A Novel Method for Embedding and Extracting Secret Messages in Textual Documents based on Paragraph Resizing

Benjamin Aziz[a], Aysha Bukhelli[b], Rinat Khusainov[c] and Alaa Mohasseb[d]
School of Computing, University of Portsmouth, Portsmouth PO1 3HE, United Kingdom
{benjamin.aziz, aysha.bukhelli, rinat.khusainov, alaa.mohasseb}@port.ac.uk

Keywords: Formal Methods, Information Hiding, Lexical Steganography, Text Steganography, Linguistic Steganography.

Abstract: The ancient technique of information hiding known as text steganography has enjoyed much research in recent years due to the rising popularity of social media platforms and the abundant availability of online literature and other text as cover media for steganography. Whilst the majority of the research approaches have focused on manipulating or replacing text, in some form or another, to embed secret information, the utilisation of the structure of the document itself for such embedding has rarely been researched. Therefore, we propose in this short paper a new approach for embedding secret messages in textual documents based on the splitting, merging, and resizing of paragraph text. The size comparison between adjacent paragraphs embeds one bit of information. We outline only the basic idea and define the syntax and semantics of the embedding language.

1 INTRODUCTION

Text steganography refers to all the techniques and methods used for hiding secret messages in textual documents (Agarwal, 2013; Lockwood and Curran, 2017; Taleby Ahvanooey et al., 2019; Kumar and Singh, 2020; Majeed et al., 2021). Unlike other media, text documents are more challenging to use as cover due to the lack of redundant information that can be used for hiding secret messages. As a result, the smallest manipulation of the text becomes immediately visible to the human eye. This further means that the original cover document cannot be assumed to be known to the reader, since differences with the modified version will be immediately detectable.

There are many approaches to text steganography. These include format-based, in which the physical features of text symbols are used to conceal a message (Xiang et al., 2007; Roy and Manasmita, 2011; Naharuddin et al., 2018; Malik et al., 2017), substitution-based, or lexical approaches, where words are substituted for others without affecting the meaning (Barmawi, 2016; Yajam et al., 2014), random statistical generation (Wu et al., 2020; Wu et al., 2019; Huanhuan et al., 2017), linguistic methods (Li et al., 2021; Yang et al., 2021b; Kang et al., 2020), as well as cover-less (Wu et al., 2018; Wang and Gao, 2019; Guan et al., 2022) and human-in-the-loop-based (Bergmair and Katzenbeisser, 2006; Grosvald and Orgun, 2012) approaches. The use of text-based steganography and steganalysis in social media (Shirali-Shahreza and Shirali-Shahreza, 2007; Wilson et al., 2015) in recent years has drawn much attention to these topics from the research community, due to the important security and safety implications that online media communications have on everyday modern life. On the other hand, several text steganalysis methods have also been proposed in literature, as documented in the work of (Samanta and Pattyanayak, 2020). In addition, datasets such as (Yang et al., 2021a) have been collected for testing text steganalysis methods, and also tools have been developed for text steganography, such as LUNABEL (Chand and Orgun, 2006) and SNOW (Kwan, 2013).

In this paper, we present a new structural approach for encoding secret messages in text documents. We do not manipulate the text itself, but rather the layout of the document. More specifically, we use the breakdown of a document into paragraphs, where the difference between the sizes of adjacent paragraphs is used to encode individual bits of information. We then propose a formal language to adjust the paragraph sizes accordingly. This paper reports on our initial ideas, and lays down the ground for future implementation and validation of these proposals.

[a] https://orcid.org/0000-0001-5089-2025
[b] https://orcid.org/0000-0001-7578-977X
[c] https://orcid.org/0000-0003-2087-5245
[d] https://orcid.org/0000-0003-2671-2199

Most relevant to our work is that of Liang and Iranmanesh (Liang and Iranmanesh, 2016), who proposed a method by adding five white space characters to randomly selected positions in a line using a key to correlate the characters required for embedding secret information. This method is advantageous because randomly-spread white spaces may encode the message differently using different keys.

The rest of the paper is structured as follows. In Section 2, we introduce the theoretical background necessary for defining our embedding method. In Section 3, we define the embedding and extraction processes. In Section 4, we suggest a suitable statistical test that can be used to attack a document embedded with content using our approach. Finally, we conclude the paper in Section 5 and discuss future work.

