EVALUATION OF STEGANOGRAPHY FOR URDU /ARABIC TEXT.

Jibran Ahmed Memon, Kamran Khowaja, Hameedullah Kazi
Department of Computer Science.
Asstt.Prof., Department of Computer Science, Isra University.
Asstt.Prof., Department of Computer Science, Isra University.
E-mail: jibmemon@hotmail.com, kamran_khowaja@hotmail.com, hammadullah.kazi@gmail.com

ABSTRACT

Establishing hidden communication and conveying information secretly has been of interest since long past ago. One of the methods introduced for establishing hidden communication is steganography. Methods of steganography have been mostly applied on images, audios, videos and text while the major characteristic of these methods are to change in the structure and features so as not to be identifiable by human users. Text documents have been widely used since very long time ago. Therefore, different methods of hiding information in texts (text steganography) are witnessed from the past to the present. In this paper a new approach for steganography in Arabic and Urdu texts is introduced. Considering the existence of harakaat/Araabs i.e. Fatha, Kasra, and Damma in Arabic and Urdu phrases, in this approach, by using Reverse Fatha, information is hidden in the texts. This approach can be categorized under feature coding methods. This method can be used for Arabic/Urdu Watermarking.
Keywords – Urdu Language, Arabic Language, Steganography, Hidden Communication

1. INTRODUCTION

Twenty first century is the era of Internet technologies. Computer techniques are progressing remarkably and the amount of electronically transmitted information is increasing everyday. Exchange of information, over far distances, is now an everyday activity.

The prevalent language for communication on the Internet is English. This may be a result of the Internet's origins, as well as English's role as the lingua franca (A lingua franca is any language widely used beyond the population of its native speakers). The Internet's technologies have developed enough in recent years, especially in the use of Unicode that good facilities are available for development and communication in most widely used languages. Communication on Internet uses different markup languages like HTML, DHTML, WML and XML etc along with plain text like English, Chinese, Japanese, German, Arabic and Urdu etc.

Today on Internet a hidden exchange of information has been an important issue since old times and the issue of information security has gained special significance. For this purpose various methods including cryptography, steganography, coding and so on have been used. In cryptography, the information is encrypted with a key and the person who has the key can decrypt and read the information which means nobody else has access to that information.

In steganography is one of the methods which have attracted more attention during the recent years. In implementing steganography, the main objective is to hide the information under cover media so that the outsiders may not discover the information contained in the said frame. This is the major distinction between steganography and other methods of hidden exchange of information. For example, in cryptography method, people become aware of the existence of information by observing coded information although they will be unable to comprehend the information. However, in steganography, nobody will understand the existence of information in the resources [1]. Most of steganography works have been carried out on pictures [2, 3], video clips [4, 5], music and sounds [6]. Text steganography is the most difficult kind of steganography [7]; this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [8].

Today, the computer systems have facilitated hiding information in texts. The range of using hiding information in text has also developed. From among the most important of these technologies, one can name of hiding information in electronic texts and documents. The use of hiding information in text for web pages is another example. Different methods are

used for hiding information in text which will be dealt with in section 2.

The present paper offers a new method for hiding information in Arabic and Urdu texts. Due to differences between languages, no single method can be used for hiding information in texts of different languages. This will be discussed in section 3.

2. PREVIOUS WORKS

Following is the list of works has been done on hiding information or text steganography carried out.

2.1. TEXT STEGANOGRAPHY IN MARKUP LANGUAGES.

In this method, one of the features of markup languages is used to hide information [10]. For instance feature of HTML document is their tags case insensitivity. For example, the tag
 can be also used as
 and
. As a result one can do text steganography in HTML documents by changing the small or large case of letters in document tags.

In some case the positions of tags are also used for text steganography. For example <B><U> </B></U> or like this <U><B> </U></B>. To extract information in first method of text steganography by comparing these words with words in normal case and in second case by comparing the tags positions. Then by using appropriate function in both hidden information is extracted. However these methods are not for all markup languages like in the WML, all tags are case sensitive and as a result first text steganography method cannot be employed on it but second text steganography method can be employed.

