Contextual Coherence Evaluation of Perfectly Secure Steganography in Text Documents

Katsuyuki Umezawa[1][0000−0003−3903−3558], Toshikatsu Kashima[1], Sven Wohlgemuth[2], and Kazuo Takaragi[3]

[1] Shonan Institute of Technology, Japan
umezawa@info.shonan-it.ac.jp
https://www.shonan-it.ac.jp/
[2] Fujisawa, Japan
[3] HISAFE, Japan

Abstract. Perfectly secure steganography is a technique designed to render the presence of embedded information undetectable by ensuring that the statistical distributions before and after embedding remain perfectly identical, thereby setting the Kullback–Leibler (KL) divergence to zero. However, in terms of textual documents, specific linguistic features, such as dialects or speech patterns associated with elderly individuals, can create noticeable inconsistencies for human readers, potentially revealing the existence of hidden information.
This study aims to evaluate stegotexts generated through perfectly secure steganography by employing alternative metrics beyond KL divergence. This is not a proposal to replace the KL divergence evaluation metric for perfectly secure steganography, but rather a proposal to conduct an additional evaluation using a different metric. Specifically, we utilize latent Dirichlet allocation (LDA) and cosine similarity to assess the coherence of the generated stegotext with a set of documents sharing the same topic. For our experiments, we collected news articles on gaming, sports, and politics from BBC News. The results indicate that gaming-related stegotexts exhibited significantly lower similarity with sports- and politics-related news articles, and a similar trend was observed for politics-related stegotexts. However, for sports-related stegotexts, no significant difference was detected.

Keywords: Perfectly secure steganography · latent Dirichlet allocation · cosine similarity.

1 Introduction

Steganography has long been studied as a method for concealing data by embedding it within seemingly harmless content such as images, audio, and text. Unlike traditional security measures such as encryption, which aim to protect the data from decrypting the contents, steganography hides the existence of the communication itself. This makes it a valuable complement to cryptographic

technologies. Its applications are varied, encompassing the secure transmission of sensitive information and privacy protection.

Traditional steganographic methods have sought to embed confidential information within a covertext (i.e., content used for steganography to embed data) while minimizing alterations to make detection by third parties challenging. Although these methods can obscure the specific contents of the embedded information when combined with encryption techniques, embedded information may still be statistically detectable. Consequently, the security level provided by traditional steganography has been limited, necessitating more advanced technologies to enhance its effectiveness.

In recent years, a concept referred to as "perfectly secure steganography" has been proposed [1]. This technique aims to modify the covertext in a manner that is statistically undetectable, thus rendering the embedded information imperceptible. Specifically, it requires that the statistical distributions of the covertext and stegotext (i.e., the content containing the embedded information) match perfectly. To achieve this, generative AI technologies, such as GPT-2 and WaveRNN, are used alongside a method called iterative minimum entropy coupling (iMEC) to minimize the Kullback–Leibler (KL) divergence towards zero. This approach effectively hides the fact that information has been embedded.

This study focuses on the steganography of text documents. Although achieving a KL divergence of zero suggests perfect statistical concealment, the possibility that human readers might notice anomalies remains. Detection can occur if the covertext, which may include specific features such as dialects or archaic phrases, reads as unnatural. These discrepancies could inadvertently indicate the presence of embedded information.

We evaluate stegotexts generated through perfectly secure steganography using metrics beyond KL divergence. Specifically, we employ latent Dirichlet allocation (LDA) [2] to model a collection of documents and calculate cosine similarity to assess the similarity between documents. This method helps verify whether the generated stegotexts align with document groups on the same topics. By incorporating this contextual consistency evaluation along with the statistical criteria (KL divergence of zero), we aim to better understand how the texts might be perceived by human readers.

2 Related work

2.1 Perfectly secure steganography

Steganography has gained recognition in the field of information security as a technique that conceals not only the contents of communications but also the act of communication itself. This technology prevents third parties from detecting data modifications, thereby addressing some limitations of cryptographic methods. Notably, information theory-based models, which assess the security of steganography through statistical analysis, are widely employed[3][4].

