mathematics

MDPI

Review

A Review on Text Steganography Techniques

Mohammed Abdul Majeed *, Rossilawati Sulaiman, Zarina Shukur and Mohammad Kamrul Hasan

Center for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia; rossilawati@ukm.edu.my (R.S.); zarinashukur@ukm.edu.my (Z.S.); mkhasan@ukm.edu.my (M.K.H.)

* Correspondence: p97938@siswa.ukm.edu.my

Abstract: There has been a persistent requirement for safeguarding documents and the data they contain, either in printed or electronic form. This is because the fabrication and faking of documents is prevalent globally, resulting in significant losses for individuals, societies, and industrial sectors, in addition to national security. Therefore, individuals are concerned about protecting their work and avoiding these unlawful actions. Different techniques, such as steganography, cryptography, and coding, have been deployed to protect valuable information. Steganography is an appropriate method, in which the user is able to conceal a message inside another message (cover media). Most of the research on steganography utilizes cover media, such as videos, images, and sounds. Notably, text steganography is usually not given priority because of the difficulties in identifying redundant bits in a text file. To embed information within a document, its attributes must be changed. These attributes may be non-displayed characters, spaces, resized fonts, or purposeful misspellings scattered throughout the text. However, this would be detectable by an attacker or other third party because of the minor change in the document. To address this issue, it is necessary to change the document in such a manner that the change would not be visible to the eye, but could still be decoded using a computer. In this paper, an overview of existing research in this area is provided. First, we provide basic information about text steganography and its general procedure. Next, three classes of text steganography are explained: statistical and random generation, format-based methodologies, and linguistics. The techniques related to each class are analyzed, and particularly the manner in which a unique strategy is provided for hiding secret data. Furthermore, we review the existing works in the development of approaches and algorithms related to text steganography; this review is not exhaustive, and covers research published from 2016 to 2021. This paper aims to assist fellow researchers by compiling the current methods, challenges, and future directions in this field.

Keywords: text steganography; data hiding; format-based; linguistic; random and statistic

check for updates

Citation: Majeed, M.A.; Sulaiman, R.; Shukur, Z.; Hasan, M.K. A Review on Text Steganography Techniques. *Mathematics* **2021**, *9*, 2829. https://doi.org/10.3390/math9212829

Academic Editor: Eckhard Pfluegel

Received: 30 August 2021
Accepted: 30 September 2021
Published: 8 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the current era, developments in digital communication play a vital role in daily life. Improvements in web-based technologies coupled with information digitalization have significantly increased the use of data transfer. Information security is an important issue in the context of protecting user information. Although several existing approaches are robust and secure, work remains in progress to make these approaches safer and more secure in terms of performance indicators. In general, there are two classes of information security systems: information hiding and encryption [1]. Both classes are capable of protecting information; however, their approaches are different. The researchers in [2] formulated different types of data encryption techniques. Figure 1 depicts a general data security mechanism classification, which interconnects the three techniques shown: steganography, watermarking, and cryptography. Steganography is divided into either linguistic or technical steganography, and watermarking is divided into robust or fragile watermarking.

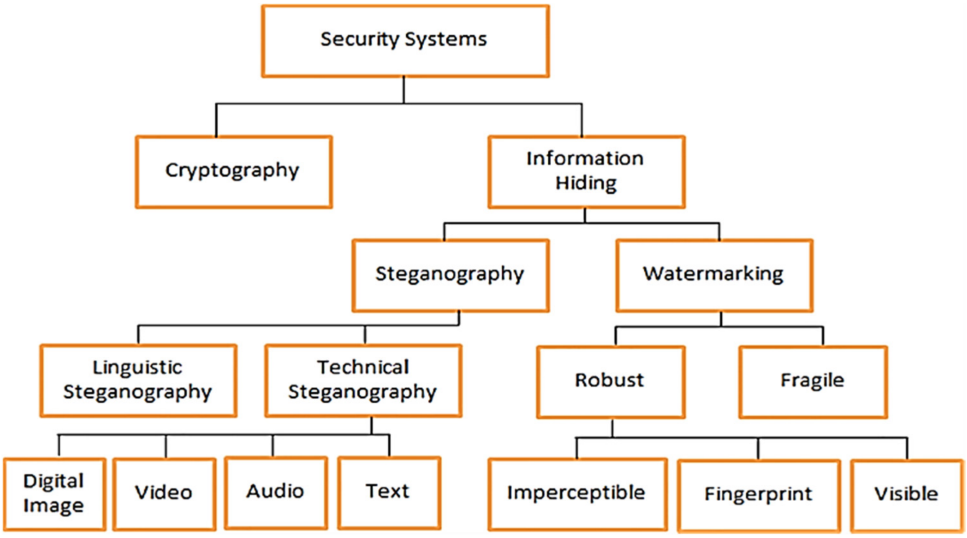


Figure 1. Classification tree of a general data security system.

Cryptography is one of the most appealing domains in securing data. In this approach, different forms of data encryption are used to transform sensitive data into an unreadable structure, called ciphertext. The ciphertext is apposite for transmission with regard to data security because it is illegible to any parties other than the intended sender and recipient. The ciphertext is produced using several substitution and permutation methods in such a manner that a third party is unable to obtain access to the actual message [3]. A key is needed for implementing cryptographic methods. This key can either be symmetric, which contains only a single key for encryption and decryption, or asymmetric, which comprises a pair of keys (public and private) for both encryption and decryption. The public key is used for encryption, and the corresponding private key is used for decryption [4].

Steganography is another means of securing messages during data communication. Although both cryptography and steganography share the same objective, the approaches vary. In contrast with cryptography, steganography retains its original data by hiding it in other media [5], whereas cryptography transforms the original data into ciphertext. The drawback of cryptography is the existence of the original data, irrespective of whether the original data are subject to encryption. Therefore, steganography methods offer a supplementary security layer for the message while communicating the data. Robustness in steganography and cryptography are viewed differently. When an attacker is able to access the actual data, then the cryptographic system is no longer secure. In contrast, if the attacker has authority over the secret data in the steganography system, the system is considered to no longer be secure [6].

In circumstances in which data security is a concern, watermarking is applicable. Watermarking is a renowned data security approach that focuses on authentication of images and protection in terms of copyright. In watermarking, digital data is embedded within multimedia data using visible or invisible watermarking, based on the visibility of files. The watermarking procedure prevents dishonest duplication of media files and claims of quasi-ownership [7], which is useful in communications among many entities. The main objective of watermarking is to produce a robust, safe, and efficient watermark on any media files and documents that are long-lasting and unchangeable for unapproved entities [8].

Watermarking and steganography overlap in a wider context, which is to hide secret data within other media. Both approaches boast attributes such as capacity, security, imperceptibility, and robustness; however, the priority is different.

Robustness is the top priority in watermarking, whereas imperceptibility is the main priority in steganography [9,10]. As mentioned previously, both steganography and watermarking have different focuses. Steganography intends to offer secure communication by hiding secret data in other media, whereas watermarking intends to provide a safeguard against the abuse of copyrights.

The similarity between watermarking and fingerprinting is that both entail the marking of objects. The only difference is that every fingerprinting entails diverse fingerprints to verify ownership, whereas in watermarking, various objects have the same watermark [2]. In steganography, secret messages are hidden within another object. For watermarking, a specific object is purposely embedded so that any third party can see it [9,11]. The main challenge in steganography is that one must keep the embedded message invisible while preserving the imperceptibility of the cover media, and maintain high capacity in terms of hiding as much of the secret message as possible. In addition, robustness is the main challenge for both watermarking and cryptography. Finally, keys must be used in cryptography but are optional in the other approaches. Table 1 compares the characteristics of steganography, watermarking, and cryptography.

Table 1. Comparison among different information security techniques.

Characteristics	**Steganography**	**Watermarking**	**Cryptography**
Goal	Protects secret data from discover	Protects legitimacy of the media	disorganizes the content of data
Cover Choosing	free cover choice	Limit	Not use
Challenges	Imperceptibility, security, and Capacity	Robustness	Robustness
Keyes	Possible	Possible	Necessary
Output	Stego-media	Watermarked-media	Encryption-text
Visibility	Definitely not	Occasionally	Constantly
The system is invalid if	Noticed	Detached or substituted	Decryption
Attacks	Steganalysis	image processing	Cryptanalysis

As a result of the growth in natural language processing (NLP) technology in recent years, researchers have begun to examine the automatic generation of steganographic text to transfer secret data. This type of steganographic method is classified as a form of natural modification of the cover media, which embeds data during text generation. This study reviews the evolution of steganography and links this development with existing categories to help researchers understand the existing methods.

This paper reviews text steganography methods and covers research published from 2016 to 2021. Languages considered included English, Arabic, Chinese, Indian, and Bengali . In addition, the current research directions in the field are shown in the later sections. The contributions of this paper are summarized as follows:

- presents a brief review of existing text steganography methods;
- summarizes text steganography classes: statistical and random generation, format-based methodologies, and linguistics, while identifying their methodologies from 2016 to 2021;
- recommends future work in the field of text steganography.

The remainder of this paper is organized as follows—Section 2 presents Materials and Methods. Section 3 offers a background study of steganography. Section 4 discusses the general procedure in steganography, followed by the properties of steganography in Section 5. Then, Section 6 discusses text steganography categories, including various existing works. Section 7 provides a discussion and outlines future directions in these fields, and finally, Section 8 concludes this paper.

2. Materials and Methods

This study's review focused on the existing techniques and methods of text steganography published from 2016 to 2021. Studies that are not related to text steganography were excluded. We only used the complete edition of the report in original formats. This section includes the subsections Data Sources, Search Process, Data Selection, and Data extraction.

2.1. Data Sources

Research papers that were related to the study were selected from the following libraries as our primary resources:

- Science Direct (www.sciencedirect.com (accessed on 25 January 2021)).
- Institute of Electrical and Electronics Engineers (IEEE) Xplore Digital Library (ieeexplore.ieee.org (accessed on 31 January 2021)).
- Springer Link (link.springer.com (accessed on 8 February 2021)).
- Taylor and Francis (www.taylorandfrancisgroup.com (accessed on 17 February 2021)).
- MDPI (www.mdpi.com (accessed on 25 February 2021)).
- Google scholar (www.scholar.google.com (accessed on 7 March 2021)).

2.2. Search Process

The search was focused on text steganography, and several keyword patterns were used during this process. Boolean operators were used to helping clarify data on keywords for each research publication. Symbols and Boolean operators were used, such as "OR" and "AND", to check for the following keywords:

- (text steganography OR (text steganographic) OR (text AND hiding)) AND (format based OR linguistic OR random OR statistical).
- ((text data AND steganography) OR (text steganography AND method) OR (steganography AND text)) AND (neural network OR deep learning) OR (Natural language OR NLP).

2.3. Data Selection

Data selection is an essential factor in reviews of existing studies. We utilized three filter processes to include the criteria for the search after receiving the results of the keywords we used. During the initial phase of the filter, the criteria were determined. We organized all of the study findings into research articles based on these keywords. Then, the next step was to complete a second filter that looks at the title and read the abstract related to the research question. Finally, the third filter included reading the content from a research paper selected from the candidate studies. The data selection criteria were as follows:

- Was the research article published between the years 2016 and 2021?
- Is the research article reported in any of the referred data sources?
- Does the research article mention or discuss one of the text steganography categories?

2.4. Data Extraction

We looked at each preliminary study to determine if there were any text steganography topics. We kept recorded all the studies we found in a spreadsheet with their names, descriptions, and justifications. In March 2021, we completed the search process, discovering 183 publications. The relevant research articles were carefully extracted using the search technique indicated in Figure 2 by following the selection and refusal parameters. In the end, 50 preliminary studies were identified.