2.2. TEXT STEGANOGRAPHY IN SPECIFIC CHARACTERS IN WORDS.

In this method, some specific characters from certain words are selected [11]. For example the first words of each paragraph are selected in a manner that by placing the first characters of the selected words side by side, as a result it forms secret or hidden information is extracted.

This method requires strong mental power along with it takes a lot of time and it also requires special text because not all type of texts can be used in this method.

2.3. LINE SHIFTING METHOD.

In this method, the lines of the text are vertically shifted to some degrees [12,13]. For example, some lines are shifted 1/300 inch up or down in the text and information are hidden by creating a hidden unique shape of the text. This method is feasible for printed texts.

However, in this method, the distances can be observed by using special instruments of distance assessment and necessary changes can be introduced to destroy the hidden information. Also if the text is retyped or if character recognition programs (OCR) are used, the hidden information would get destroyed.

2.4. WORD SHIFTING.

In this method, by shifting words horizontally and by changing distance between words, information are hidden in the text [12, 14]. This method is acceptable for texts where the distance between words is varying. This method can be identified less, because change of distance between words to fill a line is quite common. But if somebody was aware of the algorithm of distances, he can compare the present text with the algorithm and extract the hidden information by using the difference. The text image can be also closely studied to identify the changed distances. Although this method is very time consuming, there is a high probability of finding information hidden in the text. The same as in the method described under 2-3, retyping of the text or using OCR programs destroys the hidden information.

2.5. SYNTACTIC METHODS.

By placing some punctuation signs such as full stop (.) and comma (,) in proper places, one can hide information in a text file [11].

This method requires identifying proper places for putting punctuation signs. The amount of information to hide in this method is trivial.

2.6. SEMANTIC METHODS.

In this method, we use the synonym of words for certain words thereby hiding information in the text [10, 15]. A major advantage of this method is the protection of information in case of retyping or using OCR programs (contrary to methods listed under 2-3 and 2-4).

However, this method may alter the meaning of the text.

2.7. FEATURE CODING.

In this method, some of the features of the text are altered [16]. For example, the end part of some characters such as h, d, b or so on, are elongated or hortened a little thereby hiding information in the text. In this method, a large volume of information can be hidden in the text without making the reader aware of the existence of such information in the text.

By placing characters in a fixed shape, the information is lost. Retyping the text or using OCR program (as in methods 2-3 and 2-4) destroys the hidden information.

2.9. ABBREVIATION.

Another method for hiding information is the use of abbreviations.

In this method, very little information can be hidden in the text [8]. For example, only a few bits can be hidden in a file of several kilobytes.

2.10. OPEN SPACES.

In this method, hiding information is done through adding extra white-spaces in the text [8, 17]. These whitespaces can be placed at the end of each line, at the end of each paragraph or between the words. This method can be implemented on any arbitrary text and does not raise attention of the reader.

However, the volume of information hidden under this method is very little. Also, some text editor programs automatically delete extra white-spaces and thus destroy the hidden information.

2.11. VERTICAL DISPLACEMENT OF THE POINTS IN PERSIAN LETTERS.

In this method, text steganography is applied on Persian text [18, 19]. One of the characteristics of Persian language is abundance of points in its letter. In Persian 18 letters out of 32 letters have points. From these 18, 3 letters have 2 points each, 5 letters have 3 points each and the other 10 letters have 1 point each. These 1 point letters are used to hide the information by shifting position of point a little bit vertically high with respect to the standard point position in the text.

3. PROPOSED TECHNIQUE

One of the characteristics of Arabic language is the use of Araabs i.e. (Fatah, Kasra, and Damma). Where Fatha is slash like symbol and is written over the character, whereas Kasra is also a slash like symbol but is used below the character and Damma is number nine like symbol which is also placed over the character. Although these Araabs are not very commonly used now a days but still are the part of languages like Urdu. These Araabs are used for the correct pronunciation of words because in some languages a single word has multiple meanings depending upon the pronunciation which is handled by these Araabs. These Araabs are not only acceptable in Urdu but also rarely used.

These araabs are applicable on every single character of the Arabic language. In general these araabs are noteworthy in Arabic or Urdu text. In this paper, these same characteristics of Arabic or Urdu languages are used for steganography.