Cachin's model [5], a foundational approach in steganography, emphasizes the statistical similarity between covertexts and stegotexts. A covertext is defined as natural, inconspicuous data used to embed secret information, such as images, text, or audio. By contrast, a stegotext is the data produced when secret information is embedded within the covertext, designed to ensure that its statistical characteristics align with those of the covertext. This model is intended to completely eliminate the risk of detection by third parties.

As a primary approach to achieving a KL divergence of zero, the minimum entropy coupling (MEC) method has been proposed. MEC is designed to identify the optimal coupling between the distribution of a covertext and that of a stegotext, enabling the statistical properties of the stegotext to closely mirror those of the covertext. Specifically, MEC modifies the entropy structure of the covertext distribution while ensuring that it aligns with the statistical characteristics of the embedded information. This adjustment minimizes discrepancies between the two distributions, effectively avoiding statistical detection.

The emergence of generative AI models such as GPT-2, WaveRNN, and Image Transformer in recent years has significantly improved the performance of steganography. These models can accurately replicate complex covertext distributions, and when integrated with techniques like MEC, they allow for more precise modifications to the entropy structure[1]. This combination greatly enhances the naturalness and secrecy of stegotexts, offering an effective means of reducing KL divergence. In addition, the previously defined method of iMEC has been introduced to improve the alignment between the distributions of covertexts and stegotexts, not only approaching zero KL divergence but also maximizing statistical security. Consequently, the application of MEC alongside generative AI techniques not only boosts the safety of steganography but also enhances its practicality, laying a critical foundation for the advancement of this field.

2.2 Latent Dirichlet allocation (LDA)

In this section, we describe LDA, which is a method used for topic modeling and cosine similarity analysis.

Topic models operate on the premise that a document encompasses several latent topics, with each keyword either belonging to a specific topic or being produced by that topic. LDA [2] is a method for estimating these latent topics from keywords. It is a language model that assumes the probability distribution of topics (parameter θ of the multinomial distribution) that follow a Dirichlet distribution. In LDA, topics are selected according to the Dirichlet distribution, and words are chosen based on the probability distribution associated with that topic.

LDA can be represented by the graphical model shown in Figure 1. This figure illustrates the smoothed LDA model. In recent years, the term "LDA" has commonly referred to smoothed LDA, and our implementation follows this formulation.

Here, d represents the document ID, n denotes the word ID of a word in a document, and k indicates the topic ID. The total number of documents is

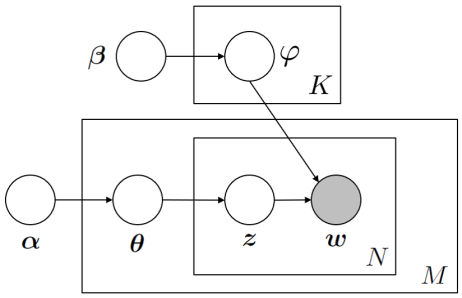

Fig. 1. Graphical model representation of smoothed LDA

M whereas N and K refer to the number of words and topics in a document, respectively. The number of words in document d is given as N_d. The ranges for d, n, and k are specified as $1 \leq d \leq M$, $1 \leq n \leq N_d$, and $1 \leq k \leq K$, respectively. w_{dn} represents the nth word of document d, and z_{dn} represents the latent topic associated with the nth word in document d. The variable θ_{dk} is the mixing ratio for the latent topic k of document d. For example, if document d has three topics with mixing ratios of 10%, 70%, and 20% for topics 1, 2, and 3, respectively, then $\theta_{d1} = 0.1$, $\theta_{d2} = 0.7$, and $\theta_{d3} = 0.2$, reflected as $\boldsymbol{\theta}_d = \{0.1, 0.7, 0.2\}$. In this context, $\boldsymbol{\alpha}$ is the parameter of the Dirichlet prior on the per-document topic distributions, and $\boldsymbol{\beta}$ denotes the parameter of the Dirichlet prior on the per-topic word distribution. $\boldsymbol{\varphi}$ represents the word distribution for topic k. A collection of documents is referred to as a corpus, denoted as $\boldsymbol{D} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_M}\}$.

