Clustering-Driven Pseudo-Labeling in Source-Free Domain Adaptation for Linguistic Steganalysis

Yufei Luo
Beijing University of Posts and Telecommunications
Beijing, China
luoyf@bupt.edu.cn

Zhen Yang*
Beijing University of Posts and Telecommunications
Beijing, China
yangzhenyz@bupt.edu.cn

Xin Xu
Tsinghua University
Beijing, China
xuxin527@tsinghua.edu.cn

Yelei Wang
Beijing University of Posts and Telecommunications
Beijing, China
yel_wang@bupt.edu.cn

Ru Zhang*
Beijing University of Posts and Telecommunications
Beijing, China
zhangru@bupt.edu.cn

Yongfeng Huang
Tsinghua University
Beijing, China
yfhuang@tsinghua.edu.cn

Abstract

Linguistic steganalysis often encounters domain shift in practice, mainly due to differences in text sources and steganographic methods. These variations create distribution discrepancies between the training (source domain) and test (target domain) sets, reducing detection accuracy. Most existing domain adaptation methods for linguistic steganalysis rely on labeled source domain data to alleviate these issues, but due to data privacy or transmission costs, source domain data is often unavailable in many real-world scenarios. Without access to this data, models cannot directly compare feature distributions between the source and target domains, hindering the model's ability to learn the target domain's features and ultimately affecting detection performance for stego texts. In this paper, we propose a Clustering-driven Pseudo-labeling method for Source-free domain adaptation in Linguistic Steganalysis (CPSLS). During the adaptation phase, we leverage the clustering structure of the target domain data to generate pseudo-labels, helping the model identify stego features in the target domain. Additionally, we use a weighted classification loss function to reduce the impact of incorrect pseudo-labels. To prevent the model from overlooking the diversity between stego and cover texts during optimization, we introduce a prediction diversity loss, improving the model's ability to differentiate between the two. Experimental results show that CPSLS not only has stronger practical applicability but also outperforms existing domain adaptation linguistic steganalysis in terms of detection accuracy.

*Corresponding authors.

IH&MMSEC '25, San Jose, CA, USA

ACM ISBN 979-8-4007-1887-8/25/06
https://doi.org/10.1145/3733102.3733114

CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Natural language processing**.

Keywords

Linguistic Steganalysis, Source-free Domain Adaptation, Clustering Structure, Pseudo Label.

ACM Reference Format:
Yufei Luo, Zhen Yang, Xin Xu, Yelei Wang, Ru Zhang, and Yongfeng Huang. 2025. Clustering-Driven Pseudo-Labeling in Source-Free Domain Adaptation for Linguistic Steganalysis. In *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSEC '25), June 18–20, 2025, San Jose, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3733102.3733114

1 Introduction

Deep learning-based linguistic steganalysis methods examine the statistical distribution of textual features to detect hidden information [17, 16, 19, 23, 31, 15]. These methods assume that the training (source domain) and testing (target domain) share the same distribution, allowing them to effectively distinguish between cover and stego texts. However, in real-world scenarios, variations in text sources or language usage can lead to domain mismatch, where the distributions of the source and target domains differ. Even small distributional discrepancies can severely impair the detection performance of these models [28].

To tackle the domain mismatch issue in linguistic steganalysis, researchers have turned to domain adaptation methods [22, 21]. These methods focus on bridging the distribution gap by transferring steganalysis-specific knowledge from labeled source domains to unlabeled target domains, thereby reducing the impact of domain mismatch. For instance, Xue et al. [22] utilized Maximum Mean Discrepancy (MMD) [5] as a distance metric to align the distributions of the source and target domains. Additionally, Xue and colleagues [21] developed a Steganographic Domain Distance Metric (SDDM) to further address the domain mismatch issue in steganalysis.

While domain adaptation methods in steganalysis have made progress in addressing the domain mismatch issue, they typically rely on the availability of labeled source domain data during the

adaptation phase, as shown in the figure 1. However, such data may be restricted due to privacy regulations or commercial confidentiality, preventing its use in the adaptation process. Moreover, steganalysis often requires the analysis of large-scale text data, and in resource-constrained environments, such as edge devices, real-time access to source data is impractical. In the domain adaptation phase, when source domain data is unavailable and only a pre-trained source domain model can be used, i.e., Source-Free Domain Adaptation (SFDA), as illustrated in figure 1, existing domain adaptation methods for linguistic steganalysis face considerable challenges. Due to the unavailability of source domain data, existing domain adaptation methods for linguistic steganalysis cannot leverage source domain data for feature alignment, resulting in an inability to effectively learn the target domain's features, which severely impacts the model's detection performance in the target domain.