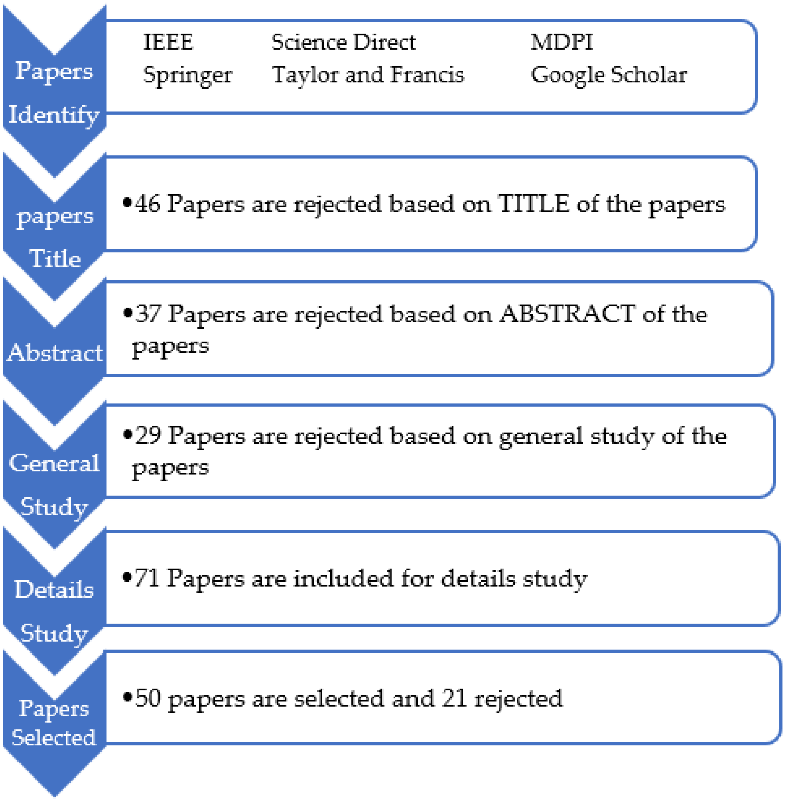

Papers Identify
IEEE
Science Direct
MDPI
Springer
Taylor and Francis
Google Scholar
papers Title
•46 Papers are rejected based on TITLE of the papers
Abstract
•37 Papers are rejected based on ABSTRACT of the papers
General Study
•29 Papers are rejected based on general study of the papers
Details Study
•71 Papers are included for details study
Papers Selected
•50 papers are selected and 21 rejected

Figure 2. Search process.

3. Background of Steganography

The term "Steganography" comprises two ancient Greek words: "Stegano" and "Graphy", and both refer to "Cover Writing". Steganography was first used centuries ago. For example, Histiaeus used steganography to send a secret message by inking secret messages (tattooing) on his slave's skull, who travelled only after the hair had grown to cover the tattoo. Greeks were well-known for sending hidden messages. For example, Demaratus utilized a wax-coated tablet. A message was written on the tablet by carving over it, and then the tabled was waxed. The message was scuffed after the wax was scraped off. The tablet was coated again with wax, making it appear like another blank tablet. Thus, the message was then sent securely without any suspicions [1,5,12,13]. Approximately 50 decades ago, Jerome Carden, a Italian mathematician, restored the writing of hidden messages, which the Chinese utilized in ancient times. The technique involved using a paper with grid holes as a mask and penning a hidden message by putting it on a blank paper. This mask was shared between the sender and receiver. After removing the grid mask, the blank section of the paper was written, with what appeared to be harmless text. In the First World War, Germans employed microdot technology based on multiple phases and utilized the waste from magazines [14]. There were numerous methods for penning hidden messages during the Second World War, such as the Enigma machine, open-coded messages, various null ciphers, and invisible inks [5]. A Saudi Arabian king also began a project for stealthy writing at the Abdulaziz City of Science and Technology, which was discovered in a manuscript that was 1200 years old. These manuscripts were gathered from Turkey and Germany [1,15].

Different literature works offer deeper perspectives regarding the history of steganography and approaches employed globally [5,12,16,17]. As a result of the advancement of interconnected multimedia setups, wireless systems, and electronic digital cameras, the digitalization of data has significantly enhanced the prospect of revival and dissemination of information. Steganography techniques have been adapted to digital means due to advancements in the processing capability of computers and the speed of the Internet.

New research works and advancements, such as those found in signal processing [18,19], encoding techniques, and the theory of information, are aiding the development of secure steganography approaches. The newest steganography approaches are not just restricted to concealing stealthy information within images, but they also help in embedding data in text [20,21], codes [22], audio [23,24], videos [25], and DNA [26,27]. They also comprise concealing information in different formats, which can be found in extensible mark-up language (XML), executables, and hypertext mark-up language (HTML) [28,29]. The literature works in [30,31] also appraised and studied different trends in digital methodologies of steganography.

4. Steganography General Procedures

The fundamental notion of any approach or technique in digital steganography is to discreetly hide private or secret data within cover media. The type of secret data can be in many forms, such as image, text, binary, or video. This also applies to the cover media, which can also be videos, images, or text. The data hidden or embedded within a host or cover media are called secret or hidden data. The outcome of the embedding process is called stego media. The intention is to communicate the secret data through any unsecured communication channels. Figure 3 shows a block diagram of a generic steganographic setup [1]. The secret data is hidden inside cover media at the sender side to generate stego media to ensure safe transmission to the receiver side.

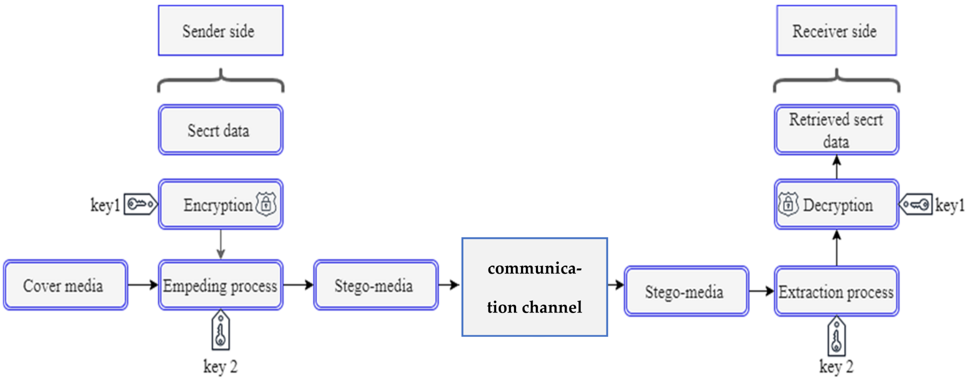


Figure 3. Block diagram of a steganographic system.

Systems looking for improved security characteristics may frequently utilize a security key, encryption structure, or both during embedding. The key may hold additional information, such as encryption passwords, embedding maps, and a threshold value used to choose a particular coefficient for the embedding process. Generally, an embedded system can be signified as:

$$c' = \mathrm{Em}\ (C, \mathrm{En}(S, k_1), k_2) \tag{1}$$

where S, C and C′ are the secret data, cover media, and stego media, respectively. For En(.) to function as encryption, two secret keys are needed (k_1 and k_2), and are used together with Em(.) to function as embedding. Then, C′ is sent to the recipient, and a decryption algorithm and reverse embedding are intended to be applied over the stego media, signified as Equation (2):

$$s_r = D\ (\mathrm{Ex}\ (c^*,\ k_2),\ k_1) \tag{2}$$

where S_r is the original data. Ex(.) and D(.) are functions for extraction and decryption, respectively. c^* is the doctored stego media (because of intruder attacks or channel noise) obtained at the recipient's side.

5. Attributes of Steganography

Security, imperceptibility, and capacity are the three attributes of steganography required to conceal secret data [32]. Research in [33–35] stated the three properties, in addition to robustness. These are the most influential factors that determine the efficacy of a steganography setup. According to its use, there are certain specific requirements for managing a number of steganographic designs. Steganography and watermarking have these attributes for data embedding. However, a common trade-off is between the size of secret data and the quality of the stego files. Suppose a large quantity of secret data is to be embedded. In that case, modifying the stego files is harder because the imperceptibility is more difficult to achieve because of a possibility of distortion [36]. Therefore, optimally maintaining these attributes should be the main objective. Robustness is not always an obligation. However, security, imperceptibility, and the capacity to conceal are always necessary. Regarding digital watermarking, greater capacity and imperceptibility are not compulsory. Venkatraman et al. [37] noted that robustness is essential when responding to malevolent and unwanted attacks. Figure 4 depicts the main requisites for a steganography setup [38]. The following segment comprehensively elucidates the discussed attributes. The maximum number of hidden messages planted in the cover text without affecting the text content is referred to as capacity, and is usually represented in bits. The ability of an attack vector to quickly deduce secret information is referred to as security. The term "robustness" refers to the ability to resist the risk of secret data being modified or destroyed.

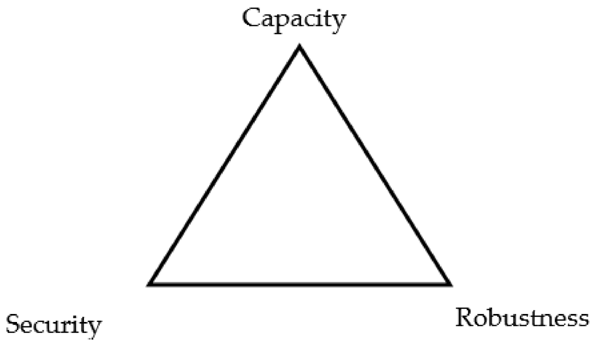


Figure 4. Main requirements for a steganography system.

5.1. Imperceptibility

Imperceptibility is the foremost priority in steganography, and aims to conceal the secret data within other media. It is not possible for the human eye to understand it even when statistical methods are applied [11]. Statistical methodologies are beneficial means for attackers to determine if secret data is being transferred with the communication between two parties. Therefore, the cover media should not be noticeably changed, in terms of the statistical standards, due to the embedding of the secret data; that is, if similar statistical data are found in both the original and the stego files, then we can consider that the security is sufficiently high to allow communication of the data. The quality of the cover media must be maintained when sharing through unsecured networks despite noise from the embedding process [36,37,39].

5.2. Security

The term "security" in steganography relates to "undetectability" or "unnoticeability". Therefore, a steganography approach is secure when the data it is hiding cannot be detected using statistical techniques by any third party. Security is the main requirement to prevent access by unlawful individuals or computers while communicating using an unsecured channel, thus ensuring that the data remain secure.

5.3. Payload Capacity

An effective steganographic setup, in general, intends to send the highest quantity of information by utilizing the least-covered media. This decreases the prospect of interception while communicating through an unsecured channel and thus typically demands a high embedding capacity. Ref. [37] described the rate of embedding the bits against the size of the cover image. A key challenge in steganography is to maintain a high payload capacity while maintaining security and imperceptibility.

5.4. Robustness

Robustness signifies the capability of the embedding and extracting method to withstand any corruption by a third entity through any processing methods [34,35]. In the case of steganography, if the stego files are not impacted or altered while being dispatched over the Internet, then it is not considered an attack, and the recipient receives the stego file as intended. Otherwise, attacks such as compression, conversion of file format, and transformation between digital and analogue format can happen during the communication process. However, for fingerprint systems, robustness is necessary when modifying or altering files on purpose.

6. Text Steganography Categories

Text steganography is considered exceedingly difficult due to the inadequate redundant data in textual files compared with other digital media, such as audio, image, or video files [38]. Text steganography can be generally split into three classes, as depicted in Figure 5: linguistic, format-based, and random and statistical generation [21].

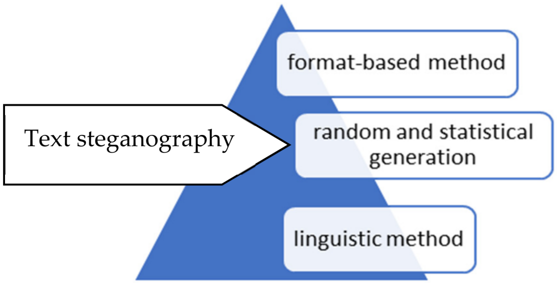


Figure 5. Categories of text steganography.

