Robust Text Watermarking Technique for Authorship Protection of Hindi Language Documents

Nitin Namdeo Patil *er_nitinpatil@rediffmail.com*
Department of Computer Engineering
R. C. Patel Institute of Technology, Shirpur
North Maharashtra University, Jalgaon
Shirpur, 425405, India

Jayantrao Bhaurao Patil *jbpatil@hotmail.com*
Professor, Department of Computer Engineering
R. C. Patel Institute of Technology, Shirpur
North Maharashtra University, Jalgaon
Shirpur, 425405, India

Abstract

Digital text documents have become a significantly important part on the Internet. A large number of users are attracted towards this digital form of text documents. But some security threats also arise concurrently. The digital libraries offer effective ways to access educational materials, government e-documents, financial documents, social media contents and many others. However content authorship and tamper detection of all these digital text documents require special attention. Till now, considerably very few digital watermarking techniques exist for text documents. In this paper, we propose a method for effective watermarking of Hindi language text documents. Hindi stands second among all languages across the world. It has widespread availability of its digital contents of various types. In proposed technique, the watermark is logically embedded in the text using 'swar' (vowel) as a special feature of the Hindi language, supported by suitable encryption. In extraction phase the Certificate Authority (CA) plays an important role in the authorship protection process as a trusted third party. The text is decrypted and watermark is extracted to prove genuine authorship. Our technique has been tested for various types of feasible text attacks with different embedding frequency.

Keywords: Authorship, Attacks, Robustness, Swar, Hindi, Corpus.

1. INTRODUCTION

Availability of various types of text contents in digital format has bridged the gap between the authors and the intended users. These users can have easy access to these digital documents at any time. At the same time, unauthentic users threaten the use of the data illegitimately resulting in authorship conflicts. The digital text watermarking is the outstanding solution which protects the intellectual property rights and the authorship of genuine authors.

Storage of documents using digital technologies have several advantages but with it illegal use of digital data is increased. The classical techniques of data protection are ineffective towards these problems [1]. Data is stored in digital format which is easy to copy, modify or to deform it by using appropriate software. This has created a great concern on digital content security and is being studied seriously for text document [2].The information hiding is one of the most popular techniques in the field of information security which accomplish the copyright protection and status monitoring. The digital watermarking technology embeds the specific information embedded into the digital media such as image, audio, video and text files for the digital media protection, in order to achieve the purpose of identification, annotation and copyright protection. In response to these challenges, many digital watermarking schemes have been proposed in the

last decade. In these schemes, a small amount of imperceptible secret information is embedded into the digital content, which can be extracted at a later stage for copyright assertion, copy control, broadcasting, authentication and content integrity verification [3]. Digital text watermarking is a recent research trend considered as the best suitable technique that provides reasonable and robust solution to these issues related to digital text. For these reasons, it has become a very crucial area of research. The constraints required by a watermarking system are distinct. First, watermark should be imperceptible and it must be robust against attempts of replacing or removing it. The quality of original data should not be lost or degraded. We must be able to correctly extract the watermark despite the deterioration of the payload [4]. Finally, the watermark existence must be detected only by an authorized person. In general, any watermarking scheme consists of three parts such as the watermark, the embedding algorithm and extraction or detection algorithm. The embedding algorithm incorporates the watermark into the text file, whereas the extraction algorithm authenticates the text file, determining both the owner and the integrity of the text file [5].

A large amount of Hindi language content like research, literature, government resolutions, educational study material and many others are accessible through Internet in digital format. Watermarking of Arabic, Korean, English, Chinese, Turkish, Latin, and Marathi language with particular approach is available [5-15]. We focus on Hindi language which is official language of India. Also it is second in most spoken language by native speakers all over the world. There are nearly 500 million native speakers of Hindi [21]. The major challenges of Hindi text watermarking are that it is free order language and handling variations in spelling is a complex task [22]. The lack of sufficient corpus and novelty in the research field of Hindi text watermarking make it difficult to devise and implement new watermarking algorithm(s) for Hindi text document. In this paper we propose a novel method to watermark Hindi language text. We identified ‘swar’ (vowel) as a distinguished unique feature of Hindi language grammar to be used for the generation of watermark. We further experiment with available Hindi corpus to test the proposed technique against all possible types of attacks. The results are encouraging to prove that the proposed technique is capable of proving authorship of the genuine user. It also meets the crucial basic characteristics of a typical text watermarking technique.