To achieve Source-Free Domain Adaptation, Luo et al. [10] proposed a novel cross-domain model called PDTS. This model adopts a progressive strategy, gradually generating pseudo-labels for target domain data, enabling the model to learn the distinguishing features of cover and stego texts in the target domain. However, PDTS still faces two key challenges. First, differences in linguistic style or steganographic strategies between the target and source domains may lead to the pseudo-labels containing incorrect labels. These errors can interfere with model training, causing misclassifications of target domain samples and degrading the model's generalization ability and detection accuracy. Second, since the distribution gap between stego and cover texts is minimal, the model may suffer from posterior collapse during optimization, causing it to focus predictions on either stego or cover texts, neglecting the diversity and complexity of the data, and ultimately diminishing its ability to effectively distinguish between the two classes.

To improve the detection performance of the model in source-free domain adaptation scenarios, we design a clustering-driven Pseudo-labeling method for linguistic steganalysis, termed CPSLS. First, although there are distribution discrepancies between the source and target domains, they still share certain similarities. For example, the News and Movie data exhibit similar language styles and expressions. Therefore, despite the presence of feature shift in the target domain, the stego and cover text features extracted by the pre-trained source model can still form natural clusters in the feature space, with high intra-class similarity, as shown in the figure 2. From the figure, it can be observed that even without label supervision in the target domain, the stego and cover samples exhibit a certain degree of intra-class cohesion. This supports the feasibility and validity of generating pseudo-labels based on the clustering structure. Based on this observation, during the domain adaptation phase, we leverage the clustering structure of the target domain data to generate pseudo-labels, thereby enhancing the reliability of the pseudo-labels. These pseudo-labels support supervised training of the model, deepening its understanding of the stego features in the target domain and effectively mitigating the impact of domain shift. Additionally, by evaluating the uncertainty of the predicted distributions, we weight the classification loss according to the reliability of the samples, ensuring that the learning process places more emphasis on reliable pseudo-labels. Furthermore, to prevent posterior collapse during optimization, we introduce a prediction

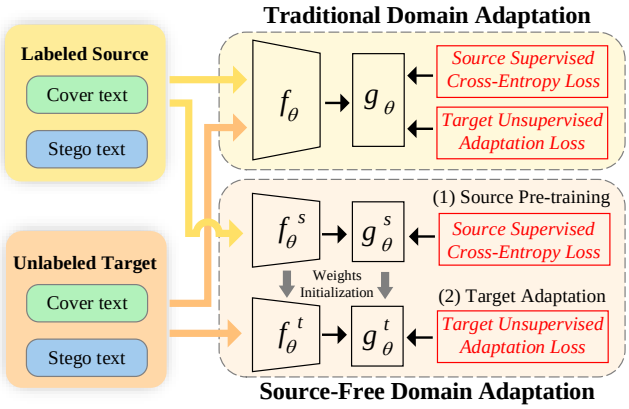


Figure 1: (Top) Traditional domain adaptation for linguistic steganalysis requires both source and target domain data. (Bottom) In source-free domain adaptation, adaptation relies on large unlabeled target data and a source-trained linguistic steganalysis model.

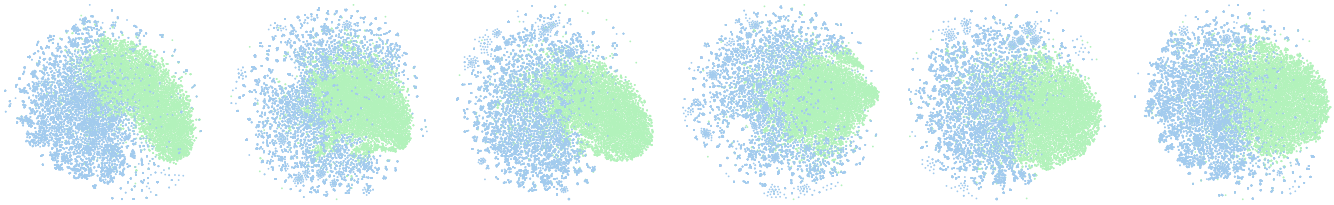
diversity loss as a regularization term. This loss preventing the model from focusing predictions too narrowly on either stego or cover texts, thus maintaining the diversity of the predicted distributions. In summary, the main contributions of this paper are as follows:

1. We leverage the clustering structure of the target domain data to generate pseudo-labels, thereby enhancing the reliability of the pseudo-labels and deepening the model's understanding of the steganographic features in the target domain.
2. We propose a weighted classification loss function and a prediction diversity loss function to mitigate the negative impact of pseudo-label noise and prevent posterior collapse during the optimization process.
3. We conducted experiments on three common natural corpora, and the results show that CPSLS not only effectively applies to source-free domain adaptation scenarios but also outperforms existing methods in detection performance for linguistic steganalysis.

2 RELATED WORK

2.1 Domain Adaptation in Linguistic Steganalysis

In practical applications of linguistic steganalysis [2, 12, 20, 24], the training data (source domain) and test data (target domain) often come from different domains, such as news articles or social media posts. These types of texts exhibit distinct language patterns, vocabulary choices, and writing styles. Furthermore, language usage is influenced by factors such as region, culture, and time, leading to significant differences in vocabulary, grammatical structures, and even overall style. Differences in text sources or language usage may result in discrepancies between the source and target domain distributions. These distributional differences make it challenging for a model trained on the source domain to maintain its detection performance when applied to the target domain, thereby diminishing its ability to detect stego texts in practical applications.



(a) $M \rightarrow N$ (b) $M \rightarrow T$ (c) $N \rightarrow M$ (d) $N \rightarrow T$ (e) $T \rightarrow M$ (f) $T \rightarrow N$

Figure 2: Visualization of the clustering structure of target domain data in the feature space: green and blue represent cover texts and stego texts in the training set, respectively, with labels derived from the ground truth. The data features are extracted using the pre-trained model M_s used in this paper. Figures (a) to (f) show six cross-domain tasks conducted on three commonly used natural corpora, where "N" stands for "News," "M" stands for "Movie," and "T" stands for "Twitter," with stego texts generated using the AC steganographic algorithm.

To address this domain mismatch issue in linguistic steganalysis, researchers have proposed various domain adaptation methods [22, 21, 10]. For example, Xue et al. [22] introduced Maximum Mean Discrepancy (MMD) as a distance measure to align the features between the source and target domains, enhancing the adaptability of the steganalysis model. Later, they [21] proposed the SANet network, which combines SDDM and an adaptive weighting selection mechanism to further reduce the domain gap. However, these methods rely on source domain data to establish feature alignment rules. When source domain data is unavailable, they are unable to effectively align the features between the source and target domains, which limits the model's generalization ability and significantly reduces its performance in detecting stego texts in the target domain.

To address the domain mismatch issue in source-free scenarios, Luo et al. [10] proposed the PDTS model, which uses a progressive strategy to generate pseudo-labels for target domain data. This approach helps the model identify the distinct features of stego and cover texts in the target domain, improving detection performance. However, it does not account for the noise present in the generated pseudo-labels or the potential posterior collapse during optimization, which results in a significant decrease in detection performance for stego texts in the target domain.

2.2 Source-free Domain Adaptation

Traditional text steganalysis methods typically rely on labeled source domain data during domain adaptation [22, 21] , as illustrated in Figure 1. However, in many real-world applications, source domain data is difficult to access due to privacy concerns, legal restrictions, or storage limitations. For example, source domain texts may involve personal privacy, sensitive information, or confidential data, such as communication content, government, or corporate internal documents, which cannot be publicly used due to privacy or legal regulations. In the absence of source domain data, existing domain adaptation methods for text steganalysis cannot effectively address the domain mismatch issue.

To address this challenge, Source-Free Domain Adaptation (SFDA) methods have been developed [8, 25, 18]. These approaches aim to adapt to the target domain without relying on source domain data, instead using a pre-trained model on the source domain and unlabeled data from the target domain, as illustrated in Figure 1. While the SFDA in text steganalysis has been relatively underexplored, it has garnered significant attention in the Computer Vision (CV) community [7, 6, 14]. For instance, Liang et al. [9] proposed the SHOT method, which refines the source model using pseudo-labels without requiring source data. Yang et al. [26] proposed the G-SFDA method, which uses local structure clustering (LSC) to cluster data based on sample similarity, making the local structure of the target domain more compact and thereby more effectively capturing the local features of the target domain data.

