# A NOVEL FEATURE EXTRACTION APPROACH: CAPACITY BASED ZERO-TEXT STEGANOGRAPHY

*Saeeda Kouser*[1], *Aihab Khan*[2]

[1] Department of CS & IT, Mirpur University of Science and Technology, MUST
[2] Department of Computing and Technology, Iqra University, Islamabad
e-mails: Saeeda.csit@must.edu.pk, Aihab@iqraisb.edu.pk
Pakistan

**Abstract:** Information hiding is a key field to ensure the secure data movement over a network. This paper presents a novel zero text steganography approach using the features of text in write-ability context. Existing feature based schemes are either successive in achieving high capacity or imperceptibility, but are failed to maintain the balance among these opposing parameters. Moreover, previous feature based methods are less robust against steg-analysis, as during embedding process, these approaches modify the external appearance of the text. In the backdrop of the limitations in the prevalent text based steganography approaches, this paper proposes simple, yet novel approach that uses the feature of English alphabets rather than modifying them, which overcome the limitation of imperceptibility. Furthermore, the embedding capacity of the proposed technique is enhanced to 3-bits per characters. Moreover, embedding process does not change the external appearance of the text, which enhances the robustness factor to some extent. The experimental results determine that the proposed technique is prevailing with maximum embedding capacity, embedding capacity ratio with minimum time overhead as compared to existing techniques.

**Key Words:** Capacity; Feature coding; Imperceptibility; Information Hiding; Text Steganography;

## 1. INTRODUCTION

Information hiding is a key field to ensure the secure data movement over a network and the two basic approaches used for this purpose are Cryptography and Steganography [1]. This paper focuses on existing steganography techniques and presents novel text steganography approaches by enhancing the basic parameters for steganography.

In Steganography different data carrier agents [2] like image, audio, video and text are used as cover text to attain secure communication over a network. In text steganography the structure of cover text is not ambiguous, therefore, it is preferred and considered secure. However, because of the lack of available redundant information in a text file, using text as the target medium is relatively difficult as compared to the other target media. The different text steganography techniques [2] comprise of lexical, syntactical, ontological and Feature coding. Beside these, format based techniques like line shifting, character shifting and paragraph shifting are also elementary techniques [1, 3]. In Feature coding method the features of specified language, for example English language, are used to hide secret information.

Different existing text steganography methods, format-based methods, random and statistical generation methods, and linguistic methods, are commonly used in text steganography [4]. A comparative analysis of different approaches, based on above mentioned methods, are discussed along with their advantages and disadvantages over one another. Furthermore, text steganography methods like CALP, VERT and QUAD are compared and evaluated based on embedding ratio (ER) and saving space ratio (SSR) [5]. ER and SSR are two important factors to evaluate the format based text steganography methods.

In existing feature coding techniques [6, 7], the statistical features of characters are modified to conceal the secret information behind those characters of language, this rises the imperceptibility issue. If the formatting is applied on cover text at any level during communication process, the secret information will be lost and the receiver is not able to extract the secret message. On the other hand techniques [8, 9] are capable of storing 1 and 2 bits of secret message against single character of cover text respectively which limits the capacity parameter of the proposed algorithm.

In this paper, text is used as data carrier agent, the proposed technique uses the features of English alphabets of cover text to conceal the secret information. Beside this, instead of modifying the features of language that arises the imperceptibility issue, the proposed algorithm only uses the features to hide the secret bits behind those alphabets of cover text that possess well-defined features of proposed technique. The proposed technique outperforms two major issues of imperceptibility and capacity in text steganography discussed above and robustness to some extent. The proposed algorithm is evaluated and compared with existing methods depending on capacity, embedding ratio (ER) and time overhead.

The rest of the paper organized as per following sections: in section 2 the literature is examined by analyzing the existing methods for text steganography along with their limitations, Section 3 demonstrates the proposed novel technique, in section 4 and 5 proposed algorithms and experimental results are demonstrated respectively. Section 6 concludes the discussion and portrays the future work.