We arrange this paper in the following manner: The section-I introduce the background of the concept and need of proposed system. In section-II, we describe the review of related work and respective approaches. The section-III presents the proposed work in detail whereas section-IV focuses on the experimental results. At last we summarize our work in conclusions, followed by the acknowledgements.

2. RELATED WORK

Gutab et al. logically used the extension Arabic character ‘Kashida’ to propose an improved method for Arabic e-text watermarking. The algorithm utilizes all the extendable characters possibly fitted in words to represent some watermark bits. It embeds bits within 'Kashida' characters in the cover text based on a secret key similar to classical cryptography. The method maintains the robustness of the original text document [5].

Yasser et al. further extended use of Kashida to propose an invisible watermarking technique based on Kashida-marks. The watermarking key is predefined where a Kashida is placed for a bit 1 and omitted for a bit 0. Kashidas are inserted in the text before a specific list of characters until the entire key is embedded. This watermark technique has two types based on the frequency-recurrence properties of the characters. The watermarking technique proved to achieve the goal of document protection and authenticity with enhanced robustness and improved perceptual similarity with the original cover-text [6].

Zaiane et al. explored the method for Korean text watermarking and developed a morpheme and syntax based scheme. In this technique, a predicate nominal is divided into a nominal and a predicate. Korean word usually consists of a content morpheme and a function morpheme.

However, predicate nominal has exceptionally two content morphemes — nominal and predicate and one function morpheme. The technique divides a predicate nominal into a nominal and a predicate. It improves security by adopting a sentence weight value which carries a watermark bit [7].

Kim et al. Suggested a method which is useful for watermarking of agglutinative languages – such as Korean and Turkish. The algorithm constructs a syntactic dependency tree of input text which chooses target syntactic constituents. The method embeds watermark bits if the watermark bit does not coincide with the movement bit of the target constituent. The syntactic constituents are converted into the syntactic tree. The modified syntactic tree, gives the marked text [8].

He lu et al. focused on watermarking of text document in Chinese language. The coding character font size is increased by 0.5 in watermarked document. In font coding, Microsoft Word has many types of fonts. Some of them are same; means their effect is not identified easily. Such fonts are used for watermarking [9].

Meral et al. developed a morpho syntax-based natural language watermarking scheme. The algorithm first transformed the text into a syntactic tree diagram where the hierarchies and the functional dependencies are made explicit. The watermarking algorithm then works on sentences in syntax tree format and avoids semantic drops. Security is provided via key-controlled randomization of morph syntactic tools and the insertion of void watermark. The algorithm uses free word property of Turkish language [10].

Zhang et al. proposed a novel text watermarking scheme for word document. The scheme embedded the secret message in the special properties of word object. The watermarking information was encrypted and divided into several groups. The encrypted information was added into watermark before embedding into the document [11].

Fu et al. successfully developed public zero knowledge watermark detection protocol. Watermarks are generated by using logistic chaotic map from the robust text features which are extracted from plain text. The watermarks are embedded into the plain text using substitution of synonyms based on the semantic adjacent words method. Correlation value is calculated for ownership detection [12].

Rizzo et al. implemented a text watermarking technique based on homoglyph characters substitution for Latin symbols. The proposed algorithm is able to efficiently embed a password based watermark in short texts without disturbing the contents. The method uses alternative Unicode symbols to ensure visual understandability and content preservation [13].