6.1. Format-Based Method

In this form of steganography, the physical features of text symbols are used. The features are altered in such a manner that the human eye cannot sense them. For example, lines in the text are moved up and down to conceal the bits of secret data. Then, words are moved left or right, or up and down. In some cases, white spaces among the words or between the paragraphs or lines are used to hide data. In feature-based encoding, the physical features of the words are altered to conceal the information. This is reliant on symbols and languages. Numerous studies into format-based approaches enhance the capacity of text steganography by changing the physical nature of the text format. For instance, ref. [40] maps the secret message's binary digits with the cover text's binary digits using the American Standard Code for Information Interchange (ASCII) characters, comprising punctuation, spaces, and symbols. The secret text is initially encrypted using a one-time pad and transformed into ciphertext. Then, each character is transformed into 7-bit binary numbers. The embedding procedure is carried out by mapping one bit of the secret text onto the first bit of the ciphertext character comprising the same quantity of bits. Each bit position for the bit of the secret text is documented as a stego key, which is placed on the bit of the ciphertext. The stego key functions as a key to extract the secret text embedded within the ciphertext.

Ref. [41] recommended a model that used color coding and deployed the Lempel–Ziv–Welch (LZW) compression method. The method utilized the forward mail as cover media to conceal the secret data. The algorithm initially compresses the secret data and then hides it in the email addresses and the email's cover message. The secret data bits are embedded within the cover text (or message) by coloring it via a color-coding table. The outcomes of the study indicated that the technique had a greater embedding capability than other methods, with lower computational complexity. Furthermore, the security of the recommended technique is considerably enhanced by using stego keys.

Ref. [42] offers a format-based text steganographic method centered on color coding using two approaches: permutation and numeration systems. With a secret message and a cover text, the method embeds the secret message in the cover text by coloring it. The stego text is then dispatched by mail to the receiver. The outcomes of the research indicated that these models demonstrate a high hiding capacity compared with other approaches.

In [43], the authors suggested a format-based text steganography technique for Arabic based on the Kashida and Unicode text (Zero-width non-joiner [ZWNJ] (discrete), zero-width joiner [ZWJ], little space, and zero-width space along with traditional spaces). The cover text can be used to conceal one bit of the secret text in each letter by transforming the letters using their position (i.e., end, middle, or beginning of a word, or an isolated word). Figure 6 demonstrates that the Arabic language letters are depicted differently despite sharing the same Unicode text corresponding to all shapes used in the file. The shapes of the letters are corrected using software that changes typographic sequence depending on letter position (i.e., isolated, end, middle, or beginning). Kashida, typed as "_", represents a character in Arabic that extends a letter but does not change word meaning. The word (احمد) may be indicated as (احـــمـــد) by concatenating two Kashida characters to embed secret data. Embedding is performed in several layers. Every layer comprises the words sharing similar tags as evaluated by the part of speech (POS) tagging tool, which divides the cover text into lists of words in which each word is correlated with the corresponding POS. Words having a common POS are combined to form a layer. Subsequently, the word count for each layer is determined to sort them in ascending order. The stego key is employed to randomly select embedding layers to improve security. The outcomes of the study indicated that this method met the hiding capacity goal.

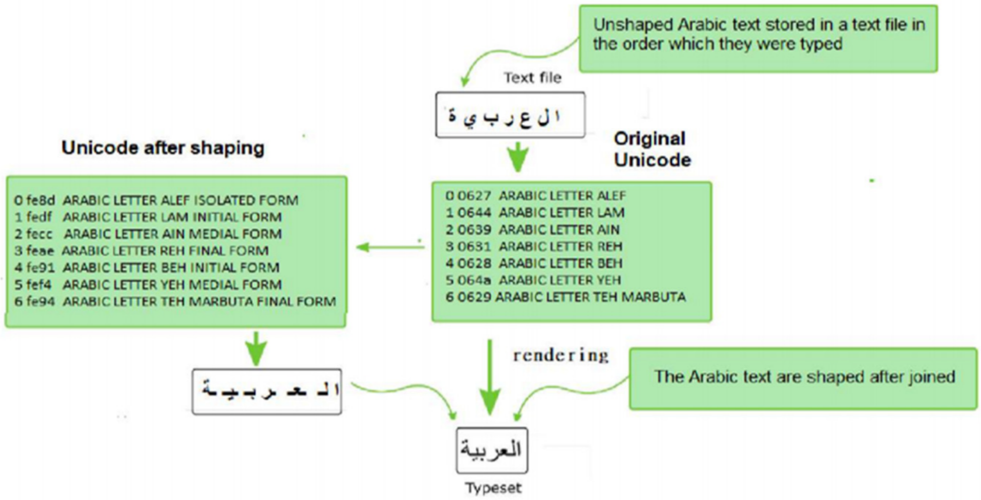

Unshaped Arabic text stored in a text file in the order which they were typed
Text file
ا ل ع ر ب ي ة
Original Unicode
Unicode after shaping
0 0627 ARABIC LETTER ALEF
1 0644 ARABIC LETTER LAM
2 0639 ARABIC LETTER AIN
3 0631 ARABIC LETTER REH
4 0628 ARABIC LETTER BEH
5 064a ARABIC LETTER YEH
6 0629 ARABIC LETTER TEH MARBUTA
0 fe8d ARABIC LETTER ALEF ISOLATED FORM
1 fedf ARABIC LETTER LAM INITIAL FORM
2 fecc ARABIC LETTER AIN MEDIAL FORM
3 feae ARABIC LETTER REH FINAL FORM
4 fe91 ARABIC LETTER BEH INITIAL FORM
5 fef4 ARABIC LETTER YEH MEDIAL FORM
6 fe94 ARABIC LETTER TEH MARBUTA FINAL FORM
الـ ـعـ ـر بـ ـيـ ـة
rendering
The Arabic text are shaped after joined
العربية
Typeset

Figure 6. Arabic text shaping.

Ref. [44] proposed a method by adding five whitespace characters to randomized positions in a line using the key to correlate to the characters required for embedding secret information. This method is advantageous because randomly spread whitespaces may encode the message differently using different keys. The whitespaces contained in the secret text regulate the embedding process.

Ref. [45] suggested a technique to enhance the embedding capacity for format-based text steganography using the font and other text characteristics for encoding secret information. This technique uses similar symbols for several codes, known as Set of High-Frequency Letters (SHFLs), used for embedding. The embedding process is based on replacing English letters with codes that share similar shapes. One pass encodes two bits, where 00, 01, 10, and 11 conceal glyph1, glyph2, glyph3, and glyph4, respectively (Table 2). The steganographic capacity of the table can be enhanced, and the technique is based on lower-case SHFL. This two-bit technique outperforms the standard text steganography because it enhances the embedding capacity of the stego file.

Table 2. Selected letters in SHFL for the hiding process.

Letters	ASCII Code	Unicode		
	S = 00	S = 01	S = 10	S = 11
e	0065	0023	0026	002A
t	0074	003C	003D	003E
a	0061	005B	005D	005E
o	006F	007B	007C	007D

The research in [46] relates to frequency modulation methods. Secret information is embedded using character spacing and font attributes. This technique typically hides one of every three characters. Given this coding frequency, there are eight distinct ways to hide one character and enhance embedding capacity.

Another study by [47] recommended an algorithm to raise the capacity of the embedding message by merging the Arabic extension character Kashida with three small space characters: hair space, thin space, and Six-PRE-EM space. Thin space occupies about 20% of the width of a typical space character. It is used for adding a narrow space. Hair space denotes the least-width space employed for distancing words or letters. Similarly, the six-PRE-EM denotes a space typically 1/6th of the space character. The secret bits are hidden inside the text using both small spaces and Kashida characters. The presence of Kashida means that a word is embedded in one bit; similarly, the presence of whitespace indicates three hidden bits corresponding to the secret message.

The character insertion technique uses the process as described. If the bits corresponding to the secret information have a bit value of "1" and there is a letter connection in addition to the presence of dots, then Kashida will follow the connected letter. Otherwise, Kashida is not concatenated. Once every character in the string is assessed, and all feasible Kashida characters are concatenated at the respective positions, whitespace is used between two consecutive words. Subsequently, three small spaces are concatenated with the typical whitespace character before progressing to the following word. All spaces are not concatenated; those that correspond to the secret bit sequence are concatenated. The secret bits are split into a sequence of three. The first, second, and third bits represent the thin space, hair space, and Six-PRE-EM space, respectively.

A single space is inserted if the bit value equals "1"; on the other hand, "0" leads to no insertion. For instance, if the bit sequence of the group is 000, none of the three spaces is added.

However, if the bit sequence is 111, all spaces are concatenated with the standard whitespace character. If the bit sequence is 010, only the hair space is concatenated with the standard whitespace. The insertion process concludes when all bits corresponding to the secret message are processed. This technique offers enhanced hiding capacity compared with the present state-of-the-art Kashida and space-based methods.

Another study in [48] suggests a technique that seamlessly concatenates Unicode Standard characters with Arabic characters to ensure safe use at the individual level. This technique is based on Arabic characters represented using the Unicode contextual characters. Characters are used in their original form to conceal the "0" bit, whereas the contextual form is represented to hide the "1" bit. Furthermore, this technique introduced additional characters (such as Kashida and ZWJs) between each pair of unconnected contextual and original characters by maximizing the use of the spaces among letters and words. This is achieved by transforming the Unicode spaces located between the words and concatenating unconnected letters using ZWNJs. The process enhances the embedding capacity on the cover text because most of the modification's corresponding cover texts comprise the conversion of original character forms to corresponding contextual representations, and whitespace Unicode to Medium Mathematical Space. Although visible and invisible additional characters (for example, ZWJs, Kashida, and ZWNJs) are used, the likelihood that they are used is much lower in this method than in other current works.

Two techniques for text steganography are proposed in [49]: using only pseudo-spaces and merging them with Kashida (extension character). The proposed approaches improved the attribute of data hiding with regard to security and capacity. The initial proposition referred to as Kashida-PS used Unicode (0640) between the letters where Kashida can be inserted, whereas PS Unicode (200C) is used between letters that cannot have Kashida. Kashida is concatenated seamlessly between two joined letters without separation. The PS character is not visible on prints; however, it is separated when inserted between two connected letters. PS is concatenated with letters or spaces in addition to dotted characters. The reader cannot detect any alteration between the actual and stego texts. Moreover, this sequence employs the NS and PS Unicode characters to provide the optimal capacity for Arabic steganography. The other suggested technique, PS-bet Words, uses NS and PS in sequence to embed a set of secret bits. This technique utilizes spaces within words for hiding procedures. The goal is to improve security. Furthermore, it can be executed in any language. Experimental outcomes indicated that the recommended algorithms attained a high-capacity ratio compared with other approaches.

Models by [50] suggested masking Arabic text using steganography based on the counting secret sharing technique. The research emphasized the integration of steganography with a counting-based secret sharing to memorize data.

These two techniques were formulated to hide secret information using Kashida features. The authors validated the steganography model, presupposing that the initial model served as a reference. Furthermore, the new enhancement margining strategy of including Kashida in specific areas was done with the intent of producing higher ambiguity to enhance the security level.

A technique that employed an integration of the zero-width character of the Unicode format and the zero-width joiner of Arabic was proposed by [51]. The objective is to address the challenges of the structure pertaining to the cover text and enhance the stego text file size. The experimental results show that hidden data capacity per word is significantly increased and the high visual similarity between both the cover and stego text can reduce the attention of intruders. The suggested technique indicated a significant increase in the capacity for a specified sample used by other researchers.

In similar work suggested by [52], a zero-width character (ZWC) algorithm-based technique was employed to hide secret data in social media messages. Furthermore, symmetric key techniques used in addition to mathematical coding techniques offer enhanced resilience against attacks.

