factors are dataset size, attack types and volumes.

4.6.1 Results Study of Attack Type Effect

Figure 10 shows how the tampering detection accuracy of HNLPZWA, ZWAFWMMM and HHZWMA approaches are influenced by the type of tampering attacks. Comparison results show that HNLPZWA outperforms ZWAFWMMM and HHZWMA in terms of tampering detection accuracy with low matching values of the original and attacked Arabic contents in all scenarios of insertion, deletion and reorder attacks. This means that HNLPZWA is a more sensitive approach for any tampering even attacks with very low volume. Based on that, HNLPZWA is a strongly recommended and applicable approach for content authentication and tampering detection of Arabic text transmitted via internet.

4.6.2 Results Study of Attack Volume Effect

Figure 11 shows how the tampering detection accuracy is influenced by low, mid and high attack volumes. Figure 11 demonstrates the lowest effect of attack volumes on detecting tampering of Arabic text transmitted via internet was detected by HNLPZWA approach with minimum matching values under all scenarios of low, mid and high volumes of all attacks types.

4.6.3 Results Study of Dataset Size Effect

The comparative results as shown in Fig. 12 reflect the tampering detection accuracy of the proposed HNLPZWA approach. The results show that the lowest effects of dataset size that lead to the best tampering detection accuracy are produced by HNLPZWA, with minimum matching values under all scenarios of low, mid and high size of all datasets. This means that HNLPZWA is a very sensitive approach and strongly recommended for content authentication and tampering detection of Arabic text transmitted via internet as it is more sensitive than ZWAFWMMM and HHZWMA in tampering detection.

5. Conclusion

HNLPZWA approach is implemented in PHP programming language using VS code IDE. The experiments are performed on various standard datasets under different volumes of insertion, deletion and reorder attacks. HNLPZWA has been compared with ZWAFWMMM and HHZWMA approaches. The comparison results show that HNLPZWA outperforms both ZWAFWMMM and HHZWMA approaches in terms of accuracy of tampering detection. The results also show that HNLPZWA is applicable to all Arabic alphabetic letters, special characters, numbers and spaces.

Fig. 9 Tampering detection evaluation of HNLPZWA under all attacks with various volumes.

Fig. 10 A compression of attack type effect on tampering detection accuracy of HNLPZWA with ZWAFWMMM and HHZWMA approaches.

Fig. 11 A compression of attack volume effect on tampering detection accuracy of HNLPZWA with ZWAFWMMM and HHZWMA approaches.

Fig. 12 A compression of dataset size effect on tampering detection accuracy of HNLPZWA with ZWAFWMMM and HHZWMA approaches.
For future work, the improvement of tampering detection accuracy should be considered for all kinds of attack.

Acknowledgements

The author extends his appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through Research Groups under grant number (R. G. P. 2/55/40 /2019).

References

[1] M. Kuriyamashi, T. Fukushima, and N. Funabiki, “Robust and secure data hiding for PDF text document,” IEICE Trans. Inf. & Syst., vol.E102-D, no.1, pp.41–47, Jan. 2019.

[2] E. Elsayed and D. Fathy, “Evaluation of Quran recitation via OWL ontology based system,” The International Arab Journal of Information Technology, vol.16, no.6, pp.970–977, 2019.

[3] N.S. Kamaruddin, A. Kamsin, L.Y. Por, and H. Rahman, “A review of text watermarking: Theory, methods, and applications,” IEEE Access, vol.6, pp.8011–8018, 2018.

[4] H. Morita, J.C.N. Schuldt, T. Matsuda, G. Hanaoka, and T. Iwata, “On the security of non-interactive key exchange against related-key attacks,” IEICE Trans. Fundamentals, vol.E100-A, no.9, pp.1910–1923, Sept. 2017.

[5] A.A. Alwan, M.A. Shahid, N.N. Amir, Sjarif, M.M. Hashim, and M.S. Mohd, “A review and open issues of diverse text watermarking techniques in spatial domain,” Journal of Theoretical and Applied Information Technology, vol.96, pp.5819–5840, 2018.

[6] I. Hamamoto and M. Kawamura, “Image watermarking technique using embedder and extractor neural networks,” IEICE Trans. Inf. & Syst., vol.E102-D, no.1, pp.19–30, Jan. 2019.

[7] N.N. Hurrah, S.A. Parah, N.A. Loan, J.A. Sheikh, M. Elhoseny, and K. Muhammad, “Dual watermarking framework for privacy protection and content authentication of multimedia,” ELSEVIER Future Generation Computer Systems, vol.94, pp.654–673, 2019.

[8] A.S. Panah, Ron van S., T. Sellis, and E. Bertino, “On the Properties of Non-media Digital Watermarking: A Review of State of the Art Techniques,” IEEE Access, vol.4, pp.2670–2704, 2016.

[9] C. Qin, C. Chang, and T. Hsu, “Effective Fragile Watermarking for Image Authentication with High-quality Recovery Capability,” KSII Trans. Internet Inf. Syst., vol.7, no.11, pp.2941–2956, 2013.

[10] S.A. Parah, J.A. Sheikh, and G.M. Bhat, “StegNmark: A Joint Stego-Watermark Approach for Early Tamper Detection,” Springer International Publishing Switzerland, vol.660, pp.427–452, 2017.