Patil et al. developed an embedding algorithm which effectively uses the unique construct of Devanagari language ‘sarvanam’ (pronoun) for generation of the required watermarks in combination with additional security phrases. The watermark has given extra level of security by encrypting it with AES algorithm. The extraction algorithm uses the sarvanam feature as a private key for ownership detection of the Devanagari text document. The authors also conducted performance analysis with suitable corpus to analyze possible types of attack [14, 15].

Mir et al. proposed a method to watermark xml document. Synonym and Acronym are used for Watermark process. Synonym or acronym databases are created or implemented at run time. The list of acronym and synonym is required at the receiver end to decode the information. The method gives the extra level of protection in case of breach of document. The xml file is well constructed into the HTML tags so that it can show the content in the browser [16].

Jalil et al. developed method for watermarking text document using zero watermarking technique. The existence of the double letters of the text is used to watermark the documents. However the structural algorithm is not applicable to all type of the text documents and the algorithm used alphabetic watermark [17].

Tyan et al. proposed new adaptive approach based on zero-watermarking for highly-sensitive documents in order to achieve content originality verification and validation without physically modifying the cover-text. The algorithm embeds the watermark in an identical duplicate of the cover-document and a private key was generated. The algorithm is applicable to all digital text document from illegal copying [18].

Xingming et al. used a new approach of noun-verb based technique for text watermarking. The technique exploits grammatical constructs like nouns and verbs in a sentence parsed with a grammar parser using semantic networks [19].

Vybornova et al. proposed a method for watermarking texts of arbitrary length using natural-language semantic structures. The key to the algorithm is to use the linguistic semantic construct of presuppositions. The sentence can be used with or without presupposition or with a different presupposition trigger. Then further used without changing relations between subjects, objects and meaning of the sentence. Distinct rules are provided to identify presupposition for each trigger. The watermark embedding is effective with longer text sentences [20].

All the above discussed techniques can be categorized as syntactic, semantic, open space and structural methods based on their way of implementation. Each of these methods has certain limitations. The syntactic methods are not uniformly applicable to all types of documents. Some of these methods are also referred as open space methods which use manipulation in inter-line, inter-word or inter-character space for targeted implementation [8, 9, 10, 11]. The semantic based methods affect the meaning of the original text. This change in meaning and presentation of respective language content is unexpected to the genuine authors [5, 16, 19]. Although the structural text watermarking approach including method like zero watermarking does not make alterations to the original text, these are difficult to implement and have limited applications [17, 18, 20].
In our proposed technique which is also a structural approach, we have attempted to overcome the above discussed limitations. We claim higher degree of embedding capacity, lower computation time and zero distortion rate. The robustness is also considerable better in our proposed work.

3. PROPOSED WORK

We propose a novel watermarking technique for embedding and extraction of data from the watermarked text documents. We use an important language construct called “swar” of Hindi. There are total eleven (11) swar (vowels) in Hindi language. We form an array of nine swar to be used for watermark generation in proposed algorithm. Our algorithm works effectively with swar to make private key which can be embedded with the original text.

In embedding algorithm, the original Hindi text (HT) is provided with unique author-id and Timestamp. The algorithm uses total count of swar in input text to calculate the digital root of its all occurrences to generate the private key. The raw key comprising of calculated swar, author-is and time-stamp is formed. Then ciphering is applied for the purpose of encryption of the raw key before embedding it into the input Hindi text.

3.1 Embedding Algorithm

Our embedding algorithm is given below:

1. Input Hindi Text File HT.
2. Input Author id Generated by any combination of capital (A-Z) and small (a-z) English alphabet.
3. Input Time-Stamp (user choice).
4. Create and define an array of Hindi Swar AS.
5. For the content of Hindi Text File do

I. Count the Occurrence of Hindi Swar from Hindi Swar text File
II. Find the Digital Root of Total Swar Count
III. For AS [i:n]

i. Digi_Root = $\sum_{i=1}^{9} \text{Digi\_Root} + i$
ii. until (i ≥ 9)

6. Select AS [j]
   until j ≤ Digi_Root
7. Raw_Key= AS + Author_id+ Time_Stamp
8. Private Key = Cipher(Raw_key)
9. Generate Watermark file by embedding cipher key into HT.
10. Stop.