However, the application of these SFDA methods in linguistic steganalysis still faces certain limitations. The differences in text sources and the complexity of steganographic techniques result in more pronounced distributional discrepancies between the source and target domains. This not only complicates the adaptation of the source domain model to the target domain but also increases the error rate in pseudo-label generation, negatively impacting training performance. Furthermore, the subtle distributional differences between stego and cover texts can lead to posterior collapse during training. In such cases, the model may become overly reliant on predictions from a single category, neglecting others, which hinders its ability to learn the full range of target domain features and destabilizes the training process. Consequently, current methods must address the challenges of managing distributional discrepancies, reducing pseudo-label noise, and preventing posterior collapse in the context of linguistic steganalysis.

3 PROPOSED APPROACH

In this section, we first introduce the definition of Source-Free Domain Adaptation in linguistic steganalysis. We then describe the framework of the steganalysis model used, followed by an overview of the overall process of the proposed CPSLS method.

3.1 Problem Definition

In this study, we explore linguistic steganalysis within the framework of source-free domain adaptation. We represent the source domain dataset as $D_s = \{(X_i^s, y_i^s)\}_{i=1}^{n_s}$, which consists of n_s labeled samples, where X_i^s is the i-th sentence from the source domain and y_i^s is the corresponding class label. The target domain dataset $D_t = \{X_j^t\}_{j=1}^{n_t}$ consists of n_t unlabeled samples, where X_j^t is the

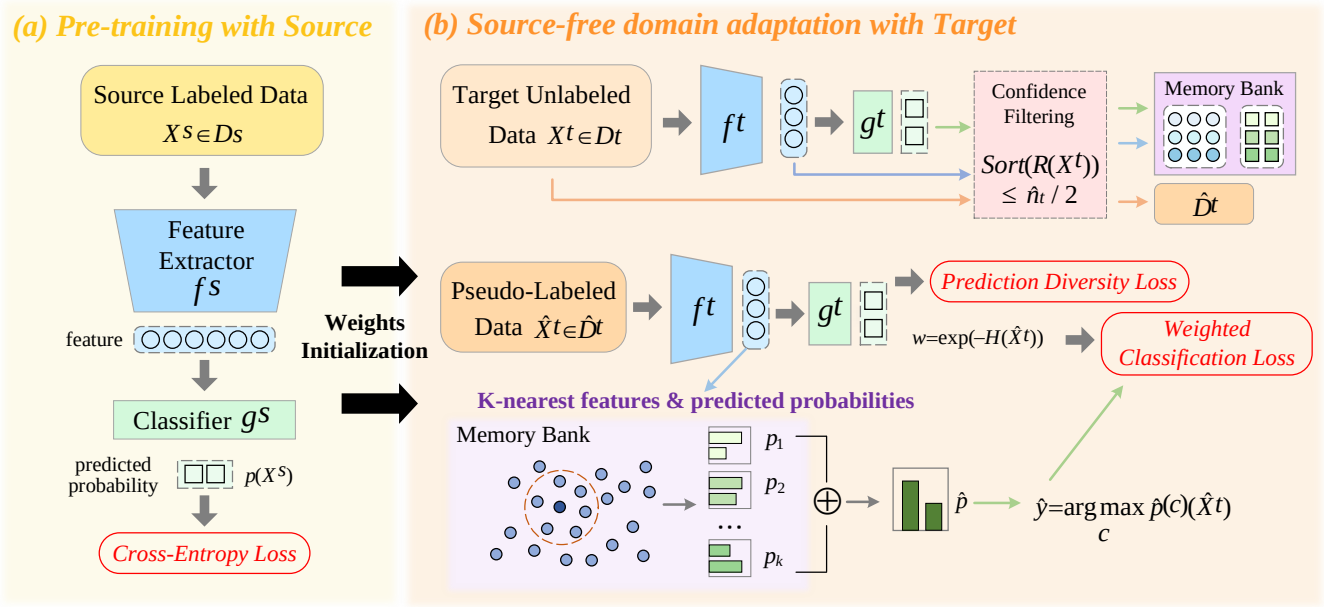


Figure 3: Overall process of CPSLS for source-free domain adaptation in linguistic steganalysis.

j-th sentence in the target domain. Both datasets share the same class labels, specifically "cover text" and "stego text".