Two systematic steganography-based techniques were suggested in [53] to hide information for English and Indian languages. The English content of the secret message was encoded using the shapes of the English alphabet's capital letters. The technique splits the letters based on their shape (e.g., vertical, horizontal, and curved). Subsequently, the three initial letters comprising the cover text are hidden, by counting the bits that have value of "1". The count value is the key value for hiding the data. After calculating the key value, the message is randomized by placing one character of the message after the key number character of the cover text. For Indian languages, the letters are split into eight sets based on pronunciation. The bits corresponding to covert text letters are used to hide the secret message within the cover text. Such techniques offer high data hiding capacity; however, the concept of using random sequences to hide valuable information is conspicuous. Therefore, it is required that the security of such methods be enhanced.

A text steganography technique based on text format was suggested in [54]. This method enhances hidden data storage using the justified formatted text contained inside PDF documents. This process initially uses Huffman coding (HC) to process the message. Next, specific lines from the cover text are designated as host lines. The embedding process comprises the spaces contained in the host lines to replace the inserted spaces. This process uses a key for enhancing communication security. This technique hides more information inside cover text compared with other techniques. Additionally, the cover file size remains the same, thereby suspicion from being raised. Moreover, because text originality is preserved, there is no chance of syntactical or grammatical errors.

A format-specific steganography technique was proposed by [55] that hides secret information by embedding it in cover documents (e.g., Microsoft (MS) Word format). This technique hides secret information bits by transforming the bits into whitespaces because these are abundantly present in most documents. In addition, these characters remain invisible if the font type and style are changed. The white spaces between words are manipulated to hide the secret data in terms of their font type and style. Thus, there is no perceptible difference in the stego file even after the secret data is hidden. The result showed that this method successfully improved the capacity of embedding data inside a cover file.

A new PDF steganography technique was formulated in [56] using the Chinese Remainder Theorem and ASCII code A0, representing the non-breaking space character. This character is invisible when embedded inside the text and viewed using most software required for viewing PDF documents. The researcher demonstrated four methods to enhance the data amount that can be hidden using a PDF cover file, while the A0 characters used between words can be minimized. Therefore, the difference in size between the stego file and the cover file is minimized. The results showed that this method was successful in improving the capacity of embedding data inside a cover file.

6.2. Linguistic

This technique uses linguistic steganography for hiding secret information inside text files. In [57], a linguistic steganography technique was proposed that is a topic-aware neural-linguistic steganography method. It can generate a steganographic paragraph with a specific topic based on knowledge graphs (KGs). A KG provides data about relevant topics and content to generate coherent multi-sentence texts for better concealment. The proposed method provides the quality of the generated steganographic text and its relevance to a specific topic.

The author in [58] proposed a method that focuses on addressing the inability to control the semantic expression in text steganographic generated by neural networks. The author addressed control cognitive imperceptibility as a new challenge, which the steganography models must attempt to overcome in the future. The author compared three encoder models for semantic extraction, namely the Gated Recurrent Unit (GRU) model, the Transformer model, and the Topic-Aware model. Categorical sampling generates steganographic sentences to construct the candidate pool that samples the words according to the overall conditional probability distribution. Experimental results show that the proposed method can additionally constrain the semantic expression of the generated steganographic text.

A linguistic steganographic method was proposed in [59] that can automatically generate the stego text based on an adaptive probability distribution and a generative adversarial network. The proposed method efficiently addressed the exposure bias produced due to the discrepancy between training and inference stages. Furthermore, the proposed method determines the candidate word space and embedding capacity related to the similarity of word probability to reduce the resulting deviation. The experiments showed the proposed method outperforms the previous models, particularly in anti-steganalysis ability, and offers a promising means to enhance steganographic security.

Ref. [60] suggested a secure generative linguistic steganographic method, which recursively embeds secret information using Adaptive Dynamic Grouping (ADG), based on the capability of an off-the-shelf language tool. The proposed method focused on improving the imperceptibility of embedded data by adaptively grouping the tokens according to their possibility at each step, thus dynamically embedding secret information in the generated stego text. Experiments showed that the proposed method has high security and the capability of generating fluent stego text.

Another linguistic steganographic method was proposed in [61], which can automatically generate the stego text based on Variational Auto-Encoder (VAE-Stega). This method focused on improving the imperceptibility and security of generated steganographic texts. The encoder in VAE-Stega is used for two primary purposes: to learn the statistical distribution features of large-size regular texts and generate steganographic sentences. Experimental results showed that the proposed method improves the imperceptibility of the generated steganographic sentences.

The author in [62] proposed a generative text steganographic method based on a long short-term memory (LSTM) network based on a large-scale regular text database to construct a language model. The word generated is tested based on the conditional possibility distribution (of words) intended by the LSTM network and the secret value embedded on the receiver side. An identical model is used at the receiver side to recover the secret data. Results showed that it outperforms similar works.

Recurrent Neural Networks (RNN-Stega) was proposed in [63], creating text covers automatically using a discrete bitstream. Fixed- and variable-length coding (FLC and VLC) are employed to analyze the conditional distribution of words used for encoding. The experiments indicated high embedding capacity and high security against malicious attempts.

A novel method was suggested in [64] to enhance the embedding capacity pertaining to linguistic steganography by employing synonym replacement. The change-tracking method was revised to hide the message inside MS Word files by utilizing HC.

To ensure that a third party remains unaware of the presence of a message, frequently used synonyms often hide information and prevent suspicion.

Another similar work suggested by [65] uses sampling to produce language with enhanced hiding capacity. Information is encoded in the cover text using arithmetic coding (AC). The results are benchmarked against FLC and VLC formulated for hiding information by creating text using deterministic techniques. Moreover, Kullback–Leibler divergence (KL) determines the difference between probability distributions of a specific variable x used to regulate FLC embedding. Furthermore, divergence- and temperature-based techniques are used for regulating VLC and AC, respectively. The AC technique's embedding performance was experimentally benchmarked against VLC and FLC and found to be superior.

Linguistic text steganography was suggested in [66], where the emphasis is on semantic text-based steganography. Information is concealed by substituting it with synonyms to address security challenges in credit letters issued by banks. Because such letters of credit (LC) can be used online, the importance of such security measures is increased.

Another work in the linguistic text steganography technique was proposed in [67], where Arabic symbols hide data. Proverbs and other figures of speech are used as datasets. A specific font (Naksh) is used along with the Aho–Corasick algorithm (AC*). The security and coverage aspects of this technique were assessed. It was found that several shapes used in Arabic can be used to fulfil steganographic requirements.

Ref. [68] suggested a new linguistic text-based steganography technique where the presence of frequently used letter sets or double letter pairs is searched for in the words of the paragraph, where the sentences are in the normalized form. The algorithm identifies the sentences to output its summary, which is also referred to as cover text. This process identifies the aspects of the letters of the source language and uses a classification using the extracted features. The English alphabet is classified using three representations. LC, LS, and LSS represent the sets having letters with slanting lines, curved lines, or standing and sleeping lines. The number of members representing each set is used to determine the secret bit and its representation.

A work proposed in [69] is another form of linguistic steganography known as the word-indexing compression technique (WIC), which can minimize the length of the experimental embedded payload. In contrast, an optimal stego content with high undetectability can be chosen from candidates via a selection strategy of stego text. The WIC technique compresses the hidden message by integrating a minimum–maximum weight algorithm with HC with the aid of the candidate cover text. This technique also attempts to enhance the ability of anti-steganalysis; 10 cover texts with small compression ratios are chosen from a huge cover text set and are embedded in the matching compressed hidden message through synonym substitutions. Only a single stego text is chosen by a given principle derived using the distance between a cover text and its stego text. Experimental results showed that the recommended compression algorithm attains better compression ratios than HC and LZW coding techniques, resulting in a higher embedding capacity, and performs better in anti-steganalysis with proposed compression and the stego text selection rule.

A linguistic text steganography technique was proposed in [70] to transform the multivariate hidden message into alphabets using the alphabetic transformation technique. This technique resolves the issue of the selection of cover messages. The proposed technique is a blind embedding technique that substitutes a character of the hidden message with one from the cover message and thus makes it zero-perceptible. Results showed that the proposed technique performance is improved with respect to information hiding capacity.

A multi-keyword carrier-free text steganography algorithm was proposed in [71] based on tagging of the POS. The concealed tags are chosen from each Chinese character element of the words. The POS is used to hide the number of keywords. Furthermore, each keyword in the sentence is labelled with a suitable POS tag. The mapping set between POS and numbers is acquired by counting the POS of words after segmentation. The POS is used to conceal the number of keywords hidden in each stego text, thus enhancing the hiding capacity because hidden tags are chosen from all Chinese character elements of the word.

6.3. Random and Statistical Generation

The statistical characteristics of a language are obtained and then employed to generate cover text. The work in [72] proposed a statistical text steganography technique based on the Markov Chain model, which focuses on transition probability, one of the most significant ideas of this model. This technique developed state transition-binary sequence illustrations based on the ideas and used them to regulate the production of new texts with embedded information. This technique uses the transition probability in the steganographic text generation process. This technique also encodes the state transition-binary sequence diagram required by the receiver to obtain the information, which further improves the security of the steganography information. The results showed that this model had greater hiding capacity than previous techniques.

Another technique was developed by [73] to increase the embedding capacity in the statistical steganography method by implementing a Markov Chain (MC) encoder/decoder integrated with HC for steganography of Arabic text. A lower bound and an upper bound are calculated for the stego text length, which is based on the parameters of the designed encoder/decoder. The capacity performance of the recommended Arabic steganography method was examined for hidden message length and various encoder parameters. The outcome showed that the application of specific constraints on MC increased the embedding capacity until it reached a maximum value.

Ref. [74] recommended a coverless statistic steganography method that is based on the MC model. This method is aimed at producing steganographic text that more closely resembles the training text. As per the value and characteristic of the model's transition probability, the technique employs maximum variable bit embedding rather than standard fixed bit embedding. This technique does not insert fixed bits of hidden data into each word of the content but retains the attributes of the training model to a greater extent. After integrating it with the transition probability, this model had greater hidden capacity outcomes than those of other techniques.

A coverless statistical steganography technique was presented in [75] using single bit rules based on the MC model. The hidden information is made to pass through the model in this technique, and the corresponding steganographic contents are later generated. In the text generation process, the significance of transition probability is stressed and used to the highest extent, with the aim of producing steganographic text, which more closely resembled the training text. The method employs maximum variable bit embedding rather than standard fixed bit embedding. This technique does not embed one bit of hidden information in each word of the text. The experimental outcomes of the proposed method showed that it had good hiding capability. However, it was not greater than the variable bit embedding technique used in the preceding technique.

A statistical text steganography technique was proposed in [76], which can automatically produce steganographic text based on the HC and MC model. It can automatically produce fluent text carriers with respect to secret information that is required to be embedded. The recommended method can learn from several samples obtained from the public and develops a good approximation of the statistical linguistic model. Experimental outcomes show that the recommended model performance is better than all preceding related techniques in terms of information hiding capacity.

A model that improved the capacity and security was presented in [77] using statistical text steganography by reducing the cost resulting from synonym substitution. Research demonstrated that synonym substitution modifies statistical attributes of the cover text wherein the cost function was described as a hybrid of statistical distortion and linguistical distortion (HSL).

Moreover, the model was designed to reduce synonym substitution distortion. The statistical distortion depends on the frequency of words. Syndrome-trellis code is used to minimize the embedding effect on the cover text. The experiments demonstrate that HSL can attain a more secure level and presents more substantial performance than conventional synonym substitution techniques.

Ref. [78] suggested a statistical text steganography technique that maximizes the hidden message capacity. For this purpose, the hidden message bits are reduced by producing metadata and inserting header information in the initial few bytes of cover contents by mapping between the ASCII values of character sequences and their corresponding binary value. The subsequent phase deals with processing the hidden message and storing it in cover bits in an optimal manner, which can assess the character sequences of the hidden message.

A coverless plain text steganography method was proposed in [79], based on the parity of the stroke number of Chinese characters and expanded to further enhance the embedding rate. Using the concept of space mapping, this method develops a binary search tree (BST) based on texts from the Internet and searches the matching texts according to the hidden binary digit string (BDS) produced. The aim is to find a universal attribute of characters representing a binary digit (0 or 1) and develop a mapping function between the attributes of characters and a BDS. Furthermore, suitable texts are extracted from the internet for steganography based on this mapping function. The Parity of Chinese Characters' Stroke Number is taken as a character attribute to illustrate the concept. The benefit of this technique is that it does not produce additional information because it is a coverless technique with better hiding strategy, high scalability, and high embedding capacity.