In extraction phase, the authentication process is conducted. To extract the watermarked contents, all the necessary security conditions should meet the requirements. Here the Certificate Authority (CA) plays an important role. CA is a trusted third party which ensures the access of the watermarked contents only by the genuine authors. Along with watermarked file, the respective watermark and timestamp is maintained by CA in its database. Also CA assigns unique key for each user and keeps it with the relative record of the author. In situation of authorship conflict arrival, CA is the responsible authority who confirms the genuine author of the document by verifying the concerned corresponding record of genuine user. Table 1 shows sample entries of the registered users.

Author Name	Watermark in Protected File	Date &Time Watermarking	Unique key for each author	Watermarked File
Nitin	MOPerRGldfKnVesWLQ92df81xmw96pYICS56IBg057	1 Aug 2016 01:32:50:000	3e0n8f19	Mydoc1.pdf
Saijyot	TUB24NPgdaKLKSDhaq78BBqP06CSghe29QpTU8La3	3 Aug 2016 03:09:44:000	A5I9LPR2	NNP9.pdf

TABLE 1: Author Records Maintained by CA.

3.2 Extraction Algorithm

The extraction algorithm is as described below:

1. Input text file HT, Auth_ID, Time_Stamp.
2. Read text file.
3. Retrieve encrypted key.
4. Apply Deciphering to get extracted key.
5. For AS[i : n]

I. Find the Digi_Root of Total Swar Count
II. For AS [i:n]

I. Digi_Root = $\sum_{i=1}^{9}$ Digi_Root + i

II.until (i≥ 9)

6. Select AS [j]
   until j ≤ Digi_Root

7. Match_Watermark (extracted key, Private Key)

8. If match found.

9. “Genuine Author”

10. Else
    “Invalid Access”

11. Stop.

4. EXPERIMENTAL RESULTS AND DISCUSSION

An implementation of digital text watermarking faces a big challenge of availability of suitable corpus to validate the results and to check the performance of the proposed technique. For testing our proposed technique, we experimented with the corpus of ‘Indian Institute of Technology, Bombay’ [23] and the other provided by ‘Technology Development for Indian Languages’ (TDIL) [24]. We preprocessed the corpus to correct some typo errors. We used ‘krutidev’ as common font for entire selected Hindi text contents. We formed four different text categories of available corpus like: Category 1 which contains 1 to 250 words, Category 2 having 251 to 750 words, the Category 3 with words in between 751 to 1200 and the Category 4 which can have minimum 1201 to maximum 2000 words. We considered 10 samples for all the categories to test them against particular insertion, deletion and combined watermark attacks. We also varied the capacity of percentage attack like 10, 20 and 30 % for each text category.

The digital text may undergo attacks like insertion, deletion and combined attack. Insertion attack inserts additional contents in the original text document. The deletion attack is the deletion of some or more contents from original text. The combined attack which is combination of insertion and deletion attack, may insert and delete contents to and from the original author’s document. If insertion or deletion attack are focused at particular location like beginning, middle or end of the document then it is localized text attack. The attack at any random location on the original text is called as dispersed attack [17].

As per our observations, the retrieving accuracy varies according to the volume of text in all four categories. More the volume of text category, better is the retrieving accuracy for it. Also the type of attack affects the retrieving accuracy. Also the nature of attack like localized and dispersed can affect the retrieving accuracy. The observations for deletion attack in Table 2 indicate the destruction of text reasonably which is contrary to observations noted in case of insertion attack. When the text experiences insertion and deletion attack as a combined attack, it can considerably damage the text.