2 THEORY

We assume that a text document, $S \in \mathcal{S}$, can be defined as a finite sequence of paragraphs, $S = \langle P_1, \ldots, P_n \rangle$. Any other content can be ignored. We call the set of all possible paragraphs, $\mathcal{P}$. Hence, any sequence would be a finite subset, $S \subseteq \mathcal{P}$. We also call the set of all possible sequences of paragraphs (i.e. documents), $\mathcal{S}$. An example of such set would be one constructed from the hypothetical *Hyperwebster* dictionary (Stewart et al., 1996). Furthermore, a paragraph, $P = \langle c_1, \ldots, c_m \rangle$, is a sequence of some finite number of characters with different multiplicities, which we can capture through a function, $ch\_of : \mathcal{P} \to \wp(C)$, where C is the multi-set of all possible characters in some language (e.g. English), and $\wp(C)$ is a power-set preserving the multiplicity (but not the order) of each character in this multi-set (as defined by Axiom V in (Blizard, 1988)):

$$ch\_of(P) = \{|c : c \in P|\}$$

Assuming every paragraph has at least one sentence defined by a punctuation mark e.g. '!', '?' or '.', the condition that $\forall P \in \mathcal{P} : |ch\_of(P)| \in \mathbb{N}^+$ will always hold. The same applies to documents, as we assume there is no empty document, $S = \langle\,\rangle$, since it would be useless for embedding any hidden content. In general, a document and its paragraphs will need to satisfy a number of conditions, such as those above, to ensure these are grammatically and structurally sound. These conditions must remain invariant once the embedding operations are defined in the next section.

We define the union of paragraphs, $\cup_P$, as follows:
$\langle x_1, \ldots, x_n \rangle \cup_P \langle y_1, \ldots, y_m \rangle =$
$\langle x_1, \ldots, x_n, \text{SPACE}, y_1, \ldots, y_m \rangle$

this simply states that the paragraph union is equivalent to the composition of two sequences of characters (i.e. two paragraphs) separated by a newline character.

Now, we can define the following function, $R : \mathcal{P} \times \mathcal{P} \to \{0, 1\}$, to be a function on pairs of paragraphs resulting in one of the values 0, 1. We can define R in any way we want, but for the rest of our model, we will define R as the *size comparison* function:

$$R(P_l, P_r) = \begin{cases} 0 & \textit{if } |ch\_of(P_l)| \leq |ch\_of(P_r)| \\ 1 & \textit{otherwise} \end{cases}$$

Using the R function, a document consisting of n paragraphs can be seen as a sequence of $(n-1)$ zeroes and ones. We call this sequence the *document's R sequence*:

$$\langle R_1, \ldots, R_{n-1} \rangle$$

R is important; it constitutes our embedding function and its definition will determine what secret message, $M \in \langle x_{i=1\ldots n-1} \mid x_i \in \{0, 1\} \rangle$, we can communicate.

3 OUR PROPOSED METHOD

We present here our method for embedding and extracting a secret message. This method can be classified under the structural (or format-based) approaches for text steganography, since it affects the structure of the document rather than its content.

3.1 The Embedding Process

In order to embed a secret message in a document, we first define a language for altering the paragraph structure of documents. We call this language Λ, and define its syntax as follows:

$$\Lambda ::= Op \mid \Lambda \circ \Lambda$$
$$Op ::= P \oplus P \mid \ominus P \mid P \curvearrowleft P \mid P \curvearrowright P$$

Informally, Λ is either a document altering operation, Op, or a sequential composition of two Λ terms, written as $\Lambda \circ \Lambda$. A document altering operation, Op, can be one of the following: $P \oplus P$ is the paragraph merge operation, which takes a pair of paragraphs and merges them into a single paragraph. $\ominus P$ is the paragraph split operation, which splits a paragraph into two paragraphs at the end of some sentence, each paragraph with random proportions. $P \curvearrowleft P$ is the left-shift operation, which takes some characters from the right (in reality, the lower) paragraph and adds them to the left (in reality, upper) paragraph, such that the number of characters in the former becomes less than that in the latter. Finally, $P \curvearrowright P$ is the right-shift operation, which takes some characters from the left (in

reality, upper) paragraph and adds them to the right (in reality, lower) paragraph, such that the number of characters in the latter becomes more than that in the former. All of these operations must respect the invariant conditions on the grammar and structure of a paragraph and the document, for example, that a paragraph must always end with a sentence as defined by a punctuation mark.