A coverless text steganography technique based on the Markov model and the half frequency crossover principle was introduced in [80], which conforms to the statistical attributes of natural language. This technique combines various language rules with training text in the half frequency crossover rule based on the Markov model. Two distinct databases were employed to enhance the efficiency and capacity of embedding.

Another statistical coverless text steganography model was suggested in [81] based on multi-rule language techniques. It differs from the conventional steganographic text generation technique based on a single-rule model. The technique recommended in the paper attempts to alternately combine models with different language principles and attempts to extract more language attributes from the training text. To confirm the effectiveness of the technique, the word sequence constraints are increased, and the accuracy of the generated text is increased with respect to the semantics. Nonetheless, because of the strong high order language model constraints, uncertainty is created regarding the increase in the capacity level.

A randomized indexed word dictionary was presented in [82], in addition to a list of emails to increase security and hiding capacity. A forward email is used as a cover. The carbon copy field contains hidden data that are encrypted using a randomized index-based word dictionary. An arbitrary bitstream (temporary stego key) is produced using the system time and transmitted separately using public-key cryptography. This provisional stego key is utilized to randomize the index values of the dictionary words. Because this method excludes noise from the actual email body content, it is safe and secure against common attacks. Moreover, it adds a security layer by randomizing the word index values each time with an 18-bit key generated using the system time, inserted in the "date" field of the forward email format. The hiding capacity is increased compared with other relevant methods because of the unique means of the message of the hidden word.

Similar work was undertaken in [83], where the secret message is hidden in numerous email addresses generated through the email body. The proposed technique uses the lossless compression methods of LZW to compress the hidden message. Various stego keys are also utilized to improve the level of security using the proposed technique. It was shown that the proposed technique results in a substantial increase in hiding capacity for a general sample, and this approach was also used in another study.

Another study [84] recommended establishing security and accuracy in the text steganography process based on the RSA algorithm, which is used to encrypt the user data by generating imperfections in the layout or the appearance of characters contained in the text. The message goes through multilevel encryption and equivalent decryption using both private and public keys, thus becoming resilient to cyber threats and security violations. The receiver obtains the modified message in the sender's format. Contrary to contemporary methods of text steganography, the encoded message is relatively limited with respect to size because it relies on the hidden message size, and is not limited by the amount of information to be transmitted.

A statistical text steganography technique was recommended in [85] that uses a combination of cryptography and steganography to ensure data transmission is secure. The Data Encryption Standard (DES) is implemented for the suggested technique. The encrypted data is hidden by changing the position of the hiding bit, which depends on the frequency of letters in the cover text. After the implementation of these two mechanisms, the security of the outcome of the recommended algorithm was analyzed by subjecting the outcome to a range of cyber-attacks. The findings of the analysis indicate that the algorithm helps to achieve high data security.

In [86], an email-based text steganography system was used, alongside the compression by Huffman, which exhibits a reversible characteristic. The email IDs are classified into a set of email addresses that had four characters before the "@" symbol. This is shared between the mail sender and the mail recipient. The set of the second parts of email IDs, for example, gmail.com, yahoo.com, etc., is then hidden in the secret message. The extra added characters as per the requirements of secret data bits are also taken from the secret data. Therefore, the hiding capacity is increased without adding any overhead to the cover text.

In [87], a novel and simple approach for steganography is presented that involves transliteration in Bengali. The key concept behind this technique is to optimally explore the special characteristics of Bengali phonetic keyboard layouts that can enable the hiding of secret information in the form of bits. In one option, the bit "0" is represented, and in another option, the bit "1" is represented in a document without involving the risk of being understood by an intermediate reader or user. The result shows the capacity of the method is increased and adequate for a text steganography system with an exceptionally low risk of detection.

In [88], a technique that increases the security aspect of data sharing is recommended using steganography and encryption, which addresses security inadequacy with respect to data privacy for a secured data-sharing environment in the cloud. A random key generation algorithm is also introduced that utilizes random-sized public/master-secret key pairs developed by the data owner. The uniqueness of this approach is that data encryption is enabled by the data owner using the Feistel structure network, which uses the public key and the data index of an image using steganography before uploading the data in the cloud server. To generate the master-secret key, the aggregate decryption key (ADK) is primarily used for data sharing with other users. The ADK is transmitted from the sender to those who are interested in accessing the contents through email. Only after completing verification of ADK and authentication through fingerprint 3D scan will the original data be downloaded.

In [89], a statistics-based steganography method was suggested that carefully evaluates the characteristics of webpages on the Internet. This approach functions on a search-based text steganography model. The suggested model was developed based on a hypothesis that involves a significant amount of data on the Internet. The model enables the sender of the secret message to find a webpage that includes all of the information necessary to describe a secret message. As a result, the sender is not required to make modifications to the webpage as cover data. Hence, only the distribution of characters in the secret message in the cover webpage is evaluated, and no modifications are made to the cover webpage. Thus, a URL that includes position information is merely required to be sent to the extractor, which ensures that the hiding capacity is significantly enhanced and imperceptibility is guaranteed. One of the preconditions for this model to function is that such a webpage exists and can be found. The findings of the experiments showed that the recommended method offers a substantial embedding capacity and considerable imperceptibility.

The preceding works on text steganography and their corresponding description are summarized in Table 3. Table 3 indicates the text steganography categories used in each technique, in addition to the improvement factor.

Table 3. Related works on text steganography.

	Title Name	Category	Technique Description	Year	Improvement
1	Aitsteg: an innovative text steganography technique for hidden transmission of a text message via social media [52]	Format based	Innovative text steganography using the ZWC algorithm.	2020	Capacity
2	Improvement of "text steganography based on Unicode of characters in multilingual" by custom font with special properties [45]	Format based	Replaces the code of English symbols with other code that has the same glyph. Two bits are hidden at once based on the set of high-frequency letters SHFL in lower case letters.	2020	Capacity
3	Text steganography using character spacing after normalization [46]	Format based	Uses frequency modulation techniques, font attributes, and character spacing to embed secret data.	2020	Capacity
4	Two high-capacity text steganography schemes based on color coding [42]	Format based	Embeds the secret message in the cover text using colored based permutation and numeration systems.	2020	Capacity

Table 3. *Cont.*

	Title Name	Category	Technique Description	Year	Improvement
5	A high-capacity algorithm for information hiding in Arabic text [47]	Format based	Unicode Arabic extension character Kashida with three small space characters: thin space, hair space, and six-PRE-EM space and use of Kashida and to use small space characters instead of the normal space character to hide the secret message.	2020	Capacity
6	Inclusion of Unicode standard seamless characters to expand Arabic text Steganography for secure individual uses [48]	Format based	Unicode standard seamless characters within Arabic text for secure individual uses, Unicode of whitespaces that are visible and invisible extra characters (such as Kashida, ZWJS, and (ZWNJS) are also used to hide secret bits.	2020	Capacity
7	Aggrandize text security and hiding data through text steganography [53]	Format based	Hides the secret message according to the shape of capital alphabets letters.	2019	Capacity
8	A multi-layer Arabic text steganographic method based on letter shaping [43]	Format based	Uses Unicode to hide a bit in each letter by reshaping the letters according to its position of the word or standalone. The hiding process is undertaken using a multi-embedding layer, where each layer contains all words with the same tag identified using the part of speech (POS) tagger.	2019	Capacity
9	A new method for pdf steganography in justified texts [54]	Format based	Justified text in a pdf file, by compressing the secret message using Huffman coding, then choosing special lines from the cover as host lines and replacing the added spaces with the normal spaces of the host lines.	2019	Capacity
10	Enhancing Arabic text steganography for personal usage utilizing Pseudo-spaces [49]	Format based	Uses two methods of pseudo-spaces alone and combines them with Unicode to maximize the use of spaces.	2019	Capacity
11	Refining Arabic text stego-techniques for shares memorization of Counting-based secret sharing [50]	Format based	Combines the counting-based secret sharing with steganography for personal remembrance using Kashida steganography.	2019	Capacity
12	A high capacity and imperceptible text steganography using binary digit mapping on ASCII characters [40]	Format based	Maps secret message binary digit onto binary digit of cover text using ASCII characters, involving spaces, punctuation, and symbols.	2018	Capacity
13	Information hiding Arabic text steganography by using Unicode characters to hide secret data [51]	Format based	Uses a combination of Unicode character's zero-width-character and zero-width-joiner in the Arabic language.	2018	Capacity
14	An efficient, secure system of data in the cloud using steganography-based Cryptosystem with FSN [88]	Format based	Combines both steganography and cryptography, and aggregate decryption key (ADK) to generate the master-secret key, where it is used to share data to other users by transferring its ADK to those who are interested in accessing the contents through e-mail by the data owner.	2018	Security
15	A high-capacity text steganography scheme based on LZW compression and color coding [41]	Format based	Employs the LZW compression technique and color coding-based approach. The technique uses the forward mail platform to hide secret data.	2017	Security
16	A space-based reversible high-capacity text steganography scheme using font type and style [55]	Format based	Hides the secret data bits into the white spaces in MS word documents.	2016	Capacity
17	Information hiding using whitespace technique in Microsoft word [44]	Format based	Uses five whitespaces within a line to represent each character for embedding secret data, which is randomized based on a key in MS word.	2016	Capacity
18	Topic-aware Neural Linguistic Steganography Based on Knowledge Graphs [57]	Linguistic	Neural linguistic steganography generates a steganographic paragraph with a specific topic based on knowledge graphs.	2021	Security

Table 3. *Cont.*

	Title Name	Category	Technique Description	Year	Improvement
19	Linguistic Generative Steganography with Enhanced Cognitive-Imperceptibility [58]	Linguistic	Uses categorical sampling to construct the candidate pool that samples the words according to the overall conditional probability distribution, to construct candidate pools in steganographic text generated by a neural network.	2021	Security
20	Linguistic Steganography Based on Adaptive Probability Distribution [59]	Linguistic	Automatically generated stego text based on adaptive probability distribution and generative adversarial network.	2021	Security
21	Provably Secure Generative Linguistic Steganography [60]	Linguistic	Improved embedded data imperceptibility by groups the tokens adaptively according to their possibility at each time step to embed secret information dynamically in stego text generated by a neural network.	2021	Security
22	VAE-Stega: Linguistic Steganography Based on Variational Auto-Encoder [61]	Linguistic	Automatically generate the stego text based on Variational Auto-Encoder (VAE-Stega) that learns the statistical distribution features of a huge number of regular texts to generate steganographic sentences.		Security
23	Generative Text Steganography Based on LSTM Network and Attention Mechanism with Keywords [62].	Linguistic	Generative text steganographic method based on long short-term memory (LSTM) network and uses a mechanism based on a large-scale regular text database to construct a language model.		Security
24	RNN-Stega: linguistic steganography based on recurrent neural networks [63]	Linguistic	Automatically generated text covers secret bitstream based on recurrent neural networks (RNN).	2020	Capacity
25	A modified approach to data hiding in Microsoft word documents by Change-tracking technique [64]	Linguistic	Synonym substitution by redesigning the change tracking technique for hiding the secret message in Microsoft word document with Huffman's codes.	2020	Capacity
26	Linguistic steganography by sampling-based Language generation [65]	Linguistic	Utilizes sampling-based language generation to improve the hiding rate. The arithmetic coding (AC) algorithm is adopted to embed messages in the cover text. Its performance is compared with fixed-length coding (FLC) and variable-length coding (VLC), which were designed for embedding messages during deterministic text generation.	2019	Capacity
27	Text steganography in the letter of credit (lc) using synonym Substitution based algorithm [66]	Linguistic	Replaces words with their synonyms to solve security risks in a letter of credit (LC) used in banking. LC information is sent online.	2019	Capacity
28	A linguistic steganography framework using Arabic calligraphy [67].	Linguistic	Uses Arabic calligraphy to hide information using Arabic poetry and proverbs as datasets with one shape of Arabic letters (Naskh font) and a modified Aho–Corasick algorithm (ac*).	2019	Capacity
29	A generalized model of text steganography by summary generation using frequency analysis [68]	Linguistic	Check for common letter pairs or double letter pairs in keywords in the paragraphs to find sentences that generate a possible summary as a cover text.	2018	Capacity
30	Linguistic steganography based on word indexing compression and candidate selection [69]	Linguistic	Establishes word indexing compression algorithm (WIC) that can reduce the length of the practical embedded payload. The best stego text with high undetectability is selected from candidates using selection strategy and compressed the secret message by combining a minimum-maximum weight algorithm with Huffman coding.	2018	Capacity