Attack Type		Localized Attack			Dispersed Attack		
Attack (%)		10	20	30	10	20	30
Insertion Attack	Category1	95.9	98.3	97.3	96.7	97.4	97.1
	Category2	94.9	96.8	97.8	94.1	97.9	97.8
	Category3	95.6	97.6	98.3	95.3	97	97.9
	Category4	95.6	97.2	97.5	97.7	96.1	97.6
Deletion Attack	Category1	96.5	98.2	97.7	96.1	96.6	97
	Category2	96.7	98.4	98	96.4	96.8	97.1
	Category3	94.9	97.5	97.8	94.9	96.1	97.3
	Category4	96.8	96.5	96.9	95.4	97.9	98
Combined Attack	Category1	98.3	96.8	96.4	96.8	94.8	96.1
	Category2	97.1	95	96.6	98.5	95.6	96.8
	Category3	97.1	96.2	96	97.8	96.3	95.5
	Category4	96.8	96.1	95.5	97	95.8	96.6

TABLE 2: Watermark Retrieving Accuracy for Different Types of Attacks.

Our proposed technique also meets the basic requirements of text watermarking specifically: robustness, imperceptibility, security and capacity. We now describe how these characteristics are achieved. In the attack analysis, the watermarked Hindi text is forced to undergo typical attacks. The attacks are of different capacity and types. Our proposed system confirms the robustness by retrieving greater accuracy to protect watermark from the damage. In our proposed technique, the watermarked Hindi text proves to be useful and valuable even after the enforced attacks. The fundamental property of text watermarking is imperceptibility. In our implementation, the watermark embedding is done invisibly. Also the text file before and after the watermarking appears perpetually indistinguishable. The security of the technique keeps the watermark undisclosed from any illegal access. The existence, size and the payload of the watermark remains covert to satisfy the security of our technique. The significantly upper limit of the watermark embedding proves the acceptable capacity of the technique. Hence the proposed technique consists of the basic characteristics of text watermarking.

We investigated attack analysis for 10, 20 and 30 % of total volume of text samples with all type of attacks for four text categories. The figures 1-6 represent the graphs of retrieved watermark accuracy against respective text categories. Here the graphical illustration verifies that the size of sample text under consideration and the type of attack has significant influence on the retrieving watermark accuracy.

The simulation results validate the watermark retrieving capacity in case of possible attacks. Any intentional or unintentional alterations to the Hindi text in form of insertion, deletion or reordering attacks can be certainly detected. Further countermeasures are enforced to retrieve the watermarks in attacked text which results in preserving the basic text watermarking properties. Also simulation executes uniformly for all size of text samples for experimental results.

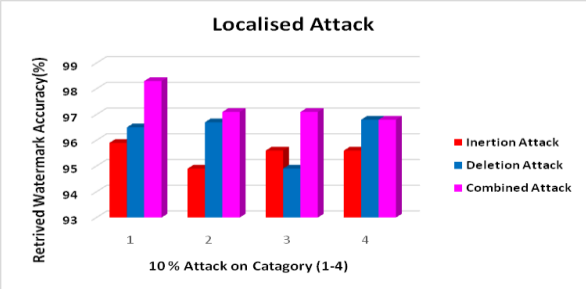

FIGURE 1: Retrieved Accuracy for 10 % Localised Attack.

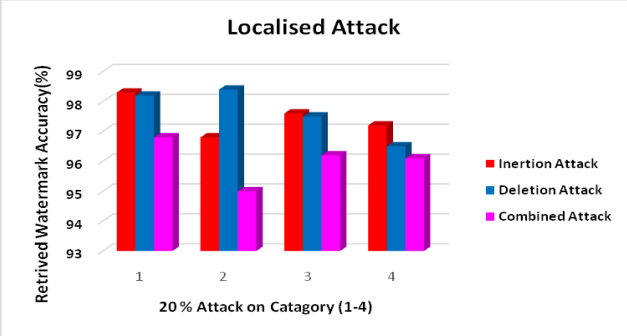

FIGURE 2: Retrieved Accuracy for 20 % Localised Attack.