For this purpose, only fatha is used in reverse order to represent the secret character in the text. For example this fatha in reverse order is named as Reverse fatha and in the text where ever this reverse fatha is used it represents a secret character below it as shown in figure 1. In figure 1 a word is written in which regular fatha is used in from the left hand side and the same word having regular fatha and reverse fatha, where this reverse fatha is on the character “RAA” which is the secret character. Therefore in an urdu text this reverse fatha is not easily visible if efficiently used.

Figure 1. Use of regular and reverse fatha.

For applying this method on any Urdu text first of all make a secret message which is probably of one line hardly 5 words containing 10-15 characters. Now let us select an article of 4 to 5 paragraph or as the number of paragraphs is increased the security is also increased. Then we put araabs on the complete text or take an article having araabs already. After this let read the secret message character by character and match it in the article, and we have to put reverse fatha where the secret characters exist sequentially try to use the reverse fatha not on the same line but on different lines. Now the article to the reader is ready to be sent in the form of letter.

When the receiver gets the letter the reader will have to read the article carefully and tries to find the reverse fatha in the letter and extract the characters exact below the reverse fatha and collect them. In the end when letter is completely read then concatenate the collected characters. Finally when the concatenated

characters are read it represents the meaningful message.

This method can be categorized under feature coding method. This steganography method can be used in Arabic language as well as in those which use araabs.

4. ADVANTAGES AND DISADVANTAGES

4.1. ADVANTAGES

1. By this method, a large volume of information can be hidden in text, because a large number of Araabs in Arabic and Urdu are used on large number of characteristics i.e. the Reverse Fatah is not clearly visible in the text.
2. Due to the lack of a strong OCR program for Urdu and Arabic languages, the printed text cannot be easily converted into a simple text thus destroying the hidden information is difficult.
3. The text containing hidden phrases is not specific to computer and the hidden information can also be extracted from printed text. In order to recover the information in case of printed text, the text should be scanned and then subjected to the relevant program.
4. The hidden text is resistant to enlargement or downsize and these changes do not destroy the hidden information.

4.2. DISADVANTAGES

1. The information is lost in case of retyping.
2. The output text has a fixed frame due to the use of only one font.
3. Due to the lack of good OCR program for Persian and Arabic languages, using this method in texts that are printed and then scanned is difficult.

5. SIMULATION AND RESULTS

There are many ways to evaluate the technique in different environment.

Following is the method used by this paper according to his environment in which most of the people are computer literate and some of them are not. Those who are computer literate then the message is given to them on intranet and those who usually don't use computers then a hard copy is given to them. In this method the Reverse Fatha technique is also compared with other previous technique that is Dot Displacement Technique. Both these techniques are implemented on the same text which is mentioned below later. The texts with the both techniques implemented are now distributed in the environment of the writer of this paper. As it is obvious that most of the people are not familiar with text steganography and readers usually not focus on the text in that sense. According to this type of environment it is obvious that negligible amount of people will notice or focus something suspicious in the text. In short writer of this paper along with his text containing secret message the writer of this paper have asked some questions which are giving some clues to the readers and trying to generate a doubtful environment so that they can see the technique or a secret message after in the text. Following is the figure of few lines from the text used in the evaluation process of Reverse Fatha Technique.