Table 3. *Cont.*

	Title Name	Category	Technique Description	Year	Improvement
31	Multilayer partially homomorphic encryption text steganography (Mlphe-ts): a zero-steganography approach [70]	Linguistic	Replaces a character of the secret message with a character of the cover message that converts the multi-variate secret message into alphabets through the alphabetic transformation process that resolves the problem of cover message selection.	2018	Capacity
32	Multi-keywords carrier-free text steganography based on the part of speech tagging [71]	Linguistic	Multi-keywords carrier-free text steganography based on the part of speech tagging. The hidden tags are selected from Chinese character components of words.	2017	Capacity
33	Stbs-Stega: coverless text steganography based on state transition-binary sequence [72]	Random and Statistical	Focuses on transition probability based on the Markov chain model, to create a state transition-binary sequence diagrams based on the concepts and used them to guide the generation of new texts with embedded secret information.	2020	Capacity
34	Enhanced least significant bit replacement algorithm in the spatial domain of steganography using character sequence optimization [78]	Random and statistical	Maps character sequences of ASCII values and their equivalent binary value while hiding secret data.	2020	Capacity
35	A forward email-based high-capacity text steganography technique using a randomized and indexed word dictionary [82]	Random and statistical	Uses a list of email addresses and forward email platform as a cover to increase the hiding capacity. Email addresses in the carbon copy (cc) field contain secret data that are encoded using a randomized index-based word dictionary.	2020	Capacity
36	Novel approaches to text steganography [84]	Random and statistical	Uses RSA algorithm to encrypt user data by generating subtle imperfections in the appearance of the characters included in the memo using both the public and private keys, thus making it robust to cyber-attacks and security breaches.	2019	Security
37	Coverless plain text steganography based on character features [79]	Random and statistical	Coverless steganography based on parity of Chinese characters' stroke.	2019	Capacity
38	Coverless text steganography based on half frequency crossover rule [80]	Random and statistical	Coverless using Markov model and the half frequency crossover rule, which accords with the statistical characteristics of natural language.	2019	Capacity
39	Research on coverless text steganography based on multi-rule language models alternation [81]	Random and statistical	Coverless based on multi-rule language models that combines models under different language rules alternately and tries to extract more language features from the training text.	2019	Capacity
40	Capacity investigation of Markov chain-based statistical text steganography: Arabic language case [73]	Random and statistical	Implements a Markov chain combined with Huffman coding for Arabic text and computed an upper bound and a lower bound for the stego-text length that depends on the designed encoder/decoder parameters.	2019	Capacity
41	Coverless text steganography based on maximum variable Bit embedding rules [74]	Random and statistical	Uses the maximum variable bit embedding instead of the usual fixed bit embedding by the Markov chain model to generate steganographic text closer to the existing text, according to the characteristic and value of transition probability in the model.	2019	Capacity
42	Research on coverless text steganography based on single bit rules [75]	Random and statistical	Uses the single variable bit embedding instead of the usually fixed bit embedding by the Markov chain model to generate steganographic text closer to the existing text according to the model's character and value of transition probability.	2019	Capacity
43	A novel steganography method using transliteration of Bengali text [87]	Random and statistic	Transliteration in the Bengali language by exploiting the special feature of Bengali phonetic keyboard layouts to hide secret information in the form of bits.	2019	Capacity

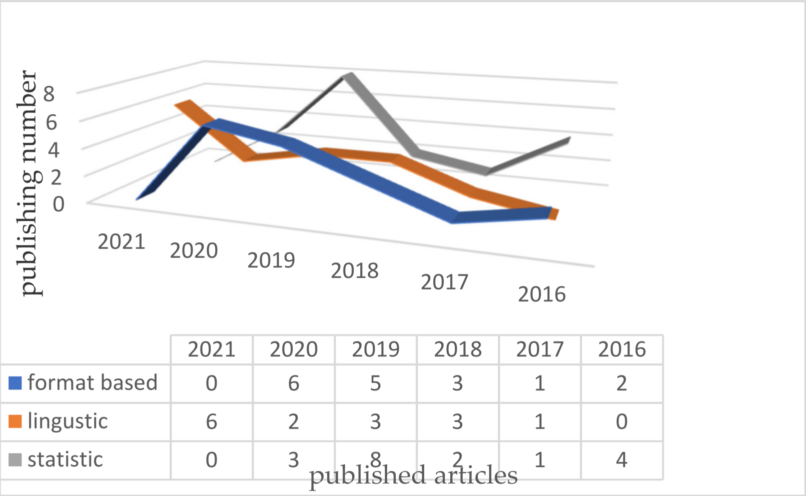
Table 3. *Cont.*

	Title Name	Category	Technique Description	Year	Improvement
44	Email-based high-capacity text steganography using repeating characters [83]	Random and statistical	Uses email in which the secret message is hidden within several email addresses generated through the body of the email and use one of the lossless compression algorithms named Lempel-Ziv-Welch (LZW) to compress the secret message.	2018	Capacity
45	Automatically generate steganographic text based on Markov model and Huffman coding [76]	Random and statistical	Automatically generates confident text carrier in terms of secret information which need to be embedded based on the Markov chain model and Huffman coding that can learn from many samples written by people and obtain a good estimation of the statistical language model.	2018	Capacity
46	Adaptive text steganography by exploring statistical and linguistical distortion [77]	Random and statistical	Minimizes the cost caused by synonym substitution, which affects the cover texts' statistical features, and minimize the distortion of synonym substitution.	2017	Capacity
47	Information hiding through dynamic text steganography and cryptography [85]	Random and statistical	Combination of steganography and cryptography using Data Encryption Standard (DES) to change the position of the hiding bit depending on the frequency of letters in the cover text.	2016	Security
48	A high-capacity email-based text steganography scheme using Huffman compression [86]	Random and statistical	Hides the secret data into the email IDs of a forward mail platform by Huffman compression, which is reversible in nature.	2016	Capacity
49	An approach to text steganography based on search on the internet [89]	Random and statistical	Analyzes the features of webpages on the internet. A search-based text steganography with a hypothesis that features of huge amount of data on the internet can create secret messages.	2016	Capacity
50	Pdf steganography based on Chinese Remainder theorem [56]	Random and statistical	Increases the amount of information that can be hidden in a cover pdf file based on the Chinese remainder theorem. While reducing the number of a 0's insertions considerably in between-character locations in that file, thus reducing the weight difference between a cover file and a stego file in which a secret message is embedded.	2016	Capacity

7. Discussion and Future Directions

Text steganography is the most difficult form of steganography due to the unavailability of redundant bits, compared with other file types, such as image, video, and audio files. The structure of the text is identical to what is observed by the human eye, whereas the structure of other files, such as images and videos, is different from what has been observed. Thus, it is relatively easy to hide information in later documents because no changes are observed compared to text. In contrast, if slight changes are made to a text document, they can be easily detected by the human eye. Figure 7 shows a count of the different text steganography methods developed to hide secret messages published from 2016 to 2021. The ultimate challenge in text steganography is the low hiding capacity, which is usually caused by insufficient redundant data in textual documents compared with other digital media, such as image, audio, and video files.

In the format-based methods shown in Table 3, many factors were required to be considered by the authors when the various techniques were applied to hide text. The first factor is the language that is used as a cover text. The language characteristics help to detect the means for hiding data inside the text cover, such as font attributes and shapes of letters for English and Indian languages. By comparison, the Arabic language uses points, word diacritics, and Kashida.



	2021	2020	2019	2018	2017	2016
🟦 format based	0	6	5	3	1	2
🟧 lingustic	6	2	3	3	1	0
◼ statistic	0	3	8	2	1	4

Figure 7. Published articles in different years (2016–2021).

For the English language, as seen in [40,44–46,53,55], various language attributes are used, such as the font, capital and lower-case letters, double letters, high-frequency letters, punctuation, and other symbols. Secret messages are also hidden according to the shape of the English capital alphabets by dividing the letters into groups depending on the letter's shape (curved, vertical, horizontal, etc.). These attributes are used to hide secret bits in the English language. Spaces are also used, via Unicode characters, to hide secret bits, as seen in [52].

Arabic, which consists of "points" or "dots", is also used in text steganography, as seen in [43,47–51,67]. Some alphabets contain points and word diacritics. A diacritic refers to all the markings that can appear above and below letters to alter their pronunciation. Thus, authors have used Kashida because it can be applied easily without affecting the integrity of the document, and Unicode characters (zero-width character and zero-width joiner) can be used to maximize the use of the spaces in points or between letters and signs, or the shape of Arabic letters.

Regarding the Indian language, as used in [53], the Indian letters are divided into eight groups based on how they are pronounced (based on the articulation of letters), and then the secret message bits are hidden inside the cover text letter based on the specific positions of the cover text letter bits, according to the shape of the alphabets. (The shapes can be divided into curves, vertical, horizontal, etc.).

The cover text characteristics can also be part of the steganography technique. For example, a text file is used as cover media in [42], which uses the spaces between letters and blank spaces in the file. The text color is also used to hide secret text. For example, in an MS Word file, the use of the font and capital or lower-case letters, or changing the colors of the letters and creating more space in the file, may be used as a pattern to hide secret texts. PDF files can also be used for text steganography, as seen in the approaches in [54,56], by removing the ragged edges of the text in a process named "Justify", which is reused during the message hiding process.

Regarding the linguistic method, numerous techniques are used to obtain the best match. Linguistic integrity, syntax, and semantics of a language are employed to accomplish this goal. In addition, as a result of the massive amounts of available data, deep learning (DL) has become increasingly popular and is widely used for numerous applications. In recent years, DL has been applied to linguistic steganography, particularly text steganography [60]. However, this poses a new challenge in terms of the security of the proposed methods. This is because the DL itself offers auto-generation of text-linguistic steganography while maintaining the semantics of the text. This is one of the most important features in successful linguistic steganography. For this purpose, ref. [57] used a knowledge graph that provides data about relevant topics and contents to generate coherent multi-sentence texts and thus improve data hiding.

Ref. [58] attempted to address the control of semantic expression in steganographic texts generated by neural networks. Moreover, [59] generated stego texts based on the adaptive probability distribution and the generative adversarial network, which focused on eliminating the exposure bias produced due to the discrepancy between training and inference stages. In [60], Adaptive Dynamic Grouping (ADG) was applied by recursively embedding secret information using an off-the-shelf language tool. Statistical distribution features were used in [61], namely Variational Auto-Encoder (VAE-Stega), using numerous regular texts to generate steganographic sentences. A long short-term memory (LSTM) network was used in [62] to generate the text. The text was then tested based on the conditional possibility distribution (of words) intended by the LSTM network and the secret value.

In [63], text covers are generated based on an RNN, which can automatically generate high-quality texts based on a secret bitstream that needs to be hidden.