## 2. LITERATURE REVIEW

Text Steganography is achieved by researchers in many ways. In [8] it is implemented using the reflection symmetry property of English alphabets. By this method, letters are bisected along X- axis and Y – axis, the letters which are symmetric on any axis are placed in one group and which are on both axis, are placed in other group. Using this technique different bit patterns are assigned to different groups. The specified technique is secure enough as it does not incorporate any change to the text used, and also uses publically available text as cover which does not gain attention of unintended recipient. It is capable of concealing large volume of data, however, if the applicability of the method is explored, it is no more secure.

Data could also be saved depending on the circular nature of English alphabets. The method [9] specifies quadruple categorization considering the curves, vertical and horizontal lines present in English alphabets. This characterization is capable of storing 2 bits behind single character at one time i.e. it works on diagrams. The applicability of these approaches on particular data set should be kept secret to secure them.

Feature coding method is not limited to English alphabets, other languages i.e. Urdu, Arabic, Hindi and Persian, have great capacity to hide data by using this approach [10, 11, 12]. Considering the existence of pointed letters in Arabic language, the techniques modify the pointed letters to keep record of secret information in cover text, these are rich for data storage, but the information is lost in case of retyping.

In [13, 14] an efficient solution is designed to generate the dynamic cover text for concealing of secret information depending on the size of secret message. These approaches merge the inter-word and inter- paragraph spacing and enhance the capacity parameter. The cover text is also generated dynamically.

## 3. PROPOSED TECHNIQUE/METHODOLOGY

The proposed technique presents a new approach for text steganography to enhance the capacity parameter and inherits specified existing techniques [9] as well. It also overrides the imperceptibility issue, present in format based text steganography techniques, as it does not applies any modification to the cover text for embedding purpose. The proposed algorithm addresses the capacity issue by enhancing the embedding capacity of cover text that is capable of embedding 3 bit per character whereas it is limited to 2 bits in existing techniques [8, 9].

The block diagram of the system is shown in fig. 1. A three level decomposition of English alphabets is proposed, the distinct feature of this decomposing is the capability of a single character to store 3 bits of secret message at a time. Depending on this grouping, bit string of secret message is embedded on cover text and a key is

generated by embedding algorithm that is used at receiver end to extract the secret data by extraction algorithm.

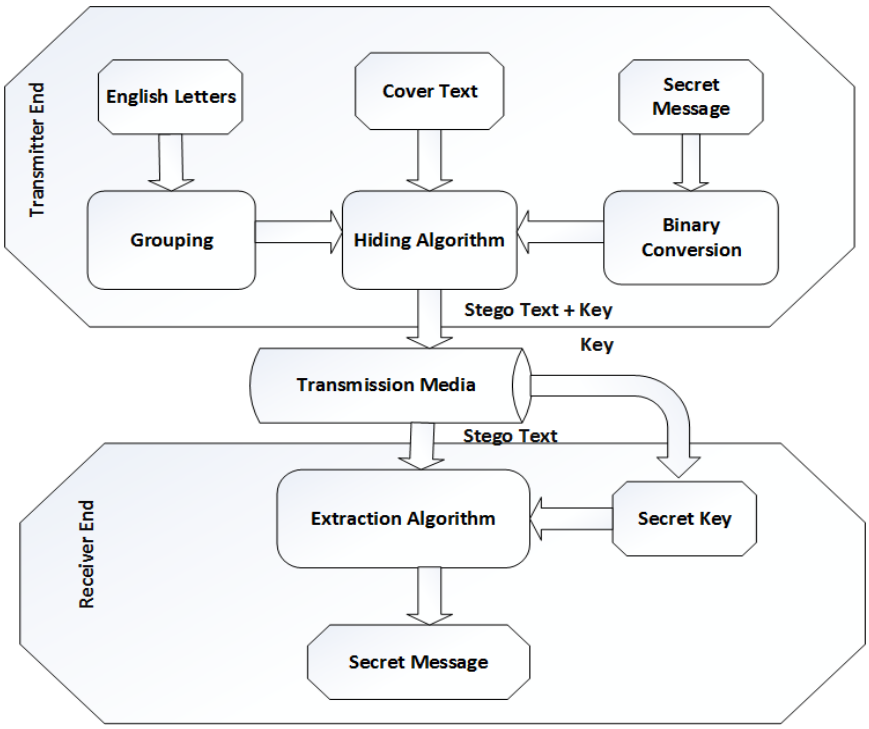


*Fig. 1. Proposed Text Steganography Model*

At first stage, the letters are divided into two groups depending on the write ability feature of English alphabets shown in table 1. Group A contains the letters that are writeable in one flow i.e. the boll point is not required to picked up in writing these letters, Whereas, group B contains the letters which do not follow the property. The former group letters need two or more attempts to write them and these groups will hide secret bits '1' and '0' respectively.