[11] S. Hakak, A. Kamsin, O. Tayan, M.Y.I. Idris, and G.A. Gilkar, “Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges,” ELSEVIER Information Processing and Management, vol.56, no.2, pp.367–380, 2017.

[12] M.T. Ahvanoey, “ANITH: A novel intelligent text hiding technique,” IEEE Dataport, vol.10, 2018.

[13] S.A. Parah, J.A. Sheikh, J.A. Akhmoon, and N.A. Loan, “Electronic Health Record hiding in Images for smart city applications: A computationally efficient and reversible information hiding technique for secure communication,” ELSEVIER Future Generation Computer Systems, vol.108, pp.935–949, 2018.

[14] R.A. Aloibai and L.A. Elrefaei, “Arabic Text Watermarking: A Review,” Int. J. Artif. Intell. Appl., vol.6, no.4, pp.1–16, 2015.

[15] K. Hameed, A. Khan, M. Ahmed, and A.G. Reddy, “Towards a formally verified zero watermarking scheme for data integrity in the Internet of Things based-wireless sensor networks,” ELSEVIER Future Generation Computer Systems, vol.167, pp.1–16, 2018.

[16] R.A. Aloibai and L.A. Elrefaei, “Improved capacity Arabic text watermarking methods based on open word space,” J. King Saud Univ.: Comput. Inf. Sci., vol.30, no.2, pp.236–248, 2018.

[17] M.S. Memon, PP and A. Shah, “A Novel Text Steganography Technique to Arabic Language Using Reverse FatStStTa,” Pakistan J. Eng. Tech. Sci., vol.1, no.2, pp.106–113, 2015.

[18] Y.M. Alginahi, M.N. Kabir, and O. Tayan, “An Enhanced Kashida-Based Watermarking Approach for Increased Protection in Arabic Text-Documents Based on Frequency Recurrence of Characters,” Int. J. Comput. Electr. Eng., vol.6, no.5, pp.381–392, 2014.

[19] A.A. Shaker, F. Ridzuwan, and S. Ali Pitchay, “Text Steganography using Extensions Kashida based on the Moon and Sun Letters Concept,” Int. J. Adv. Comput. Sci. Appl., vol.8, no.8, 2017.

[20] A.M.S. Rahma, W.S. Bhaya, and D.A. Al-Nasrawi, “Text steganography based on unicone of characters in multilingual,” Int. J. Eng. Res. Appl., vol.3, no.4, pp.1153–1165, 2013.

[21] N.A.A.S. Al-Maweri, W.A.W. Adnan, A.R. Ramli, K. Samsudin, and M.S.M.A.A. Rahman, “Robust digital text watermarking algorithm based on unicone extended characters,” Indian Journal of Science and Technology, vol.9, no.48, pp.1–14, 2016.

[22] M. Bashardoost, M.S.M. Rahim, T. Saba, and A. Rehman, “Replacement attack: A new zero text watermarking attack,” 3D Res., vol.8, no.1, 2017, DOI: 10.1007/s13319-017-0118-y.

[23] Y. Liu, Y. Zhu, and G. Xin, “A zero-watermarking algorithm based on merging features of sentences for Chinese text,” J. Chin. Inst. Eng., vol.38, no.3, pp.391–398, 2015.

[24] P. Zhu, G. Xiang, W. Song, A. Li, Y. Zhang, and R. Tao, “A text zero watermarking algorithm based on Chinese phonetic alphabets,” Wuhan Univ. J. Nat. Sci., vol.21, no.4, pp.277–282, 2016.

[25] Z. Ali, M.S. Hossain, G. Muhammad, and M. Aslam, “New Zero-Watermarking Algorithm Using Hurst Exponent for Protection of Privacy in Telemedicine,” IEEE Access, vol.6, pp.7930–7940, 2018.

[26] O. Tayan, Y.M. Alginahi, and M.N. Kabir, “An adaptive zero-watermarking approach for text documents protection,” Int. J. Image Process. Techn., vol.1, no.1, pp.33–36, 2014.

[27] M.M. Ghilan, F.M. Ba-Alwi, and F.N. Al-Wesabi, “Combined Markov model and zero watermarking techniques to enhance content authentication of Arabic text documents,” International Journal of Computational Linguistics Research, vol.5, no.1, pp.26–42, 2014.

[28] F.N. Al-Wesabi, A.Z. Alsakaf, and K.U. Vasantrao, “A zero text watermarking algorithm based on the probabilistic patterns for content authentication of text documents,” International Journal of Computer Engineering & Technology, vol.4, no.1, pp.284–300, 2014.

[29] H.M. Ahmed and M.A. Khodher, “Comparison of Eight Proposed Security Methods using Linguistic Steganography Text,” Int. J. Comput. Inf. Sci., vol.12, no.2, pp.243–251, 2016.

[30] F.N. Al-Wesabi, K. Mahmood, and N. Nemri, “A Zero Watermarking Approach for Content Authentication and Tampering Detection of Arabic Text based on Fourth Level Order and Word Mechanism of Markov Model,” J. Inf. Secur. Appl., vol.52, 2020.