The semantic synonym substitution method was used in [64] by redesigning the change-tracking technique to hide the secret message. The AC algorithm was adopted in [65] to embed messages in the cover text in a syntax manner. The performance was compared with FLC and VLC, which were designed for embedding messages during deterministic text generation. Ref. [66] focused on semantic-based text steganography to hide secret information by replacing words with their synonyms. In [67], Arabic calligraphy was used to hide information using Arabic poetry and proverbs as datasets, with one shape of Arabic letters (Naskh font) and a modified AC*. Ref. [69] used WIC, which can reduce the length of the practical embedded payload by selecting the best stego text with high undetectability from candidates using a stego text selection strategy. Ref. [70] used conversion of the multi-variate secret message into alphabets, which resolves the problem of cover message selection. Ref. [71] used a method based on POS tagging. The hidden tags are selected from all the Chinese character components of words.

The last technique is the random and statistical method. Among the statistical methods, numerous techniques applied the MC model, and focus on the transition probability to create a state transition-binary sequence diagram. This approach was used to guide the generation of new texts with embedded secret information, such as in [72–76,81]. Ref. [77] used statistical text steganography to minimize the cost of synonym substitution and reduce the impact of embedding on the cover text. Ref. [78] generated the metadata and embedded the header information in the first few bites of the cover media by mapping between the character sequences of the ASCII values and their equivalent binary value, minimizing the secret message bits while increasing the capacity of the hidden message. Ref. [79] exploited the space mapping concept to initialize a BST based on texts from the Internet and searched the corresponding texts according to the secret BDS generated based on the parity of the stroke number of Chinese characters, and extended it to further improve the embedding rate. Ref. [89] also analyzed the features of webpages on the Internet and a proposed model for search-based text steganography. The model is based on a hypothesis that features a huge amount of data on the Internet, and requires the sender of the secret message to find a specific webpage that contains all the information to describe a secret message.

Ref. [87] applied a technique through transliteration in the Bengali language by exploiting the special feature of the Bengali phonetic keyboard layouts to hide secret information in the form of bits.

Regarding random methods, ref. [82] applied a randomized indexed word dictionary and a list of email addresses to increase the hiding capacity and security. A forward email platform can also be used as cover media, such as in [83,86], and hides the secret bits within several email addresses generated through the body of the email. Refs. [83,86] also used a compression technique to compress the secret message.

All of the techniques described previously were applied to improve the capacity rate of a cover text. Some authors also tried improving security, such as in [84]. The security process relies on the RSA algorithm to encrypt user data by generating subtle imperfections in the appearance or layout of the characters. In [85], the authors combined both steganography and cryptography for secure data transmission. The proposed method used DES, a symmetric key algorithm for cryptography. Ref. [88] also attempted to improve security in data sharing in the cloud via a random key generation algorithm that used a random-sized public/master-secret key pair created by the data owner. Less attention has been paid by researchers to robustness because there is a presumption that stego files can be dispatched through the Internet and arrive at the recipient unmodified. Thus, active attack situations that may cause distortions to the stego file do not occur.

The following are some of the aspects that can be considered for future directions:

- A hybrid method from the classical approach and deep learning approaches;
- Semantic control through auto text generation by deep learning, particularly for long texts.
- Spatial format-based steganography using Unicode characters for data hiding with minimal stego file size.
- Few researchers have utilized compression to increase the effectiveness of their techniques because it decreases the amount of hidden information. This could be further examined.
- As a performance metric, several approaches have investigated hiding capacity, security, and robustness. However, there is a risk of attack if data is sent over untrusted channels. In addition to other metrics, the effectiveness of a well-designed algorithm against various attacks may be evaluated.
- Combining encryption and steganography techniques to provide an additional layer of security for the embedding algorithm. These combined methods can be studied further.
- Sequential selection of embedding positions makes the algorithm vulnerable to attack. Therefore, an extra security layer for embedding techniques that offers non-sequence or random embedding positions can be explored.

8. Conclusions

This study provides an extensive review of recent text steganography methods. It also provides a classification of steganography based on each method. The methods of text steganography are divided into three categories: format-based, random and statistical generation, and linguistics methods. Moreover, comparisons between these techniques are highlighted. This study shows that increasing the capacity factor in the format-based method and increasing security in linguistic steganography remain popular topics in text steganography. However, robustness in this field has been paid less attention by researchers. This study traced the development in the field of text steganography and showed the employment of new technologies in this area. By compiling the existing methods, these technologies were linked with the existing steganography categories to help researchers. Consequently, this paper indicates new research areas in text steganography by providing aspects that can be considered for future directions.

Author Contributions: M.A.M.: Collected, analyzed, and summarized all related materials regarding steganography techniques in several databases.; Z.S.: Provided an idea about the background and discussion of this article.; R.S.: Undertook a thorough revision of the structure and provided improvements of the manuscript for publication; M.K.H.: Undertook the final revision and provided an overall opinion on the contributions of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Malaysian Fundamental Research Grant Scheme (FRGS) under the code FRGS/1/2020/ICT01/UKM/02/4, and Universiti Kebangsaan Malaysia under the codes PP-FTSM-2021 and FTM1.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheddad, A.; Condell, J.; Curran, K.; Mc Kevitt, P. Digital image steganography: Survey and analysis of current methods. *Signal. Process.* **2010**, *90*, 727–752. [CrossRef]
2. Anderson, R.; Petitcolas, F. On the limits of steganography. *IEEE J. Sel. Areas Commun.* **1998**, *16*, 474–481. [CrossRef]
3. Srikumar, R.; Malarvizhi, C.S. Strong encryption using steganography and digital watermarking. In Proceedings of the 22nd Picture Coding Symposium, Seoul, Korea, 25–27 April 2001; pp. 425–428.
4. Al-Daweri, M.S.; Abdullah, S.; Ariffin, K.A.Z. A homogeneous ensemble based dynamic artificial neural network for solving the intrusion detection problem. *Int. J. Crit. Infrastruct. Prot.* **2021**, *34*, 100449. [CrossRef]
5. Majeed, M.A.; Sulaiman, R. An improved LSB image steganography technique using BIT-inverse in 24 BIT colour image. *J. Theor. Appl. Inf. Technol.* **2015**, *80*, 2.
6. Johnson, N.F.; Jajodia, S. Exploring steganography: Seeing the unseen. *Computer* **1998**, *31*, 26–34. [CrossRef]
7. Premaratne, P.; DeSilva, L.C.; Burnett, I. Low frequency component-based watermarking scheme using 2D data matrix. *Int. J. Inf. Technol.* **2006**, *12*, 1–12.
8. Le, T.H.N.; Nguyen, K.H.; Le, H.B. Literature survey on image watermarking tools, watermark attacks, and benchmarking tools. In Proceedings of the 2nd International Conference on Advance Multimedia, IEEE, Athens, Greece, 13–19 June 2010; pp. 67–73. [CrossRef]
9. Cox, I.J.; Miller, M.L.; Bloom, J.A.; Fridrich, J.; Kalker, T. *Digital Watermarking and Steganography*; Morgan Kaufmann: Burlington, MA, USA, 2008. [CrossRef]
10. Shih, F.Y. *Digital Watermarking and Steganography: Fundamentals and Techniques*; CRC Press: Boca Raton, FL, USA, 2017.
11. Al-Naqeeb, A.B.; Nordin, M.J. Robustness Watermarking Authentication Using Hybridisation DWT-DCT and DWT-SVD. *Pertanika J. Sci. Technol.* **2017**, *25*, 73–86.
12. Judge, J.C. *Steganography: Past, Present, Future*; Lawrence Livermore National Lab.: Livermore, CA, USA, 2001.
13. Kamil, S.; Ayob, M.; Abdullah, S.N.H.S.; Ahmad, Z. Challenges in multi-layer data security for video steganography revisited. *APIJTM* **2018**, *07*, 53–62. [CrossRef]
14. Stefan, K.; Fabien, A.P.P. *Information Hiding Techniques for Steganography and Digital Watermarking (Artech House Computer Security Series)*; Artech House: London, UK, 2000.
15. Mishra, M.; Mishra, P.; Adhikary, M.C. Digital image data hiding techniques: A comparative study. *arXiv* **2014**, arXiv:1408.3564.
16. Provos, N.; Honeyman, P. Hide and seek: An introduction to steganography. *IEEE Secur. Priv. Mag.* **2003**, *1*, 32–44. [CrossRef]
17. Petitcolas, F.A.P.; Anderson, R.J.; Kuhn, M. Information hiding-a survey. *Proc. IEEE* **1999**, *87*, 1062–1078. [CrossRef]
18. Du, J.-X.; Huang, D.-S.; Wang, X.-F.; Gu, X. Computer-aided plant species identification (CAPSI) based on leaf shape matching technique. *Trans. Inst. Meas. Control* **2006**, *28*, 275–285. [CrossRef]
19. Zheng, C.-H.; Huang, D.-S.; Sun, Z.-L.; Lyu, M.R.; Lok, T.-M. Nonnegative independent component analysis based on minimizing mutual information technique. *Neurocomputing* **2006**, *69*, 878–883. [CrossRef]
20. Bhattacharjya, A.K.; Ancin, H. Data embedding in text for a copier system. In Proceedings of the 2018 IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018; pp. 245–249.
21. Baawi, S.S.; Mokhtar, M.R.; Sulaiman, R. A comparative study on the advancement of text steganography techniques in digital media. *ARPN J. Eng. Appl. Sci.* **2018**, *13*, 1854–1863.
22. Awais, M.; Müller, H.; Tang, T.B.; Meriaudeau, F. Reversible data embedding in Golomb Rice code. In Proceedings of the 2011 IEEE Inter-national Conference on Signal and Image Processing Applications, Kuala Lumpur, Malaysia, 16–18 November 2011; pp. 515–519. [CrossRef]
23. Kadhim, I.J. A new audio steganography system based on auto-key generator. *AL-Khwarizmi Eng. J.* **2012**, *8*, 27–36.
24. Santhi, B.; Radhika, G.; Reka, S.R. Information security using audio steganography—A survey. *Res. J. Appl. Sci. Eng. Technol.* **2012**, *4*, 2255–2258.
25. Limkar, S.; Nemade, A.; Badgujar, A.; Kate, R. Improved Data Hiding Technique Based on Audio and Video Steganography. *Smart Comput. Inform.* **2017**, 581–588. [CrossRef]
26. Jeyasheeli, P.G.; Selva, J.J. A survey on DNA and image steganography. In Proceedings of the 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 January 2017. [CrossRef]
27. Haughton, D.; Balado, F. A modified watermark synchronisation code for robust embedding of data in DNA. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 1148–1152. [CrossRef]
28. Odeh, A.; Elleithy, K.; Faezipour, M.; Abdelfattah, E. Novel steganography over HTML code. In *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 607–611.
29. Memon, A.G.; Khawaja, S.; Shah, A. Steganography: A new horizon for safe communication through xml. *J. Theor. Appl. Inf. Technol.* **2008**, *4*, 187–202.
30. Zielińska, E.; Mazurczyk, W.; Szczypiorski, K. Trends in steganography. *Common. ACM.* **2014**, *57*, 86–95. [CrossRef]
31. Subhedar, M.S.; Mankar, V.H. Current status and key issues in image steganography: A survey. *Comput. Sci. Rev.* **2014**, *13–14*, 95–113. [CrossRef]
32. Li, B.; He, J.; Huang, J.; Shi, Y.Q. A survey on image steganography and steganalysis. *J. Inf. Hiding Multimed. Signal Process.* **2011**, *2*, 142–172.