*Table 1. 1-bit Classification*

| ***Group Id*** | ***Group Name*** | ***Group Property*** | ***Secret Bits*** | ***Characters*** |
|---|---|---|---|---|
| *1* | *A* | *Letters writable in one flow* | *1* | *C, G, I, J, L, M, N, O, P, R, S, U, V, W, Y, Z* |
| *2* | *B* | *Letters not writeable in one flow* | *0* | *A, B, D, E, F, H, K, Q, T, X* |

After 1 bit classification, the letters will go for 2 bits and 3 bits classification respectively. Table 2 and table 3 elaborate all the resulting groups with their distinct properties respectively.

*Table 2. 2-bits Classification*

| ***Group Id*** | ***Group Name*** | ***Group Property*** | ***Secret Bits*** | ***Characters*** |
|---|---|---|---|---|
| ***1*** | *A* | *Letters writable in one flow and has vertical or horizontal lines* | *11* | *I, J, L, M, N, P, U, Y, Z* |
| ***2*** | *B* | *Letters writable in one flow but has no vertical or horizontal lines* | *10* | *C, G, O, S, V, W* |
| ***3*** | *C* | *Letters not writable in one flow and has vertical or horizontal lines* | *01* | *A, B, D, E, F, H, K, T,* |
| ***4*** | *D* | *Letters not writable in one flow and has no vertical or horizontal lines.* | *00* | *Q, X* |

*Table 3. 3-bits Classification*

| ***Group Id*** | ***Group Name*** | ***Group Property*** | ***Secret Bits*** | ***Characters*** |
|---|---|---|---|---|
| ***1*** | *A* | *Letters writable in one flow, has vertical or horizontal lines or both and has full/partial curves.* | *111* | *P, R, U* |
| ***2*** | *B* | *Letters writable in one flow, has vertical or horizontal lines or both and has no curves.* | *110* | *I, J, L, M, N, Y, Z* |
| ***3*** | *C* | *Letters writable in one flow, has no vertical or horizontal lines and has full/partial curves.* | *101* | *C, G,O, S* |
| ***4*** | *D* | *Letters writable in one flow, has no vertical or horizontal lines and has no curves.* | *100* | *V, W* |
| ***5*** | *E* | *Letters not writable in one flow and has vertical or horizontal lines and has partial/full curves.* | *011* | *A, B, D* |
| ***6*** | *F* | *Letters not writable in one flow and has vertical or horizontal lines and has no curves.* | *010* | *E, F, H, K, T* |
| ***7*** | *G* | *Letters not writable in one flow, has no vertical or horizontal lines and has full/partial curves.* | *001* | *Q* |
| ***8*** | *H* | *Letters not writable in one flow, has no vertical or horizontal lines and has no curves.* | *000* | *X* |

The final decomposition encapsulates 3 bits per character and constructs eight groups which is the distinct feature of proposed algorithm. The three level partitioning enhances the capacity of the algorithm, so an efficient algorithm is designed that produces optimal results.

## 4. ALGORITHMS

The proposed algorithms comprise of embedding and extraction algorithms. Embedding algorithm demonstrates how the secret bits are concealed against the cover text characters at sender side. The extraction algorithm tells the way secret message is extracted from cover text at receiver end by applying the secret key.

### 4.1. Embedding algorithm

1) **Start**
2) **Inputs:** groups[i], secret message, cover text.
3) **Procedure:**
   a) **Select** the cover text if it contains at least one char from each group
   b) **Convert** the Secret text to binary and apply padding if required.
   c) **Divide** the binary string to three bit patterns
   d) **Map** the three bit patterns to the cover text characters using group information.
   e) **Save** the characters position to an array to generate key.
4) **Outputs:** key, Groups
5) **END**

The detailed embedding model is depicted in figure 2 graphically. It elaborates all the steps involved in embedding phase.