We define the formal operational meaning of Λ as in Figure 1, using the semantic operator $[\![\Lambda\ S]\!] \in \mathcal{S}$, defined inductively over the terms of the language and with respect to an existing document S. We now explain the semantic rules. Rule (*Op1*) defines the meaning of a merge operation on a pair of paragraphs that belong to a document by simply merging their sentences and preserving the position of the new paragraph as the position of the two old ones. Rule (*Op2*) defines the meaning of a split operation for a paragraph that belongs to a document, by splitting the paragraph into two paragraphs, such that the position of the new paragraphs occupies the position of the old one. The two new paragraphs have random sizes, but must add up to $|ch\_of(P)|$. Rule (*Op3*) defines the meaning of the left-shift operation on a pair of paragraphs that belong to a document as the modification of those paragraphs such that the number of characters of the left paragraph now exceeds the right one. Similarly, Rule (*Op4*) defines the meaning of the right-shift operation on a pair of paragraphs such that the new pair has more characters in the right paragraph than the left. Finally, Rule (Λ0) defines the meaning of the sequential composition of two Λ terms as the application of the left term first on a document and then the right one after that on the resulting document.

The embedding process now consists of applying terms of the Λ language to any document S, such that $[\![\Lambda\ S]\!] = \langle P'_1, \ldots, P'_n \rangle$ and:

$$\langle R(P'_1, P'_2), \ldots, R(P'_{n-1}, P'_n) \rangle = M$$

where M is the secret message being communicated. It is also possible to simply either *choose* or *construct from fresh* a textual document, $\langle P_1, \ldots, P_n \rangle$, such that the document satisfies the above equation with M.

As a simple example, let us consider the excerpt, $S_{Dickens}$, defined in Figure 2 and taken from Charles Dickens's "Oliver Twist". If we want to transmit the message $M = \langle 0, 1 \rangle$, then we must alter $S_{Dickens}$ such that $R(P_1, P_2) = 0$ and $R(P_2, P_3) = 1$. One way of achieving this would be to apply:

$$(P_2 \curvearrowleft P_3)\ S_{Dickens}$$

under an invariant condition that states that a paragraph must end with a full stop. This then results in a new excerpt, $S'_{Dickens}$, as shown in Figure 3.

3.2 The Extraction Process

The extraction process reverses the embedding process. In order to extract a message, the recipient will need to have agreed on the definition of R with the sender beforehand. With that in mind, the extraction logic can be defined as follows:

$$\mathbf{Y}\ \omega\ \langle P_1, \ldots, P_n \rangle\ \langle\ \rangle = \omega\ (\mathbf{Y}\ \omega)\ \langle P_1, \ldots, P_n \rangle\ \langle\ \rangle$$

where, $\mathbf{Y}$ is Curry's *fixed-point combinator* (Curry and Feys, 1958, p.178), $\langle P_1, \ldots, P_n \rangle$ is the received text document, $\langle\ \rangle$ is the empty sequence, to be filled with the bits of the secret message, and ω is defined as the following λ-calculus expression (Church, 1932):

$$\omega = \lambda f.\lambda s.\lambda \ell.\ \textit{if}\ |s| \leq 1\ \textit{then}\ \ell \\ \textit{else}\ f\ (s \backslash fst(s))\ (\ell : R(fst(s), snd(s)))$$

Here, $fst : \mathcal{S} \nrightarrow \mathcal{P}$ is a partial function that returns the first paragraph element in a sequence, $snd : \mathcal{S} \nrightarrow \mathcal{P}$ is a partial function that returns the second paragraph element in a sequence and $\backslash : \mathcal{S} \times \mathcal{P} \nrightarrow \mathcal{S}$ is a partial function that takes a sequence and a paragraph, and returns a new sequence resulting from the removal of that paragraph from the input sequence. Finally, $(\ell : n)$ joins an element n to the tail of an existing sequence ℓ such that n becomes the last element of the new sequence. For example, $\langle 1, 0, 1 \rangle : 1 = \langle 1, 0, 1, 1 \rangle$. Both *fst* and *snd* are partial since they are not defined over the empty sequence and the 1-element sequence, respectively. $\backslash$ is partial since the element being removed from a sequence may not be a member. The definition of ω returns the current initial message sequence unaltered if the size of the document is one or zero paragraphs, since such document cannot hold any secret messages. Moreover, this is also used as the condition to terminate the fixed point calculation.