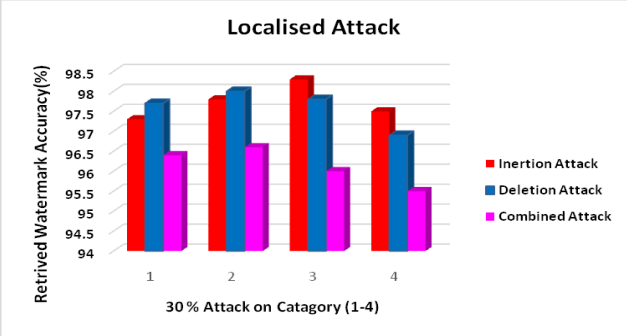

FIGURE 3: Retrieved Accuracy for 30 % Localised Attack.

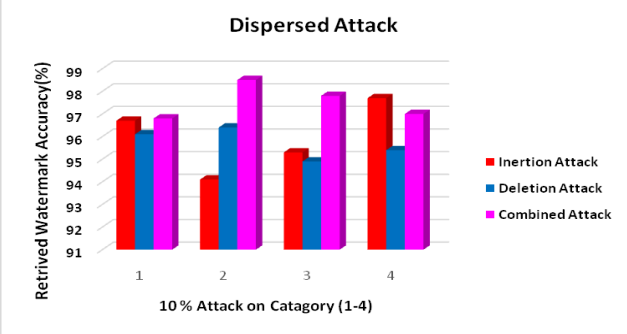

FIGURE 4: Retrieved Accuracy for 10 % Dispersed Attack.

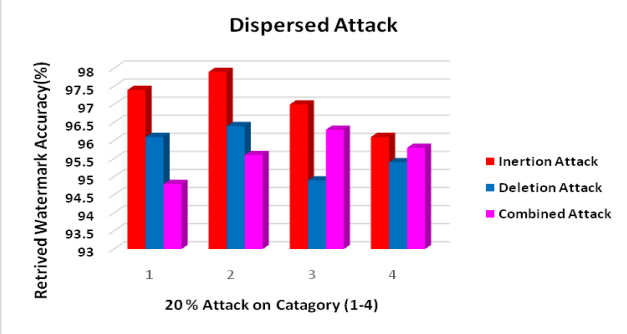

FIGURE 5: Retrieved Accuracy for 20 % Dispersed Attack.

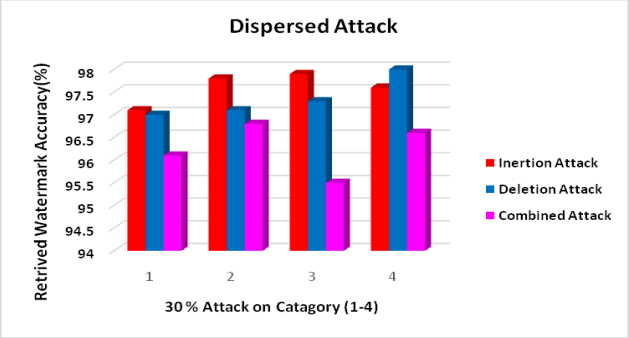

FIGURE 6: Retrieved Accuracy for 30 % Dispersed Attack.

To confirm our obtained results, we need to compare them with some other linguistic text watermarking techniques. Due to varying nature of different natural languages, we cannot have direct comparison of the related watermarking techniques. The language corpus, language specific features and the watermarking implementation approach have significant impact on the outcomes. In general, we compare our current results with English language text watermarking proposed by Jalil et al. [17] and also with our linguistic implementation i.e. Marathi language text watermarking [14, 15]. The first comparison confirms our Hindi language text watermarking to be better than the English language text watermarking with respect to watermark retrieving accuracy. Another comparison with Marathi language text watermarking is considerably similar on the basis of watermark retrieving capacity. In both the results, for the localized insertion attack, watermark retrieving accuracy is good for all types of text samples. For text samples with greater volumes, the dispersed insertion attack has better retrieving capacity compared to those for smaller text samples. Also the watermark recovery for localized deletion, smaller text volumes undergo more vulnerability than the larger text volumes. For dispersed deletion, larger the text volumes, better is the watermark retrieving accuracy.