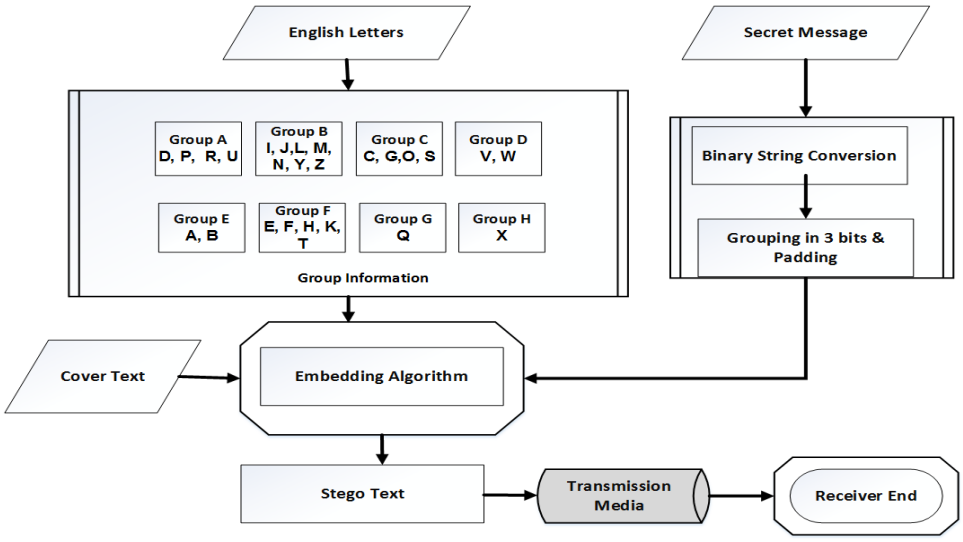


*Fig. 2. Embedding Process*

### 4.2 Extraction Algorithm

1) **Start**
2) **Inputs:** groups[i], Key, cover text
3) **Procedure:**
   a) **Extract** the cover text characters positions using Key.
   b) **Extract** the secret bits by mapping recovered characters to corresponding bit patterns using Group information.
   c) **Generate** binary string from extracted bit sequences.
   d) **Get** the ASCII values of seven bit sub strings
   e) **Convert** the ASCII values to corresponding characters
   f) **Generate** the character string.
4) **Output:** Secret Message
5) **END**

The extraction process is demonstrated in detail in figure 3. It shows how the steps involved to extract the secret information.