To extract the message from the excerpt of Figure 3, we apply the following:

$$\mathbf{Y}\ \omega\ (\mathbf{Y}\ \omega\ (\mathbf{Y}\ \omega\ S'_{Dickens}\ \langle\ \rangle)) = \langle 0, 1 \rangle$$

4 PROPOSED STEGANALYSIS

According to (Taleby Ahvanooey et al., 2019), there are generally three methods for attacking text documents with hidden content: visual attacks that involve a human in comparing two documents, structural attacks that involve modifying the structure of the suspected documents hence destroying its embedded content and finally, statistical attacks where the attacker uses statistical methods to estimate the probability or possibility that a document has some hidden content.

$(Op1)$	$[\![P_l \oplus P_r \langle P_1, \ldots, P_{l-1}, P_l, P_r, P_{r+1}, \ldots, P_n \rangle]\!]$	$=$	$\langle P_1, \ldots, P_{l-1}, P, P_{r+1}, \ldots, P_{n-1} \rangle$ *where,* $P = P_l \cup_P P_r$
$(Op2)$	$[\![\ominus P \langle P_1, \ldots, P_{l-1}, P, P_{r+1}, \ldots, P_n \rangle]\!]$	$=$	$\langle P_1, \ldots, P_{l-1}, P_l, P_r, P_{r+1}, \ldots, P_{n+1} \rangle$ *where,* $P = P_l \cup_P P_r$
$(Op3)$	$[\![P_l \curvearrowleft P_r \langle P_1, \ldots, P_l, P_r, \ldots, P_n \rangle]\!]$	$=$	$\langle P_1, \ldots, P'_l, P'_r, \ldots, P_n \rangle$ *where,* $\lvert ch\_of(P_l)\rvert \leq \lvert ch\_of(P_r)\rvert$ *and* $\lvert ch\_of(P'_l)\rvert > \lvert ch\_of(P'_r)\rvert$
$(Op4)$	$[\![P_l \curvearrowright P_r \langle P_1, \ldots, P_l, P_r, \ldots, P_n \rangle]\!]$	$=$	$\langle P_1, \ldots, P'_l, P'_r, \ldots, P_n \rangle$ *where,* $\lvert ch\_of(P_l)\rvert \geq \lvert ch\_of(P_r)\rvert$ *and* $\lvert ch\_of(P'_l)\rvert < \lvert ch\_of(P'_r)\rvert$
$(\Lambda 0)$	$[\![(\Lambda_l \circ \Lambda_r)\ S]\!]$	$=$	$[\![\Lambda_r\ [\![\Lambda_l\ S]\!]]\!]$

Figure 1: Semantics of the document alteration language Λ

```
P1: "Bow to the Board," said Bumble. Oliver brushed away two or three tears that were lingering
in his eyes, and seeing no board but the table, fortunately bowed to that.

P2: "What's your name, boy?" said the gentleman in the high chair.

P3: Oliver was frightened at the sight of so many gentlemen, which made him tremble; and the
beadle gave him another tap behind, which made him cry; and these two causes made him answer in
a very low and hesitating voice; whereupon a gentleman in a white waistcoat said he was a fool.
Which was a capital way of raising his spirits, and putting him quite at ease.
```

Figure 2: A three-paragraph excerpt, $S_{Dickens}$, taken from Charles Dickens's "Oliver Twist"

```
P1: "Bow to the Board," said Bumble. Oliver brushed away two or three tears that were lingering
in his eyes,and seeing no board but the table, fortunately bowed to that.

P2: "What's your name, boy?" said the gentleman in the high chair. Oliver was frightened at the
sight of so many gentlemen, which made him tremble; and the beadle gave him another tap behind,
which made him cry; and these two causes made him answer in a very low and hesitating voice;
whereupon a gentleman in a white waistcoat said he was a fool.

P3: Which was a capital way of raising his spirits, and putting him quite at ease.
```

Figure 3: The excerpt of Figure 2 with a new layout after encoding the secret message $\langle 0, 1 \rangle$

In general, the first method always succeeds in detecting hidden content if the attacker has access to the cover document. Therefore, our embedding method is unable to withstand such attacks. On the other had, our method resists structural or stylistic changes unless these involve resizing paragraphs, in which case the hidden message is affected. The most interesting method is the statistical one, and we outline below one such attack based on the chi-squared test (Pearson, 1900; Plackett, 1983). The generality of this test renders it a suitable one, albeit a rough one, as it is based only on detecting similarities in numbers of 0s and 1s with the norm, but not other attributes, for example their ordering. Other popular statistical tests, such as Jaro-Winkler's similarity test (Jaro, 1989; Winkler, 1990), are less suitable since they rely on changes in the textual content itself (e.g. character or word replacement), which we avoid in our Λ embedding.