In this framework, we observe the word $\boldsymbol{w_d}$ and aim to determine the topic from which these words originated. Specifically, for document d, we calculate and estimate the topic distribution $\boldsymbol{\theta}_d$, the topic assignment for each word $\boldsymbol{z}$, and the word distribution for each topic $\boldsymbol{\varphi}$. Various methods, such as the Bayesian method [2], Markov chain Monte Carlo (MCMC) [6], and Gibbs sampling [7] (a specific type of MCMC), have been proposed to estimate these latent variables ($\boldsymbol{\theta}$, $\boldsymbol{z}$ and $\boldsymbol{\varphi}$).

2.3 Cosine similarity

The similarity between documents d and d' can be evaluated by calculating the cosine similarity of $\boldsymbol{\theta}_d$, which is approximately derived using the aforementioned LDA method. The cosine similarity is expressed by the following equation.

$$sim_{cos}(d, d') = \frac{\sum_{k=1}^{K} \theta_{dk} \theta_{d'k}}{\sqrt{\sum_{k=1}^{K} {\theta_{dk}}^2} \sqrt{\sum_{k=1}^{K} {\theta_{d'k}}^2}} \tag{1}$$

3 Proposal

In this study, we utilize LDA to evaluate the similarity and contextual consistency of stegotexts generated through perfectly secure steganography. This approach assesses how well the stegotexts integrate with their respective textual environments, ensuring they maintain a natural appearance while effectively embedding hidden information.

Figure 2 provides an overview of the proposal. It illustrates that articles from three categories, namely, games, sports, and politics, are randomly selected, with 100 articles from each category. The process consists of the following steps: 1) a news article about games is chosen and used as the covertext to generate stegotext through perfectly secure steganography; 2) a topic model is created from news articles across all categories, and this model is used to compute the cosine similarity between the generated stegotext and the articles from each category (games, sports, politics); 3) the similarity is calculated for the sports category; 4) the similarity calculation is repeated for the politics category. This method assesses how effectively the stegotext integrates into the context of each category while preserving its concealed information.

If a high similarity is identified between the stegotext generated from a game-related news article and the original game-related article, it indicates that the stegotext was produced on the same topic. This also holds true for sports and politics. However, even if perfectly secure steganography results in a KL divergence of zero, if the covertext is from the 'game' category and the stegotext is related to 'sports' or 'politics,' readers may notice a discrepancy, as the topic of the stegotext deviates from that of the covertext.

4 Experiment

4.1 Selecting a dataset

In this study, the dataset consists of articles that were randomly selected from BBC News, with 100 articles from each of the three categories: games [9], sports [10], and politics [11].

4.2 Generation of stegotext

The stegotext was generated by executing the source code published on GitHub [12]. The main parameters used are shown in Table 1. Table 4 presents the news articles utilized as covertexts, along with the generated stegotexts.

4.3 Creation of the LDA model

In total, 300 news articles as explained in Section 4.1 were used to create the LDA model, with the number of topics, one of the parameters, manually set to eight.

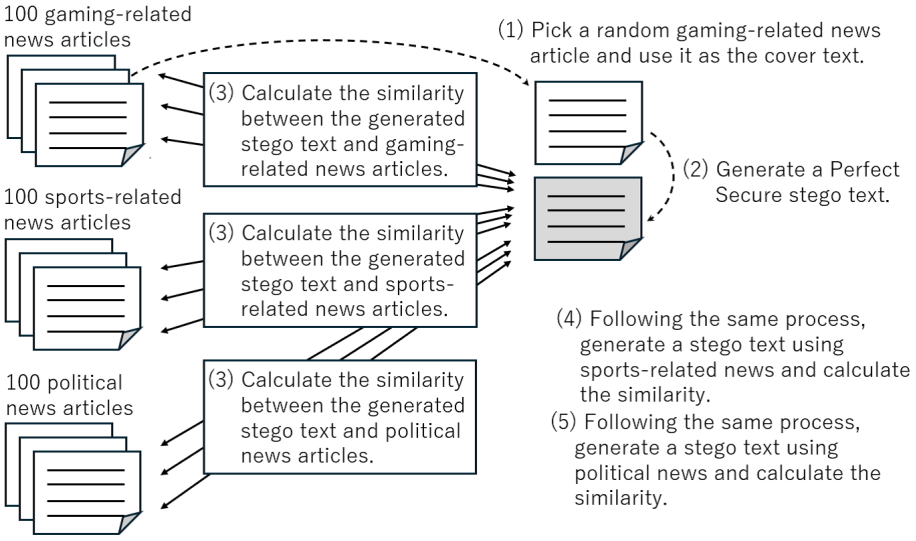

Fig. 2. Overview of the proposal

Table 1. Parameters used when executing perfectly secure steganography.