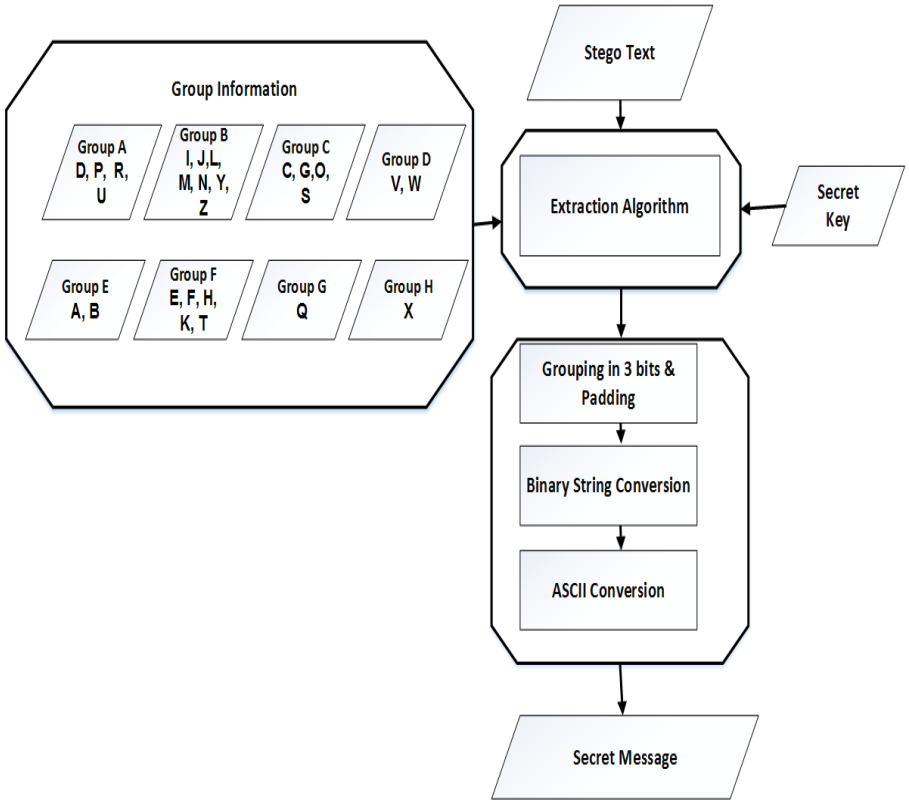


*Fig. 3. Extraction Process*

**For instance** the secret message is **'C'** and the cover text is **"TEXT STEGANOGRAPHY".** For embedding and extraction process following steps will be executed

1. Generate the binary string of secret message i.e. **01000011**
2. Make pair of three bits each, apply padding on last pair if not of three bits i.e. **010, 000, 011.**
3. Conceal the 3-bit secret pairs over cover text by finding the letters in cover corresponding to the secret bit patterns using group information. Save the index of those letters to generate key for extraction process. Following output is achieved by step 3.

| Secret bits | Cover Letters | Index Value (Key) |
|---|---|---|
| **010** | **T** | **0** |
| **000** | **X** | **2** |
| **011** | **A** | **9** |

The key (0, 2, 9) will be used for extraction process at receiver side.

4. Extraction algorithm will use the key values and extract the corresponding characters from cover i.e. T, X, A.
5. The corresponding bit patrons representing extracted characters will be retrieved using group information and a bit string will be generated i.e. **01000011**. If padding is applied during embedding process those bits will be eliminated from bit patron of last pair.
6. The resultant bit string is mapped to corresponding character value and secret information is retrieved i.e. **'C'**.

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

Three main parameters for text steganography i.e. capacity, imperceptibility and robustness are taken into consideration by proposed technique. The proposed technique is rich enough to outperform the limitations of existing techniques [8, 9]. The embedding capacity of single character is enhanced from 2 bits [8, 9] to 3 bits per character by exponentially enhancing the possible combinations of three bit patterns from $2^2$ to $2^3$ i.e. 4 to 8. The greater, the no of combinations, the greater will be the efficiency of cover text to embed data. It is analyzed that the imperceptibility is also no more an issue for the proposed algorithm as it only uses the features of English alphabets instead of modifying them for embedding purpose. The chance of revelation of proposed approach is also minimal without the knowledge of key and group information, so the robustness is implemented to some extent as well.

Table 4 and 5 give the comparative analysis of existing techniques against the proposed technique on the basis of concealing capacity, embedding ratio (ER) and time overhead using different data sets. The proposed technique is mature enough to

hide complete secrete message against specified cover text with considerably less time overhead.

### 5.1. Embedding Ratio

Embedding ratio is used to determine the total fitness of hidden text that can be embedded in cover text. This analysis is very important for steganography to understand the fitness capability of cover text.

$$ER = \left[ \frac{Total\ Number\ of\ Embedded\ Bits}{Total\ Bits\ of\ Expected\ Stego\ Text} \right] \quad X \quad 100\%$$

where:
Total bits of Stego-Text = Total Bits of Cover Text + Total Bits of Embedded Text.

*Table 4. Concealing power, embedding ratio and time overhead comparison*

| ***Text Steganography Approaches*** | ***Message Text Size (Bytes)*** | ***Cover Text Size (Bytes)*** | ***No. of Bytes can Hide*** | ***Embedding Ratio (ER) %*** | ***Time Overhead (ms)*** |
|---|---|---|---|---|---|
| ***Proposed Technique*** | *1000* | *3564* | *1000* | *21.91* | *330-335* |
| ***Zero Distortion Technique+ Abbreviation Method*** | *1000* | *3564* | *1000* | *21.91* | *27,602* |
| ***Method Based on Curve*** | *1000* | *3564* | *232* | *4.60* | *32,599* |
| ***Method Based on vertical Straight Lines*** | *1000* | *3564* | *220* | *4.32* | *30,617* |
| ***Quadruple Characterization*** | *1000* | *3564* | *205* | *3.93* | *32,269* |
| ***Feature Coding*** | *1000* | *3564* | *90* | *1.82* | *17,850* |
| ***Inter word Space*** | *1000* | *3564* | *79* | *1.60* | *21,926* |
| ***Random Character*** | *1000* | *3564* | *76* | *1.33* | *32,504* |