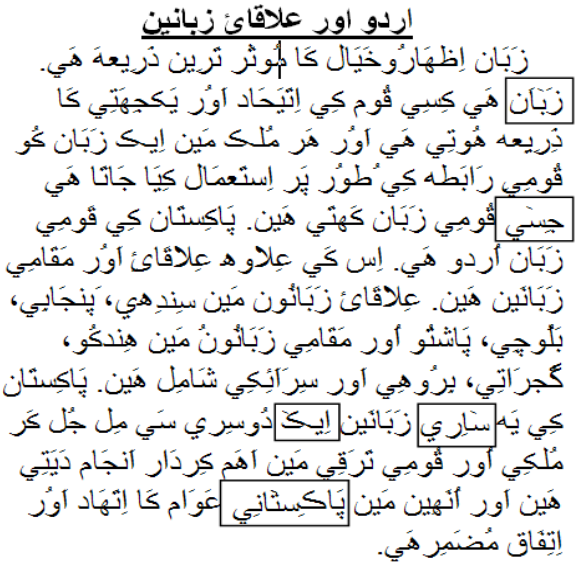
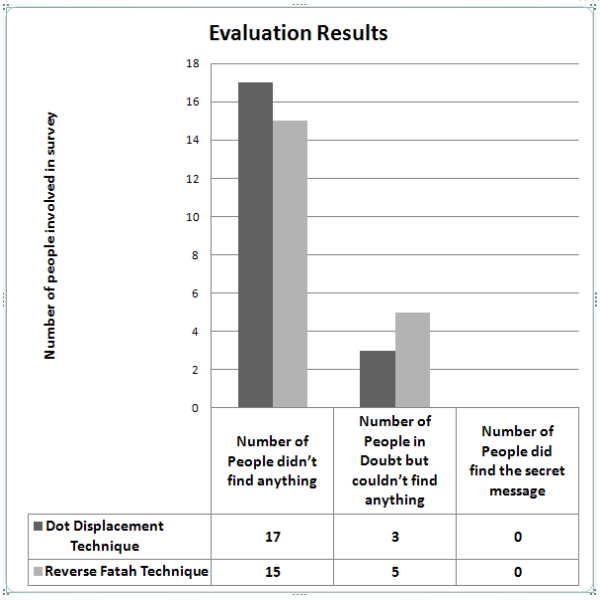

اردو اور علاقائ زبانين
زَبَان اِظهَارُوخَيَال كَا هُوثر ترين ذَرِيعهَ هَي.
زَبَان هَي كِسِي قُوم كِي اِتّيحَاد اوُر يَكجِهتِي كَا
ذَرِيعه هُوتِي هَي اوُر هَر مُلک مَين اِيک زَبَان کُو
قُومِي رَابَطه كِي طُورُ پَر اِستعمَال كِيَا جَاتَا هَي
جِسّي قُومِي زَبَان كَهتِي هَين. پَاكِستَان كِي قُومِي
زَبَان اُردو هَي. اِس كَي عِلاوه عِلاقائ اوُر مَقَامِي
زَبَانين هَين. عِلاقَائ زَبَانُون مَين سِندِهي، پَنجَابِي،
بَلُوچِي، پَاشتُو اُور مَقَامِي زَبَانُونُ مَين هِندكُو،
گجرَاتِي، برُوهِي اور سِرَائِكِي شَامِل هَين. پَاكِستَان
كِي يَه سَارِي زَبَانين اِيک دُوسرِي سَي مِل جُل كَر
مُلكِي اُور قُومِي تَرَقِي مَين اهَم كِردَار انجَام دَيتِي
هَين اور اُنهين مَين پَاكِسِتَانِي عَوَام كَا اِتّهَاد اوُر
اِتِفَاق مُضمَر هَي.

Figure 2. Article using reverse fatha.

The results achieved from the survey are shown in the following table 1 the evaluation results are shown.

Table 1 Evaluation results



6. CONCLUSION

In this paper a new approach for steganography of information in Arabic and Urdu texts is introduced. This method is based on the existence of Araabs in majority of letters of Arabic alphabets. On this basis, information was hidden in text by placing the Reverse Fatah. This method can be used in hidden exchange of information through text documents and text watermarking.

In addition to establishing secret communication, this method can be used for preventing illegal duplication and distribution of texts especially electronic texts [20, 21].

In addition to use this method for electronic texts, it can be applied on hard copy documents. To this end, they print the document after hiding data in it. For extracting data from the hard copy document, they scan it and unhide the embedded data by computer.

Considering the similarity of Persian script (official language of Iran) and other languages of Pakistan with Urdu and Arabic, this method can be used in other languages of Asian countries as well.

In addition to reverse fatha, the araabs or harakaat can be used with different sizes or width as well and thus two bits of information can be hidden in each letter. By combining the above method with other methods such as line shifting and word shifting, the volume of hidden information can be increased. By employing a font editing software, the program can be enabled dynamically to produce necessary fonts for hiding information so that the output form of the text is not homogenous and conform to the input form of the text.