In the source-free domain adaptation setting, the source domain dataset D_s is only used to pre-train the steganalysis model, and it is not accessible during adaptation to the target domain. Our goal is to optimize this steganalysis model, pre-trained on source domain dataset, so that it can effectively distinguish between stego and cover texts using the unlabeled data in the target domain dataset.

3.2 The Framework of CPSLS

The overall process of the proposed CPSLS method is shown in Figure 3. The linguistic steganalysis model M mainly consists of two components: the feature extractor $f(\cdot)$ and the linear classifier $g(\cdot)$. The feature extractor $f(\cdot)$ is responsible for extracting useful features from the input data, while the linear classifier $g(\cdot)$ is used to classify the extracted features into target categories, such as cover text or stego text. Through the collaboration of these two components, the model M can effectively learn the features of target domain data, thus improving its ability to detect stego texts.

In the pre-training phase, we train the model M using labeled source domain data through empirical risk minimization. This enables the model to gain an initial understanding of the target domain's features and helps form a clustering structure of target domain data in the feature space, resulting in the pre-trained source domain model M_s. In the domain adaptation phase, we leverage the clustering structure of the target domain and use the pre-trained model M_s to generate pseudo-labels for the target domain data, allowing the model to learn the stego features of the target domain more effectively and resulting in the domain-adapted model M_t. To further enhance the model M_t's adaptation to the target domain, we introduce a weighted classification loss function and a prediction diversity loss function. These mechanisms significantly improve the model M_t's ability to effectively detect stego texts in the target domain.

3.3 The Feature Extractor $f(\cdot)$

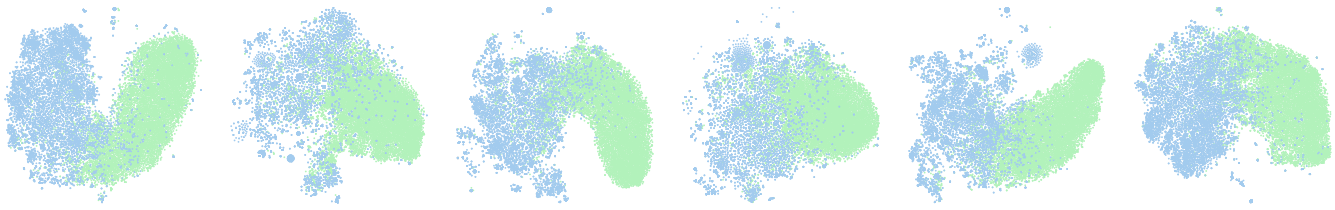
To capture the rich semantic information in text, we utilize a pre-trained BERT model [1] combined with a single-layer Bi-LSTM [4] as the feature extractor $f(\cdot)$. The BERT model consists of L layers of transformer encoders, with each layer having a hidden dimension of d_b. For an input sentence $X = \{x_1, x_2, \ldots, x_n\}$ from either the source or target domain, we first obtain the initial representation $H_0 = \{e_1, e_2, \ldots, e_n\}$ through the embedding layer. Here, n is the sentence length, x_i is the i-th word, and e_i is its corresponding embedding, which includes token, positional, and segment embeddings. Then, H_0 is passed through BERT for processing, with the output at the l-th layer represented as:

$$H_l = \text{Transformer}_l(H_{l-1}), 1 \leq l \leq L \tag{1}$$

Finally, the resulting feature matrix from BERT can be represented as $H(X) \in \mathbb{R}^{n \times d_b}$. During training, the parameters of BERT are fixed. Next, the Bi-LSTM processes the feature matrix $H(X)$ from BERT to extract temporal features and capture long-range dependencies. The output feature representation from the Bi-LSTM is denoted as $f(X)$, and is calculated as:

$$f(X) = [\overrightarrow{\text{LSTM}}(H(X)), \overleftarrow{\text{LSTM}}(H(X))] \tag{2}$$

Here, the dimension of $f(X)$ is $\mathbb{R}^{2 \times n \times d_l}$, where d_l is the LSTM hidden layer dimension and n is the sequence length. This approach allows $f(X)$ to integrate contextual semantic information, providing a comprehensive feature representation.