Parameter	Value
block-size	10
method	imec
top-k	40
message-mode	randombits
repetition	1

4.4 Calculation of cosine similarity

First, we calculated the cosine similarity between the stegotext generated using a game-related news article as the covertext and 100 other game-related news articles. We refer to this as game–game similarity. Next, we determined the cosine similarity between the same stegotext and 100 sports-related news articles, which we call game–sports similarity. We also measured the cosine similarity between the same stegotext and 100 politics-related news articles, which we term game–politics similarity. The same methodology was applied to stegotexts generated from sports- and politics-related news.

These terms are summarized in Table 2. A similarity value close to 1 indicates a strong resemblance of the stegotext to articles from that category, whereas a value close to 0 indicates minimal resemblance.

Table 2. Abbreviations for similarities between Articles A and B

Article A	Articles B	Abbreviation
Similarity between the stegotext generated from gaming-related news and	100 gaming-related news articles	game–game similarity
	100 sports-related news articles	game–sports similarity
	100 political news articles	game–politics similarity
Similarity between the stegotext generated from sports-related news and	100 gaming-related news articles	sports–game similarity
	100 sports-related news articles	sports–sports similarity
	100 political news articles	sports–politics similarity
Similarity between the stegotext generated from political news and	100 gaming-related news articles	politics–game similarity
	100 sports-related news articles	politics–sports similarity
	100 political news articles	politics–politics similarity

5 Experimental results

5.1 Similarity histogram

The histograms of similarities for game-, sports-, and politics-related content are presented in Figures 3, 4, and 5, respectively. The horizontal axis in each figure represents the range of cosine similarity (from 0 to 1), whereas the vertical axis indicates the number of texts within each range.

From Figure 3, we observe that game–game similarity generally yields high values, with a significant number of data points in the range above 0.8. This suggests that stegotexts generated through perfectly secure steganography closely resemble news articles in the same category as the covertext. By contrast, game–sports and game–politics similarities are predominantly concentrated below 0.4, with very few instances exceeding a value of 0.8. Notably, this indicates that game stegotexts have minimal similarity with news articles from different categories.

Figure 4 clearly shows that sports–sports similarity generally shows high values, with a significant number of data points in the range above 0.85. By

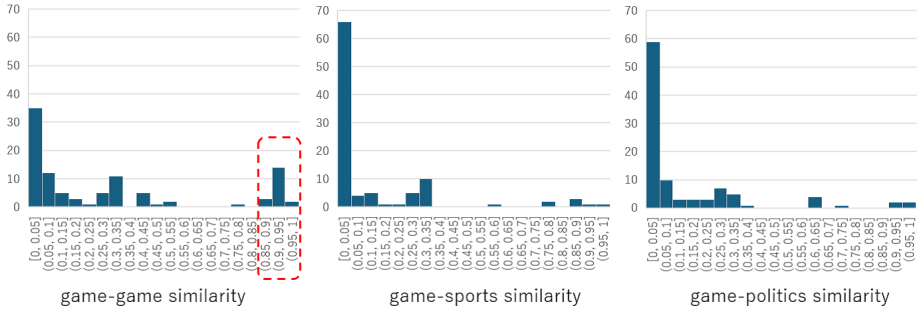

Fig. 3. Histogram of game-related similarities.

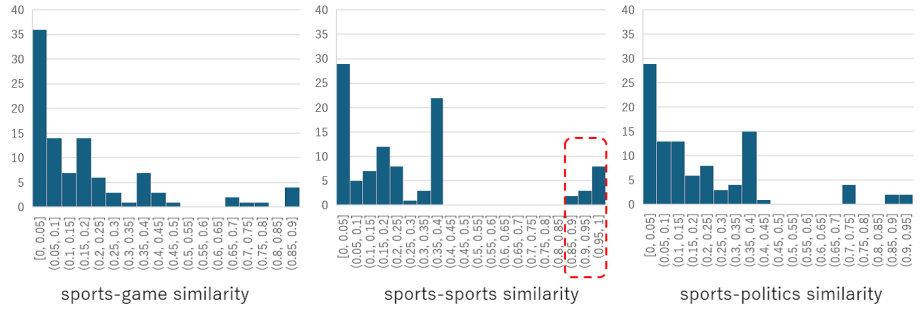

Fig. 4. Histogram of sports-related similarities.