The self-explanatory graphical representations of all the eight text steganography approaches specified in table 4 and 5 with analyzed parameters hidden bytes and embedding ratio are given in fig. 4, 5, 6 and fig. 7 respectively.

*Table 5. Concealing power, embedding ratio and time overhead comparison*

| Text Steganography Approaches | Message Text Size (Bytes) | Cover Text Size (Bytes) | No. of Bytes can Hide (Bytes) | Embedding Ratio(ER) % | Time Overhead (ms) |
|---|---|---|---|---|---|
| **Proposed Technique** | 800 | 2640 | 800 | 23.25 | 280 -285 |
| **Zero Distortion Technique + Abbreviation Method** | 800 | 2640 | 800 | 23.25 | 29,806 |
| **Method Based on Curve** | 800 | 2640 | 172 | 6.12 | 37,996 |
| **Method Based on vertical Straight Lines** | 800 | 2640 | 161 | 5.75 | 27,533 |
| **Quadruple Characterization** | 800 | 2640 | 146 | 5.24 | 26,562 |
| **Feature Coding** | 800 | 2640 | 66 | 2.44 | 18,180 |
| **Inter word Space** | 800 | 2640 | 58 | 2.15 | 20,825 |
| **Random Character** | 800 | 2640 | 48 | 1.78 | 31,292 |