(a) $M \rightarrow N$ (b) $M \rightarrow T$ (c) $N \rightarrow M$ (d) $N \rightarrow T$ (e) $T \rightarrow M$ (f) $T \rightarrow N$

Figure 4: Visualization of the clustering structure of target domain data in the feature space: green and blue represent cover text and stego text in the training set, respectively, with labels derived from the ground truth. The data features are extracted using the domain-adapted model M_t employed in this paper. Figures (a) to (f) show six cross-domain tasks conducted on three commonly used natural corpora, where "N" represents "News," "M" represents "Movie," and "T" represents "Twitter," with stego texts generated using the AC steganographic algorithm.

3.4 The Linear Classifier $g(\cdot)$

To determine the probability distribution of the text, the classifier $g(\cdot)$ consists of a fully connected layer followed by a Softmax function. The fully connected layer takes the feature vector $f(X)$ output by the feature extractor and maps it to a vector $g(X)$, where the dimension corresponds to the number of classes:

$$g(X) = Wf(X) + b \tag{3}$$

Here, W is the weight matrix, b is the bias term. The Softmax function then processes the vector $g(X)$, normalizing the scores g_i for each class into probabilities p_i:

$$p_i = \frac{\exp(g_i)}{\sum_{j=1}^{C} \exp(g_j)} \tag{4}$$

where C is the total number of classes, g_i is the unnormalized score for class i, and p_i is the corresponding predicted probability.

During testing, a detection threshold θ is applied to classify the test text. For an input test text x_{test}, if its predicted probability p exceeds the threshold, it is classified as stego text. Otherwise, it is classified as cover text. The classification rule is:

$$x_{test} \in \begin{cases} \text{stego text}, & \text{if } p > \theta \\ \text{cover text}, & \text{if } p \leq \theta \end{cases} \tag{5}$$

3.5 Pre-training with Source Domain Labeled Data

During the pre-training phase, we use the labeled source domain data to help the model gain an initial understanding of the target domain data. The model is trained with the cross-entropy loss function, defined as:

$$\mathcal{L}_{pre} = -\frac{1}{n_s} \sum_{i=1}^{n_s} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{6}$$

Here, n_s denotes the number of samples in the source domain, y_i is the true label of the i-th sample (0 for cover text and 1 for stego text), and p_i represents the model's predicted probability that the i-th sample is stego text. By minimizing this loss, this phase helps the model effectively learn the source domain features. Since both the source and target domains come from natural language corpora and share certain similarities, the model can naturally organize the target domain data into clusters of stego texts and cover texts in the feature space, thereby laying a solid foundation for adapting to the target domain data.

Algorithm 1 Clustering-Based Pseudo-Label Generation for Source-Free Domain Adaptation in Linguistic Steganalysis

Input: Unlabeled target domain dataset $D_t = \{X_i^t\}_{i=1}^N$, pre-trained model parameters θ, pseudo-label selection ratio p, number of target domain samples N, number of pseudo-label samples $\hat{n}_t$, memory bank $\mathcal{B}$.

Output: Pseudo-labeled dataset $\hat{D}^t = \{(\hat{X}_i^t, \hat{y}_i^t)\}_{i=1}^{\hat{n}_t}$.

Initialize memory bank: $\mathcal{B} = \emptyset$.
Compute confidence for all samples: $R(X_i^t) = \max p^{(c)}(X_i^t)$, where c belongs to the class set C, which includes cover and stego text classes.
Sort D_t based on $R(X_i^t)$ in descending order.
Initialize pseudo-labeled dataset: $\hat{D}^t = \emptyset$.
for each class $c \in C$ **do**
    for each of the top $\hat{n}_t/2$ samples X_i^t in class c **do**
        Add X_i^t to $\hat{D}^t$ and store its feature representation $f(X_i^t)$ and predicted distribution $p(X_i^t)$ in $\mathcal{B}$.
    end for
end for
for each $\hat{X}_i^t \in \hat{D}^t$ **do**
    Select the K nearest neighbors of $\hat{X}_i^t$ from the memory bank $\mathcal{B}$, denoted as $\mathcal{I}_i = \text{KNN}(\hat{X}_i^t, \mathcal{B})$.
    Assign the pseudo-label using soft voting:

$$\hat{p}(\hat{X}_i^t) = \frac{1}{K} \sum_{j \in \mathcal{I}_i} p(\hat{X}_j^t), \quad \hat{y}_i^t = \arg\max_c \hat{p}^{(c)}(\hat{X}_i^t)$$

end for
Return: the pseudo-labeled dataset $\hat{D}^t = \{(\hat{X}_i^t, \hat{y}_i^t)\}_{i=1}^{\hat{n}_t}$.