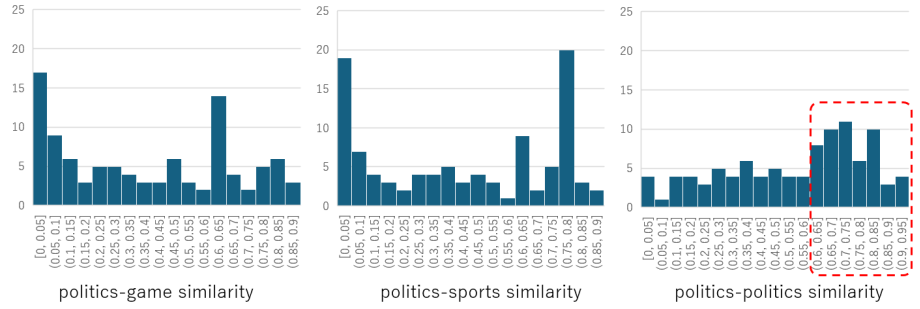

Fig. 5. Histogram of political similarities.

contrast, the similarities between sports–game and sports–politics are mainly concentrated below 0.4, and only a few instances exceed a 0.85. In addition, over 70% of the data is below 0.4, indicating that sports stegotexts have minimal similarity with news articles in the game and politics categories.

Figure 5 illustrates that politics–politics similarity is relatively high, with a significant number of data points above 0.6. Most of the data is concentrated between 0.6 and 0.9, suggesting that stegotexts within the politics category show strong similarity with articles from the same category. By contrast, although politics–game and politics–sports similarities contain data points below 0.2, a significant number of instances are above 0.6. The proportion of data above 0.6 in politics–sports similarity is particularly high, indicating a certain level of association between political stegotexts and sports news. This trend may be due to overlaps in the reporting of politics and sports news.

Table 3 presents the average values of each similarity measure.

Table 3. Average values of each similarity

Category	Similarity	Average value (Mean)
game	game-game similarity	0.293400307
	game–sports similarity	0.126264980
	game–politics similarity	0.135203711
sports	sports–game similarity	0.180721423
	sports–sports similarity	0.274772562
	sports–politics similarity	0.206046409
politics	politics–game similarity	0.384843029
	politics–sports similarity	0.419023576
	politics–politics similarity	0.540932822

6 Analysis and evaluation

The previous section described the visual assessment conducted in this study. This section describes our statistical verification of the differences between categories. Specifically, we tested whether the average similarity values shown in Table 3 differed significantly.

The histograms from Figure 3 to Figure 5 indicate that the data do not follow a normal distribution. Therefore, we conducted a Wilcoxon rank-sum test, which is a non-parametric test that can be used even when the data does not follow a normal distribution, to assess statistical differences.

The tests were conducted pairwise within each category to determine differences in the mean values. Because a total of nine tests were required, we addressed the issue of multiple testing. To do this, we applied the Bonferroni correction, one of the methods to avoid the issue of multiple testing. Specifically, the significance level, or p-value, was adjusted using the Bonferroni method to a threshold of $0.05/9 = 0.0055556$.

The results of the tests are shown in Figure 6. As illustrated, stegotexts generated from game news articles exhibited high similarity with other game news articles. The same was true for sports and politics. Significant differences were found between game and politics categories. However, for sports, no significant differences were observed among all similarity measures.