33. Marvel, L.M.; Retter, C.T.; Boncelet, C.G. A methodology for data hiding using images. In Proceedings of the IEEE Military Communications Conference, Los Angeles, CA, USA, 19–21 October 1998; pp. 1044–1047.
34. Mathkour, H.; Al-Sadoon, B.; Touir, A. A new image steganography technique. In Proceedings of the 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 12–17 October 2008; pp. 1–4.
35. Altaay, A.A.J.; Sahib, S.B.; Zamani, M. An introduction to image steganography techniques. In Proceedings of the 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT), Kuala Lumpur, Malaysia, 26–28 November 2012; pp. 122–126.
36. Ramu, P.; Swaminathan, R. Imperceptibility—Robustness tradeoff studies for ECG steganography using continuous ant colony optimization. *Expert Syst. Appl.* **2016**, *49*, 123–135. [CrossRef]
37. Abraham, A.; Paprzycki, M. Significance of steganography on data security. In Proceedings of the ITCC 2004 International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, 5–7 April 2004; pp. 347–351.
38. Baawi, S.S.; Mokhtar, M.R.; Sulaiman, R. Enhancement of text steganography technique using lempel-ziv-welch algorithm and two-letter word technique. In Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018), Kuala Lumpur, Malaysia, 23–24 July 2018; pp. 525–537. [CrossRef]
39. Li, L.; Qian, J.; Pan, J.-S. Characteristic region based watermark embedding with RST invariance and high capacity. *AEU—Int. J. Electron. Commun.* **2011**, *65*, 435–442. [CrossRef]
40. Naharuddin, A.; Wibawa, A.D.; Sumpeno, S. A high capacity and imperceptible text steganography using binary digit mapping on ASCII characters. In Proceedings of the 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA), Bali, Indonesia, 30–31 August 2018; pp. 287–292. [CrossRef]
41. Malik, A.; Sikka, G.; Verma, H.K. A high capacity text steganography scheme based on LZW compression and color coding. *Eng. Sci. Technol. Int. J.* **2017**, *20*, 72–79. [CrossRef]
42. Sadié, J.K.; Metcheka, L.M.; Ndoundam, R. A high capacity text steganography scheme based on permutation and color coding. *arXiv* **2020**, arXiv:2004.00948.
43. Al-Azzawi, A.F. A multi-layer arabic text steganographic method based on letter shaping. *Int. J. Netw. Secur. Its Appl. (IJNSA)* **2019**, *11*. Available online: https://ssrn.com/abstract=3759471 (accessed on 29 September 2021).
44. Liang, O.W.; Iranmanesh, V. Information hiding using whitespace technique in Microsoft word. In Proceedings of the 2016 22nd International Conference on Virtual System & Multimedia (VSMM), Kuala Lumpur, Malaysia, 17–21 October 2016; pp. 1–5.
45. Baawi, S.S.; Nasrawi, D.A. Improvement of "text steganography based on unicode of characters in multi-lingual" by custom font with special properties. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Jonkoping, Sweden, 22–23 June 2020; Volume 870, p. 012125.
46. Shah, S.T.A.; Khan, A.; Hussain, A. Text steganography using character spacing after normalization. *Int. J. Sci. Eng. Res.* **2020**, *11*, 949–957. [CrossRef]
47. Taha, A.; Hammad, A.S.; Selim, M.M. A high capacity algorithm for information hiding in Arabic text. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *32*, 658–665. [CrossRef]
48. Alanazi, N.; Khan, E.; Gutub, A. Inclusion of unicode standard seamless characters to expand arabic text steganography for secure individual uses. *J. King Saud Univ. Comput. Inf. Sci.* **2020**. In press. [CrossRef]
49. Al-Nofaie, S.; Gutub, A.; Al-Ghamdi, M. Enhancing Arabic text steganography for personal usage utilizing pseudo-spaces. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**, *33*, 963–974. [CrossRef]
50. Gutub, A.A.-A.; Alaseri, K.A. Refining Arabic text stego-techniques for shares memorization of counting-based secret sharing. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**. [CrossRef]
51. Ditta, A.; Yongquan, C.; Azeem, M.; Rana, K.G.; Yu, H.; Memon, M.Q. Information hiding: Arabic text steganography by using Unicode characters to hide secret data. *Int. J. Electron. Secur. Digit. Forensics* **2018**, *10*, 61–78. [CrossRef]
52. Ahvanooey, M.T.; Li, Q.; Hou, J.; Mazraeh, H.D.; Zhang, J. AITSteg: An innovative text steganography technique for hidden transmission of text message via social media. *IEEE Access* **2018**, *6*, 65981–65995. [CrossRef]
53. Chaudhary, S.; Dave, M.; Sanghi, A. Aggrandize text security and hiding data through text steganography. In Proceedings of the 2016 IEEE 7th Power India International Conference (PIICON), Bikaner, India, 25–27 November 2016; pp. 1–5.
54. Khosravi, B.; Khosravi, B.; Khosravi, B.; Nazarkardeh, K. A new method for pdf steganography in justified texts. *J. Inf. Secur. Appl.* **2019**, *45*, 61–70. [CrossRef]
55. Kumar, R.; Malik, A.; Singh, S.; Kumar, B.; Chand, S. A space based reversible high capacity text steganography scheme using font type and style. In Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 29–30 April 2016; pp. 1090–1094. [CrossRef]
56. Ekodeck, S.G.R.; Ndoundam, R. PDF steganography based on Chinese Remainder Theorem. *J. Inf. Secur. Appl.* **2016**, *29*, 1–15. [CrossRef]
57. Li, Y.; Zhang, J.; Yang, Z.; Zhang, R. Topic-aware neural linguistic steganography based on knowledge graphs. *ACM/IMS Trans. Data Sci.* **2021**, *2*, 1–13. [CrossRef]
58. Yang, Z.; Xiang, L.; Zhang, S.; Sun, X.; Huang, Y. Linguistic generative steganography with enhanced cognitive-imperceptibility. *IEEE Signal. Process. Lett.* **2021**, *28*, 409–413. [CrossRef]
59. Zhou, X.; Peng, W.; Yang, B.; Wen, J.; Xue, Y.; Zhong, P. Linguistic steganography based on adaptive probability distribution. *IEEE Trans. Dependable Secur. Comput.* **2021**. [CrossRef]

60. Zhang, S.; Yang, Z.; Yang, J.; Huang, Y. Provably secure generative linguistic steganography. *arXiv* **2021**, arXiv:2106.02011.
61. Yang, Z.-L.; Zhang, S.-Y.; Hu, Y.-T.; Hu, Z.-W.; Huang, Y.-F. VAE-Stega: Linguistic steganography based on variational auto-encoder. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 880–895. [CrossRef]
62. Kang, H.; Wu, H.; Zhang, X. Generative text steganography based on LSTM network and attention mechanism with keywords. *Electron. Imaging* **2020**, *2020*, 291. [CrossRef]
63. Yang, Z.-L.; Guo, X.; Chen, Z.-M.; Huang, Y.-F.; Zhang, Y.-J. RNN-Stega: Linguistic steganography based on recurrent neural networks. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 1280–1295. [CrossRef]
64. Mahato, S.; Khan, D.A.; Yadav, D.K. A modified approach to data hiding in Microsoft Word documents by change-tracking technique. *J. King Saud Univ.-Comput. Inf. Sci.* **2020**, *32*, 216–224. [CrossRef]
65. Yang, R.; Ling, Z.H. Linguistic Steganography by Sampling-based Language Generation. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1014–1019.
66. Chaw, A.A. Text steganography in Letter of Credit (LC) using synonym substitution based algorithm. *Int. J. Adv. Res. Dev.* **2019**, *4*, 59–63.
67. Hamzah, A.A.; Khattab, S.; Bayomi, H. A linguistic steganography framework using Arabic calligraphy. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *33*, 865–877. [CrossRef]
68. Majumder, A.; Changder, S. A generalized model of text steganography by summary generation using frequency analysis. In Proceedings of the 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 29–31 August 2018; pp. 599–605. [CrossRef]
69. Xiang, L.; Wu, W.; Li, X.; Yang, C. A linguistic steganography based on word indexing compression and candidate selection. *Multimed. Tools Appl.* **2018**, *77*, 28969–28989. [CrossRef]
70. Naqvi, N.; Abbasi, A.T.; Hussain, R.; Khan, M.A.; Ahmad, B. Multilayer partially homomorphic encryption text steganography (Mlphe-ts): A zero-steganography approach. *Wirel. Pers. Commun.* **2018**, *103*, 1563–1585. [CrossRef]
71. Liu, Y.; Wu, J.; Xin, G. Multi-keywords carrier-free text steganography based on part of speech tagging. In Proceedings of the 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, China, 29–31 July 2017; pp. 2102–2107. [CrossRef]
72. Wu, N.; Yang, Z.; Yang, Y.; Li, L.; Shang, P.; Ma, W.; Liu, Z. STBS-Stega: Coverless text steganography based on state transition-binary sequence. *Int. J. Distrib. Sens. Netw.* **2020**, *16*. [CrossRef]
73. Alghamdi, N.; Berriche, L. Capacity investigation of Markov chain-based statistical text steganography: Arabic language case. In Proceedings of the 2019 Asia Pacific Information Technology Conference, Jeju Island, Korea, 25–27 January 2019; pp. 37–43.
74. Wu, N.; Shang, P.; Fan, J.; Yang, Z.; Ma, W.; Liu, Z. Coverless Text Steganography Based on Maximum Variable Bit Embedding Rules. *J. Phys. Conf. Ser.* **2019**, *1237*, 022078. [CrossRef]
75. Wu, N.; Shang, P.; Fan, J.; Yang, Z.; Ma, W.; Liu, Z. Research on coverless text steganography based on single bit rules. *J. Physics Conf. Ser.* **2019**, *1237*. [CrossRef]
76. Yang, Z.; Jin, S.; Huang, Y.; Zhang, Y.; Li, H. Automatically generate steganographic text based on Markov model and Huffman coding. *arXiv* **2018**, arXiv:1811.04720.
77. Huanhuan, H.; Xin, Z.; Weiming, Z.; Nenghai, Y. Adaptive text steganography by exploring statistical and linguistical distortion. In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26–29 June 2017; pp. 145–150.
78. Jayapandiyan, J.R.; Kavitha, C.; Sakthivel, K. Enhanced least significant bit replacement algorithm in spatial domain of steganography using character sequence optimization. *IEEE Access* **2020**, *8*, 136537–136545. [CrossRef]
79. Wang, K.; Gao, Q. A Coverless plain text steganography based on character features. *IEEE Access* **2019**, *7*, 95665–95676. [CrossRef]
80. Wu, N.; Ma, W.; Liu, Z.; Shang, P.; Yang, Z.; Fan, J. Coverless Text Steganography Based on Half Frequency Crossover Rule. In Proceedings of the 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Hohhot, China, 5–27 October 2019; pp. 726–7263. [CrossRef]
81. Wu, N.; Liu, Z.; Ma, W.; Shang, P.; Yang, Z.; Fan, J. Research on coverless text steganography based on multi-rule language models alternation. In Proceedings of the 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Hohhot, China, 5–27 October 2019; pp. 803–8033. [CrossRef]
82. Maji, G.; Mandal, S. A forward email based high capacity text steganography technique using a randomized and indexed word dictionary. *Multimedia Tools Appl.* **2020**, *79*, 26549–26569. [CrossRef]
83. Fateh, M.; Rezvani, M. An email-based high capacity text steganography using repeating characters. *Int. J. Comput. Appl.* **2021**, *43*, 226–232. [CrossRef]
84. Alanazi, N.; Khan, E.; Gutub, A. Efficient security and capacity techniques for Arabic text steganography via engaging Unicode standard encoding. *Multimed. Tools Appl.* **2020**, *80*, 1403–1431.
85. Bhat, D.; Krithi, V.; Manjunath, K.N.; Prabhu, S.; Renuka, A. Information hiding through dynamic text steganography and cryptography. *Comput. Inform.* **2017**, 1826–1831. [CrossRef]
86. Kumar, R.; Malik, A.; Singh, S.; Chand, S. A high capacity email based text steganography scheme using huffman compression. In Proceedings of the 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 11–12 February 2016; pp. 53–56.

87. Khairullah, M. A novel steganography method using transliteration of Bengali text. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**, *31*, 348–366. [CrossRef]
88. Shanthi, S.; Kannan, R.; Santhi, S. Efficient secure system of data in cloud using steganography based cryptosystem with FSN. *Mater. Today Proc.* **2018**, *5*, 1967–1973. [CrossRef]
89. Shi, S.; Qi, Y.; Huang, Y. An Approach to Text Steganography Based on Search in Internet. In Proceedings of the 2016 International Computer Symposium (ICS), Chiayi, Taiwan, 15–17 December 2016; pp. 227–232. [CrossRef]