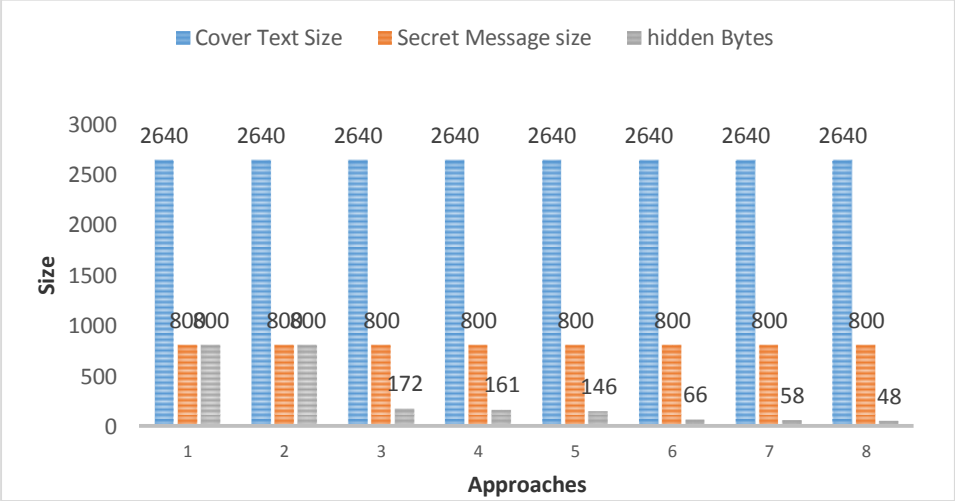


*Fig. 4. Embedding Bits Distribution w.r.t. cover and secret text size*

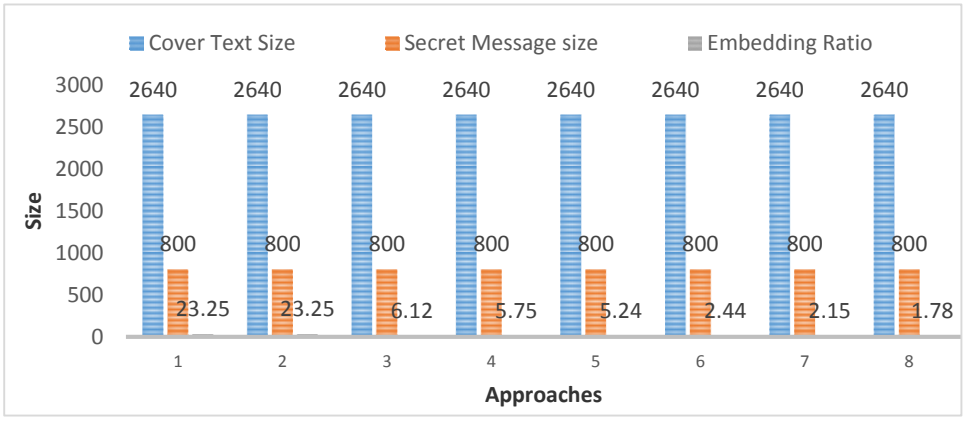


*Fig. 5. Embedding Ratio Distribution w.r.t. cover and secret text size*

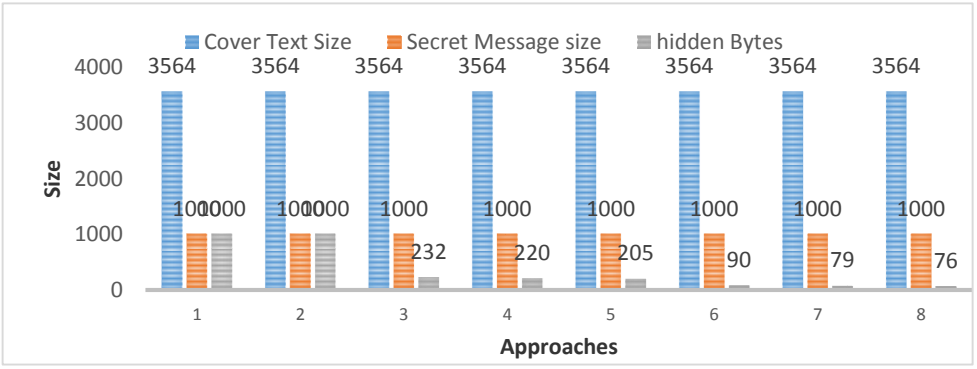


*Fig. 6. Embedding Bits Distribution w.r.t. cover and secret text size*

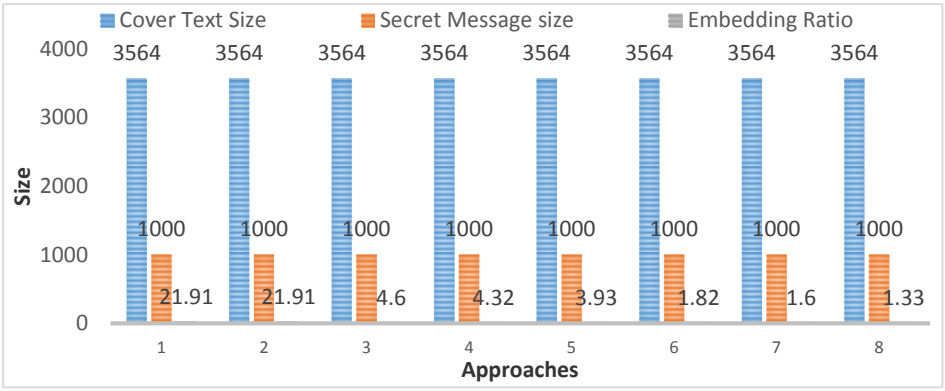


*Fig. 7. Embedding Ratio Distribution w.r.t. cover and secret text Size*

These graphical representations depict the strength of proposed technique in all the fields. It demonstrates that the proposed method has maximum embedding capacity and maximum embedding ratio with minimum time overhead as compared to all existing format based techniques [9, 15]. Hidden capacity of the proposed method is equal to the size of the message text provided to the method i.e. it can hide all the bytes of secrete text over the cover text that is very rare in format based approaches as analyzed from table 4 and 5. Moreover, the enhanced percent embedding ratio and minimum time overhead is also a powerful feature of proposed technique.

## 5. CONCLUSION

This paper presents a novel approach based on features of English alphabets for text steganography. The shapes of English alphabets are exploit to a well-defined criteria to hide secret bits. The algorithm is capable of hiding relatively large amount of data with maximum embedding ratio by using proposed method and minimizes the time overhead. In addition, it is immune to text formatting as the imperceptibility problem is outperformed.