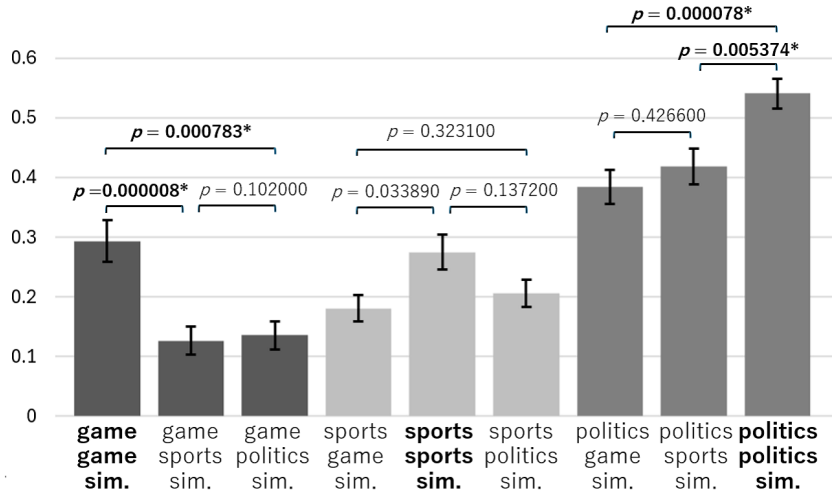

Fig. 6. The results of the tests

7 Discussion

In this study, we conducted experiments using the LDA model and cosine similarity to assess the theoretical perfection of perfectly secure steganography by evaluating the semantic similarity between stegotexts and news articles. The results indicated that stegotexts related to games and politics exhibited a higher cosine similarity to news articles within the same category than to those from different categories. Notably, strong similarities were observed between game stegotexts and game news, as well as between political stegotexts and political news, confirming that the content in these two categories was well-aligned.

By contrast, stegotexts in the sports category showed similarities to articles in other categories, with no statistically significant differences found. This may be attributed to the diversity of vocabulary in sports news and its commonalities with other categories. Specifically, overlapping topics may exist between sports and political news, leading to stegotexts that are not strictly confined to sports content but that also resemble other categories. In addition, the brevity of the news articles or the excessive number of topics defined in the LDA model may have blurred the distinctions between categories. When articles are short, LDA

may struggle to extract sufficient features for precise topic estimation, resulting in ambiguous topic distributions. In this study, we set the number of topics for the LDA to eight, which may not have been adequate for clear categorization.

8 Conclusion

In this study, we utilized LDA and cosine similarity to evaluate the similarity between documents, verifying whether the generated stegotexts resembled document groups with shared topics. The results confirmed that stegotexts derived from game-related news articles had significantly lower similarity to news articles in other categories (sports and politics). Likewise, stegotexts generated from political news articles also demonstrated significantly lower similarity to news articles in different categories. However, no significant differences were identified among stegotexts related to sports. Although the findings did not entirely meet expectations across all categories, two out of three categories demonstrated the anticipated similarity levels. These results suggest that perfect secure steganography not only achieves statistical perfection but also produces texts that resemble articles within the same category when read by humans.

In the future, expanding the variety of categories and the lengths of individual news articles is essential. In addition, providing an appropriate number of topics when training the LDA model is critical. Finally, utilizing data such as dialects and archaic phrases is necessary to ensure that the generated documents do not create a sense of incongruity for native speakers. These steps will enhance the accuracy and effectiveness of steganography in preserving the naturalness and relevance of covertexts.

Acknowledgments

Part of the work was supported by JDC Foundation for the Promotion of Academic Research.

References

1. Schroeder de Witt, C., Sokata, S., Kolter, Z., Forester, J., and Strohmeier, M., "PERFECTLY SECURE STEGANOGRAPHY USING MINIMUM ENTROPY COUPLING," Published as a conference paper at ICLR 2023, (2023).
2. Blei, D., Ng, A. and Jordan, M., "Latent Dirichlet Allocation," Journal of Machine Learning Research, 2003, pp. 1107–1135.
3. Sallee, P. "Model-Based Steganography." In: Kalker, T., Cox, I., Ro, Y.M. (eds) Digital Watermarking. IWDW 2003. Lecture Notes in Computer Science, vol 2939. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24624_12
4. Chhikaraa, S. and Kumarb, R., "An Information Theoretic Image Steganalysis for LSB Steganography." Acta Cybernetica 24 (2020) p.p. 593–612. 2020, DOI: 10.14232/actacyb.279174