In our case, Pearson's chi-squared test would be defined by the following equation:

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value, i.e. the number of 0s and 1s contained in a document's R sequence, and E_i is the expected value of those 0s and 1s, based, for example, on the results of some empirical study that would be carried out over some text corpus. There are only two categories to sum over, 0 and 1. If we assume that such empirical study produced a result so that the rate of occurrence of 0s in a document's R sequence was Pr_0, then the *null hypothesis* (H_0) for our chi-squared test would be stated as follows:

H_0 : *The average rate at which a* 0 *occurs in a document's R sequence is* $0 \leq Pr_0 \leq 1$

By contrast, the *alternative hypothesis* (H_1) is:

H_1 : *The average rate at which a* 0 *occurs in a document's R sequence is some other value,* $0 \leq Pr'_0 \leq 1$ *where* $Pr'_0 \neq Pr_0$

In testing a suspicious document, S, we then perform the following test:

$$\chi^2 = \frac{(O_0 - (Pr_0 \times |S|))^2}{Pr_0 \times |S|} + \frac{(O_1 - ((1 - Pr_0) \times |S|))^2}{(1 - Pr_0) \times |S|}$$

where $1 - Pr_0$ is the rate of occurrence of 1s in the normal case. If after this test, H_0 holds, meaning that the upper-tail critical value is $\chi^2 \leq 3.841$ for a significance level of 0.05, then this implies that S is clean. On the other hand, if H_1 is proven to be correct, meaning that the upper-tail critical value is $\chi^2 > 3.841$, then this indicates that S is more likely to contain some secret message.

5 CONCLUSION

We presented in this paper briefly a new structural method for embedding and extracting secret messages in text documents. The new method manipulates the layout of a document through the resizing of the documents and then using the size of two adjacent paragraphs to embed an information bit. In general, our method is capable of embedding $n - 1$ number of bits, for a document consisting of n paragraphs. We defined formally the semantics of the embedding language and suggested the chi-squared test as a suitable attack against a document with a hidden message.

Future work will focus on validating the new method through carrying out extensive experiments on various text corpora. This will involve establishing the normal distribution of 0s and 1s to define what the value of Pr_0 is for clean documents. Another direction of future work would be to enrich the formal language and its semantics to include a logical theory defining invariant conditions needed for maintaining the structural and grammatical integrity of paragraphs and documents. Finally, it is also possible to give different other definitions of R, in particular, ones that compare the sizes of non-adjacent paragraphs. Such definitions of R would be *keyed*, where the key indicates which paragraphs are to be compared. Although such keyed versions would be more secure, they introduce the additional problem of key distribution.

REFERENCES

Agarwal, M. (2013). Text steganographic approaches: A comparison. *International Journal of Network Security and its Applications*, 5:91–106.

Barmawi, A. (2016). Linguistic based steganography using lexical substitution and syntactical transformation. In *2016 6th International Conference on IT Convergence and Security (ICITCS)*, pages 1–6.

Bergmair, R. and Katzenbeisser, S. (2006). Content-aware steganography: About lazy prisoners and narrow-minded wardens. In *Information Hiding*, volume 4437, pages 109–123.

Blizard, W. D. (1988). Multiset theory. *Notre Dame Journal of Formal Logic*, 30(1):36 – 66.

Chand, V. and Orgun, C. O. (2006). Exploiting linguistic features in lexical steganography: Design and proof-of-concept implementation. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pages 126b–126b.

Church, A. (1932). A set of postulates for the foundation of logic. *Annals of Mathematics*, 33(2):346–366.

Curry, H. B. and Feys, R. (1958). *Combinatory Logic*. Number v. 1 in Combinatory Logic. North-Holland Publishing Company.

Grosvald, M. and Orgun, C. O. (2012). Human-versus computer-generated text-based steganography: Real-world tests of two algorithms. *Journal of Inf. Hiding and Multimedia Signal Processing*, 3(1):24–33.

Guan, B., Gong, L., and Shen, Y. (2022). A novel coverless text steganographic algorithm based on polynomial encryption. *Sec. and Comm. Networks*, 2022.

Huanhuan, H., Xin, Z., Weiming, Z., and Nenghai, Y. (2017). Adaptive text steganography by exploring statistical and linguistical distortion. In *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*, pages 145–150. IEEE.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Kang, H., Wu, H., and Zhang, X. (2020). Generative text steganography based on lstm network and attention mechanism with keywords. *Electronic Imaging*, 2020(4):291–1.