Three basic text steganography parameters i.e. capacity, imperceptibility and robustness are addressed by the proposed method. The technique is suitable for confidential and secret information transformation over an unsecure communication channel. It enhances the integrity of data being an important parameter for secure data communication.

## REFERENCES

[1] S. S. Pawar, & V. Kakde, Review On Steganography For Hiding Data. *International Journal of Computer Science and Mobile Computing*, **4** (Vol. 3), April 2014, pp. 225-229.

[2] A. Koluguri, S. Gouse, & P. B. Reddy, Text Steganography Methods and its Tools, *Int. J. Adv. Sci. Tech. Res*, **4** (vol. 2), 2014, pp. 888-902.

[3] E. Zielińska, W. Mazurczyk, & K. Szczypiorski, Development Trends in Steganography. *Communications of the ACM*, **3** (Vol. 57), 2014, pp. 86-95.

[4] V. Saraswathi, & M. S. Kingslin. Different Approaches to Text Steganography: A Comparison. *International Journal of Emerging Research in Management &Technology*, **11** (Vol. 3), 2014.

[5] B. Osman, R. Din, & M. R. Idrus, Capacity Performance of Steganography Method in Text Based Domain. *ARPN Journal of Engineering and Applied Sciences*, 2015.

[6] S. Changder, S. Das, & D. Ghosh, Text Steganography through Indian Languages using Feature Coding Method. *2nd International Conference on Computer Technology and Development (ICCTD)*, Nov. 2010, pp. 501-505.

[7] A. Odeh, A. Alzubi, Q. B. Hani, & K. Elleithy. Steganography by Multipoint Arabic Letters. In *IEEE Systems Applications and Technology Conference (LISAT)*, Long Island. May 2012, pp. 1-7.

[8] A. Majumder, & S. Changder. A Novel Approach for Text Steganography: Generating Text Summary Using Reflection Symmetry. *Procedia Technology* (10), CIMTA, 2013, pp. 112-120.

[9] S. Dulera, D. Jinwala, & A. Dasgupta. Experimenting with the Novel Approaches in Text Steganography. *International Journal of Network Security & Its Applications (IJNSA)*, **6** (Vol.3), November 2011.

[10] S. Kingslin, & N. Kavitha. Evaluative Approach towards Text Steganographic Techniques. *Indian Journal of Science and Technology*, **29** (Vol. 8), Nov. 2015.

[11] A. T. Abbasi, S. N. Naqvi, A. Khan, & B. Ahmad. Urdu Text Steganography: Utilizing Isolated Letters. *13th Australian Information Security Management Conference*, Nov. 2015, pp. 37-46.

[12] M. H. Shirali-Shahreza, & M. Shirali-Shahreza. A New Approach to Persian/Arabic Text Steganography. In *5th IEEE/ACIS International Conference Computer and Information Science & 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR 2006)*; July 2006, pp. 310-315.

[13] L. Y. Por, & B. Delina. Information Hiding: A New Approach in Text Steganography. *7th WSEAS Int. Conf. on Applied Computer & Applied Computational Science (ACACOS '08)*, Vol. 7, April 2008.

[14] L. Y. Por, T. F. Ang, & B. Delina. Whitesteg: a new scheme in information hiding using text steganography". *WSEAS Transactions on Computers*, **6** (Vol. 7), June 2018, pp. 735-745.

[15] V. K. Yadav, & S. Batham. A Novel Approach of Bulk Data Hiding using Text Steganography. *3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*, Procedia Computer Science, 57, 2015, pp. 1401-1410.

***Information about the authors:***

**Saeeda Kouser** – MS student at Iqra University Islamabad campus and working as lecturer at Mirpur University of Science and Technology, MUST, Mirpur Azad Kashmir. The area of research is Information Security and conducting the research work under the Supervision of Dr. Aihab Khan.

**Dr. Aihab Khan** – Working as associate Professor at Iqra University Islamabad. The area of research is Information Security and Watermarking, the number of research papers are published in different international Journals in specified fields.

**Manuscript received on 28 May 2017**