As no such technique exists till now for Hindi language text watermarking, our proposed work is a novel approach for Hindi language digital text documents. This approach demonstrates improvements over the limitations like lower robustness, deficient imperceptibility and inadequate security of existing digital text watermarking techniques. Subsequently, it exhibits comparatively greater degree of all the above discussed watermarking characteristics. This leads to effective applications of Hindi language digital text watermarking such as authorship protection, document authentication, tamper proofing and fingerprinting.

5. CONCLUSIONS AND FUTURE WORK

Digital content on Internet has become popular because of the easy access. It has encouraged most of the authors to publish their data in digital formats. On the other hand security threats to these data have become a crucial issue. We have developed a robust watermarking technique for a popular language Hindi. We have attempted to incorporate all fundamental characteristics of a text watermarking technique. Also we conducted text attack analysis in support of viability and performance of the proposed technique. The results are specifically limited and applicable to Hindi language based on the properties of the language construct. Successful implementation of the technique may lead to discover generalized approach for linguistic text watermarking.

Availability of adequate corpus of any natural language is a challenging task for natural language text watermarking. The unavailability of benchmarking tools is another critical problem. Also, there is need of standard guidelines to evaluate any implemented text watermarking technique. Thus the adequate corpus, quality benchmarking tools and evaluation process are the future research challenges in the context of digital text watermarking.

6. ACKNOWLEDGEMENTS

This research work is supported under the scheme 'Research Grant Assistance for Science and Technology Projects' (RGS &TC) by Government of Maharashtra (India) and North Maharashtra University, Jalgaon. We are thankful them for providing funding and resources to undertake this work.

We express our sincere gratitude to 'Center for Indian Language Technology' (CFILT) set up by 'Indian Institute of Technology, Bombay' (IITB) [23] and 'Technology Development for Indian Languages' (TDIL) initiated by the Ministry of Electronics & Information Technology (Meity), Govt. of India for providing Hindi language corpus [24].

7. REFERENCES

[1] Stefan Katzenbeisser Fabien A. P. Petitcolas, Information Hiding Techniques for Steganography and Digital Watermarking. Security Technologies for the World Wide Web, Artech House Computer Security Series, 2000.

[2] Kim, Y., Moon, K. and Oh, I. "A text watermarking algorithm based on word classification and inter-wordspace statistics". Proceedings Seventh International Conference on Document Analysis and Recognition, 2003, pp.775 -779.

[3] F. A. P. Petitcolas, R. Anderson, M. G. Kuhn. "Information hiding - A survey". Proceeding of the IEEE, 1999, pp. 1062- 1077.

[4] F. Hartung, M. Kutter. "Multimedia watermarking techniques". Proceeding of the IEEE, 1999, pp. 1079-1107.

[5] Adnan Gutub, Fahd Al-Haidari, Khalid Al-Kahsah, JameelHamodi, "e-Text Watermarking: Utilizing 'Kashida' Extensions in Arabic Language Electronic Writing", Journal of Emerging Technologies in Web Intelligence (JETWI), pp. 48-55, 2010.

[6] Alginahi, Yasser M., Muhammad N. Kabir, Omar Tayan. "An Enhanced Kashida-Based Watermarking Approach for Increased Protection in Arabic Text-Documents Based on Frequency Recurrence of Characters." International Journal of Computer and Electrical Engineering, no. 5 pp. 381-392, 2014.