Kumar, R. and Singh, H. (2020). Recent trends in text steganography with experimental study. In *Handbook of Computer Networks and Cyber Security*, pages 849–872. Springer.

Kwan, M. (2013). The SNOW Home Page.

Li, Y., Zhang, J., Yang, Z., and Zhang, R. (2021). Topic-aware neural linguistic steganography based on knowledge graphs. *ACM Trans. on Data Science*, 2(2):1–13.

Liang, O. W. and Iranmanesh, V. (2016). Information hiding using whitespace technique in Microsoft word. In *Proceedings of the 2016 International Conference on Virtual Systems and Multimedia, VSMM 2016*. Institute of Electrical and Electronics Engineers Inc.

Lockwood, R. and Curran, K. (2017). Text based steganography. *International Journal of Information Privacy, Security and Integrity*, 3(2):134–153.

Majeed, M. A., Sulaiman, R., Shukur, Z., and Hasan, M. K. (2021). A review on text steganography techniques. *Mathematics*, 9(21).

Malik, A., Sikka, G., and Verma, H. K. (2017). A high capacity text steganography scheme based on lzw compression and color coding. *Engineering Science and Technology, an International Journal*, 20(1):72–79.

Naharuddin, A., Wibawa, A. D., and Sumpeno, S. (2018). A high capacity and imperceptible text steganography using binary digit mapping on ascii characters. In *2018 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 287–292. IEEE.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Plackett, R. L. (1983). Karl Pearson and the Chi-Squared Test. *International Statistical Review / Revue Internationale de Statistique*, 51(1):59–72.

Roy, S. and Manasmita, M. (2011). A novel approach to format based text steganography. In *Proceedings of the 2011 International Conference on Communication, Computing and Security*, ICCCS '11, page 511–516, New York, NY, USA. Association for Computing Machinery.

Samanta, S. and Pattyanayak, S. (2020). A significant survey on text steganalysis techniques. *Int. Journal on Computer Science and Engineering*, pages 187–193.

Shirali-Shahreza, M. H. and Shirali-Shahreza, M. (2007). Text steganography in chat. In *2007 3rd IEEE/IFIP International Conference in Central Asia on Internet*, pages 1–5.

Stewart, I. et al. (1996). *From here to infinity*. Oxford Paperbacks.

Taleby Ahvanooey, M., Li, Q., Hou, J., Rajput, A. R., and Chen, Y. (2019). Modern text hiding, text steganalysis, and applications: A comparative analysis. *Entropy*, 21(4).

Wang, K. and Gao, Q. (2019). A coverless plain text steganography based on character features. *IEEE Access*, 7:95665–95676.

Wilson, A., Blunsom, P., and Ker, A. (2015). Detection of steganographic techniques on twitter. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2564–2569.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. Technical report, U.S. Bureau of the Census.

Wu, N., Shang, P., Fan, J., Yang, Z., Ma, W., and Liu, Z. (2019). Coverless text steganography based on maximum variable bit embedding rules. In *Journal of Physics: Conference Series*, volume 1237:2, page 022078. IOP Publishing.

Wu, N., Yang, Z., Yang, Y., Li, L., Shang, P., Ma, W., and Liu, Z. (2020). Stbs-stega: Coverless text steganography based on state transition-binary sequence. *International Journal of Distributed Sensor Networks*, 16(3):1550147720914257.

Wu, Y., Chen, X., and Sun, X. (2018). Coverless steganography based on english texts using binary tags protocol. *Journal of Internet Technology*, 19(2):599–606.

Xiang, L., Sun, X., Luo, G., and Gan, C. (2007). Research on steganalysis for text steganography based on font format. In *Third International Symposium on Information Assurance and Security*, pages 490–495.

Yajam, H. A., Mousavi, A. S., and Amirmazlaghani, M. (2014). A new linguistic steganography scheme based on lexical substitution. In *2014 11th International ISC Conference on Information Security and Cryptology*, pages 155–160.

Yang, Z., He, J., Zhang, S., Yang, J., and Huang, Y. (2021a). Tstego-thu: Large-scale text steganalysis dataset. In *International Conference on Artificial Intelligence and Security*, pages 335–344. Springer.

Yang, Z., Xiang, L., Zhang, S., Sun, X., and Huang, Y. (2021b). Linguistic generative steganography with enhanced cognitive-imperceptibility. *IEEE Signal Processing Letters*, 28:409–413.