REFERENCES

[1] J.C. Judge, "Steganography: Past, Present, Future", *SANS white paper*, November 30, 2001, http://www.sans.org/rr/papers/index.php?id=552, last visited: 1 May 2006.

[2] R. Chandramouli, and N. Memon, "Analysis of LSB based image steganography techniques", *Proceedings of theInternational Conference on Image Processing*, vol. 3, 7-10 Oct. 2001, pp. 1019 - 1022.

[3] M. Shirali Shahreza, "An Improved Method for Steganography on Mobile Phone", *WSEAS Transactions on Systems*, vol. 4, Issue 7, July 2005, pp. 955-957.

[4] G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", *Signal Processing: Image Communication*, vol. 18, Issue 4, 2003, pp. 263-282.

[5] G. Doërr and J.L. Dugelay, "Security Pitfalls of Frameby-Frame Approaches to Video Watermarking", *IEEE Transactions on Signal Processing*, Supplement on Secure Media, vol. 52, Issue 10, 2004, pp. 2955-2964.

[6] K. Gopalan, "Audio steganography using bit modification", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP '03)*, vol. 2, 6-10 April 2003, pp. 421-424.

[7] J.T. Brassil, S. Low, N.F. Maxemchuk, and L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", *IEEE Journal on Selected Areas in Communications*, vol. 13, Issue. 8, October 1995, pp. 1495-1504.

[8] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM Systems Journal*, vol. 35,Issues 3&4, 1996, pp. 313-336.

[9] N. F. Maxemchuk and S. Low, "Marking Text Documents", *Proceedings of the IEEE International Conference on Image Processing*, Santa Barbara, CA, USA,Oct. 26-29, 1997, pp. 13-16.

[10] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report 2004-13.

[11] T. Moerland, "Steganography and Steganalysis", May 15,2003,www.liacs.nl/home/tmoerlan/privtech.pdf, last visited: 1 May 2006.

[12] S.H. Low, N.F. Maxemchuk, J.T. Brassil, and L. O'Gorman, "Document marking and identification using both line and word shifting", *Proceedings of the FourteenthAnnual Joint Conference of the IEEE Computer andCommunications Societies (INFOCOM '95)*, 2-6 April 1995,vol.2, pp. 853 - 860.

[13] A.M. Alattar and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing ", *Proceedings of SPIE -- Volume5306, Security, Steganography, and Watermarking ofMultimedia Contents VI*, June 2004, pp. 685-695.

[14] Y. Kim, K. Moon, and I. Oh, "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics", *Proceedings of the Seventh InternationalConference on Document Analysis and Recognition(ICDAR'03)*, 2003, pp. 775–779

[15] M. Niimi, S. Minewaki, H. Noda, and E.Kawaguchi, "A Framework of Text-based Steganography Using SD-Form Semantics Model", *Pacific Rim Workshop on Digital Steganography 2003*, Kyushu Institute of Technology, Kitakyushu, Japan, July 3-4, 2003.

[16] K. Rabah, "Steganography-The Art of Hiding Data", *Information Technology Journal*, vol. 3, Issue 3, pp.245-269, 2004.

[17] D. Huang, and H. Yan, "Interword Distance Changes Represented by Sine Waves for Watermarking Text Images", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 12, December 2001, pp. 1237-1245

[18] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", Proceedings of 5th IEEE/ACIS international Conference on Computer and Information Science and 1st IEEE/ACIS, June 2006.

[19] M. H. Shirali-Shahreza, and S. Shirali-Shahreza, "A Robust Page Segmentation Method for Persian/Arabic Document", *WSEAS Transactions on Computers*, vol. 4,Issue 11, Nov. 2005, pp. 1692-1698.

[20] J.T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents", *Proceedings of the IEEE*, vol. 87, Issue. 7, July 1999, pp. 1181-1196.

[21] J. T. Brassil, S . Low, N. F. Maxemchuk, and L. O'Gorman, "Marking Text Features of Document Images to Deter Illicit Dissemination", *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 1994, vol. 2, 9-13 Oct. 1994, pp. 315-319.