5. Cachin, C., "An information-theoretic model for steganography," In David Aucsmith (ed.), Information Hiding, pp. 306–318, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-49380-8 (1998).
6. Griffiths, T. and Steyvers, M., "Finding scientific topics," Proceedings of the National Academy of Science, 2004, pp. 5228-5235.
7. Teh, Y., Newman, D. and Welling, M., "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation," Proceedings of Advances in Neural Information Processing Systems 19, NIPS '07, Cambridge, MA, pp. 1353–1360, 2007.
8. Spooky_Maskman, "Topic analysis by LDA with Gensim," https://qiita.com/Spooky_Maskman/items/0d03ea499b88abf56819, last accessed June 6, 2025.
9. BBC Gaming, https://www.bbc.com/news/topics/c008ql15vdxt, last accessed June 6, 2025.
10. BBC SPORT, https://www.bbc.com/sport, last accessed June 6, 2025.
11. BBC politics, https://www.bbc.com/search?q=politics&edgeauth=eyJhbGciOiAiSFMyNTYiLCAidHlwIjogIkpXVCJ9.eyJrZXkiOiAiZmFzdGx5LXVyaS10b2tlbi0xIiwiZXhwIjogMTczODExMDk5OSwibmJmIjogMTczODExMDYzOSwicmVxdWVzdHVyaSI6ICIlMkZzZWFyY2glM0ZxJTNEcG9saXRpY3MifQ.nvt2u8NmlOLjrAzzFiW_ScVYJlGQrijzbNqAoqEjv8g, last accessed June 6, 2025.
12. GITHUB: perfectly-secure-steganography, https://github.com/schroederdewitt/perfectly-secure-steganography, last accessed June 6, 2025.

A Covertext and generated stegotext

Table 4 presents the news articles used as covertexts and the generated stegotexts.

Table 4. The covertext and the generated stego text

Category	Kind	Text
Game	covertext	The forthcoming film comes during a renaissance for gaming films, which were once thought to be almost guaranteed flops. In the 90s, film adaptations of popular video games such as Super Mario Bros and Mortal Kombat were critical and commercial failures when they were released.But Nintendos star-studded Super Mario Bros Movie, which was released earlier this year, became one of the highest-grossing animated films of all time, earning more than $1bn (£800m) at the box office. While Mario was the most successful, it is not a one-off. Films based on Segas Sonic the Hedgehog, starring Jim Carrey, Ben Schwartz and James Marsden, have made hundreds of millions in box office takings, and Nintendos Detective Pikachu, starring Ryan Reynolds as the titular Pok?mon, was a box office hit too.
	Generated stego text	But with Nintendo, with its huge gaming audience and lucrative licensing deals, the company has been facing tough financial struggles. In the last ten years it has been embroiled in a string of scandals that led to massive losses, most famously for its gaming assets. In May 2015, the film released the first-ever leak of sensitive information about the Nintendo Switch console.
Sports	covertext	Meanwhile, Naomi Osaka retired from the final of the Auckland Classic with an abdominal injury.The four-time Grand Slam winner, who was playing in her first WTA final for three years after taking a 15-month break and giving birth to her first child midway through 2023
	Generated stego text	with a populist and anti-immigration message that included references to the euro, free trade and the end of collective bargaining, despite his staunch support for the party. "I think we're a very different country from Germany," he said, "but I also think the world needs to realize how Germany is more than a German colony." But at the end, the talk was dominated by the two
Politics	covertext	Elon Musk took his endorsement of Germanys far-right party to the next level on Thursday, hosting a live chat with its frontwoman, Alice Weidel. The 74-minute conversation ranged across energy policy, German bureaucracy, Adolf Hitler, Mars and the meaning of life. The worlds richest man unequivocally urged Germans to back Alternative f_r Deutschland (AfD) in forthcoming elections. Its the tech billionaires latest controversial foray into European politics. Thered been a considerable build-up to this discussion as Elon Musk faced accusations of meddling in Germanys snap election
	Generated stego text	with a populist and anti-immigration message that included references to the euro, free trade and the end of collective bargaining, despite his staunch support for the party. "I think we're a very different country from Germany," he said, "but I also think the world needs to realize how Germany is more than a German colony." But at the end, the talk was dominated by the two