[7] Kim Mi-Young, Osmar R. Zaiane, and Randy Goebel. "Natural language watermarking based on syntactic displacement and morphological division." Computer Software and Applications Conference Workshops (COMPSACW), 2010 IEEE 34th Annual. IEEE, 2010.

[8] Kim, Mi-Young. "Text watermarking by syntactic analysis." Proceedings. of WSEAS International Conferences on Computers. 2008, pp. 904-909.

[9] H. Lu, F. DingYi, G. XiaoLin, C. XiaoJiang, X. XinBai and L. Jin' An, "A New Chinese Text Digital Watermarking for Copyright Protecting Word Document." Communications and Mobile Computing, 2009. CMC '09. WRI International Conference on, Yunnan, 2009, pp. 435-439.

[10] Hasan Mesut Meral, Bulent Sankur, A. Sumru Ozsoy, TungaGungor, EmreSevinç. "Natural language watermarking via morphosyntactic alterations."Comput. Speech Lang. 23, 2009, pp.107-125.

[11] Zhang, Yingli, Huaiqing Qin, Tao Kong. "A novel robust text watermarking for word document." Proceeding of Image and Signal Processing (CISP), 2010 3rd International Congress on, vol. 1 IEEE, 2010, pp. 38-42.

[12] Fu, Zhangjie, Xingming Sun, JiangangShu, Lu Zhou. “Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption.” Advanced Science and Technology Letters 48, 2014.

[13] Rizzo, Stefano Giovanni, FlavioBertini, DaniloMontesi. “Content-preserving Text Watermarking through Unicode Homoglyph Substitution.” Proceedings of the 20th International Database Engineering & Applications Symposium, ACM, 2016, pp. 97-104.

[14] Patil Nitin N, J B Patil. “Implementation of a Novel Watermarking Technique for Devanagari Text.” International Journal of Information and Electronics Engineering 5(5), 2015.

[15] Nitin Namdeo Patil, Jayantrao Bhaurao Patil. “Performance Analysis of a Novel Text Watermarking Technique for Devanagari Text” In book Proceeding of International Conference on Intelligent Communication, Control and Devices, volume 479, editors Rajesh Singh, Sushabhan Choudhury, Springer Singapore, 2016, pp. 325-333.

[16] Mir, Nighat, and Sayed Afaq Hussain. "Web page watermarking: XML files using synonyms and acronyms." World Academy of Science, Engineering and Technology 73 (2011).

[17] Jalil, Zunera, Anwar M. Mirza, Tahir Iqbal. "A zero-watermarking algorithm for text documents based on structural components." In Information and Emerging Technologies (ICIET), 2010 International Conference on IEEE, 2010, pp. 1-5.

[18] Tayan, Omar, Yasser M. Alginahi, and Muhammed N. Kabir. "An adaptive zero-watermarking approach for authentication and protection of sensitive text documents." In Proceedings of the International Conference on Advances in Computer and Information Technology (ACIT’13). 2013.

[19] Sun, Xingming, and Alex Jessey Asiimwe. "Noun-verb based technique of text watermarking using recursive decent semantic net parsers." In International Conference on Natural Computation, pp. 968-971. Springer Berlin Heidelberg, 2005.

[20] Vybornova, Olga, and Benoit Macq. "A method of text watermarking using presuppositions." In Electronic Imaging 2007, pp. 65051R-65051R. International Society for Optics and Photonics, 2007.

[21] “Top 30 Languages by Number of Native Speakers.” Internet Source: http://www.vistawide.com/languages/top_30_languages.html[22 Aug. 2016]

[22] Sharma, Richa, Shweta Nigam, Rekha Jain. "Opinion mining in Hindi language: a survey." International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.2, March 2014, pp. 41-47.

[23] “Center for indian language technology.” Internet Source: http://www.cfilt.iitb.ac.in/Downloads.html. [22 Aug. 2016]

[24] “Technology Development for Indian Languages Programme” Internet Source: http://tdil.mit.gov.in/ [22 Aug. 2016]