3.6 Domain Adaptation with Target Domain Clustering Structure

Our approach generates pseudo-labels using the clustering structure of the target domain data, thereby improving the reliability of

the pseudo-labels. Additionally, we introduce a weighted classification loss and a prediction diversity loss to ensure the model can effectively distinguish between stego text and cover text, further enhancing the generalization ability and detection performance of the linguistic steganalysis model in the target domain. The details is described in the following.

3.6.1 Clustering-Based Pseudo-Label Generation. To generate pseudo-labels, we aggregate information from the nearest neighbor samples. The process of pseudo-label generation consists of two main steps: First, we use the pre-trained model to compute the confidence for each target domain sample X_i^t, defined as:

$$R(X_i^t) = \max_c p^{(c)}(X_i^t) \tag{7}$$

where $p^{(c)}(X_i^t)$ represents the predicted probability that sample X_i^t belongs to class c. The confidence score $R(X_i^t)$ is determined by selecting the highest probability among all classes.

Based on this confidence, we select the top $\hat{n}_t/2$ samples with the highest confidence from both cover texts and stego texts, where $\hat{n}_t = N \times P \times \text{Epoch}$, with N representing the total number of samples in the target domain and P being a parameter that controls the proportion of pseudo-label generation. A new dataset is then formed: $\hat{D}^t = \{\hat{X}_i^t\}_{i=1}^{\hat{n}_t}$ The feature representations $f(\hat{X}_i^t)$ and predicted distributions $p(\hat{X}_i^t)$ of these samples are stored in a memory bank $\mathcal{B}$, filtering out samples with lower confidence.

Next, we leverage the clustering structure of the target domain data to assign pseudo-labels to the samples in $\hat{D}^t$. We begin by calculating the Euclidean distance between the feature representation of each sample $\hat{X}_i^t$ and the samples in the memory bank. We then select the K nearest neighbors from $\mathcal{B}$ to provide a clustering context for each sample $\hat{X}_i^t$. Using a soft voting strategy, we generate the pseudo-label for each sample $\hat{X}_i^t$ by aggregating the predicted probability distributions of its K nearest neighbors:

$$\hat{p}(\hat{X}_i^t) = \frac{1}{K} \sum_{j \in \mathcal{I}_i} p(\hat{X}_j^t) \tag{8}$$

where $\mathcal{I}_i$ represents the index set of the nearest neighbors of $\hat{X}_i^t$, and $p(\hat{X}_j^t)$ is the predicted probability distribution of the j-th neighbor sample. Here, $\hat{p}(\hat{X}_i^t)$ denotes the aggregated probability distribution obtained through soft voting, which provides a more robust estimation of the true label of $\hat{X}_i^t$. Instead of relying solely on the potentially noisy individual prediction of $\hat{X}_i^t$, we leverage the consensus among its neighboring samples to refine the pseudo-label assignment.

Ultimately, the final pseudo-label is determined by selecting the class with the highest probability from $\hat{p}(\hat{X}_i^t)$, expressed as follows:

$$\hat{y}_i^t = \arg\max_c \hat{p}^{(c)}(\hat{X}_i^t) \tag{9}$$

where $\hat{p}^{(c)}(\hat{X}_i^t)$ represents the probability of class c in the aggregated distribution $\hat{p}(\hat{X}_i^t)$. Through this process, we obtain the pseudo-labeled dataset for the target domain: $\hat{D}^t = \{(\hat{X}_i^t, \hat{y}_i^t)\}_{i=1}^{\hat{n}_t}$ These pseudo-labels $\hat{y}_i^t$ serve as self-supervised signals to compute the standard classification loss for the target domain data. For clarity, we outline the key steps involved in generating pseudo-labels in Algorithm 1.

3.6.2 Loss Functions. Due to the distributional differences between the source and target domains, the boundary between the clustering structures of stego and cover texts is not clearly defined (as shown in the figure 2). As a result, $\hat{D}^t$ may still contain a certain number of erroneous pseudo-label samples. To mitigate this, we introduce a weighting mechanism based on the entropy of each sample's predicted distribution. This method helps reduce the impact of uncertain samples during training. The entropy of a sample $\hat{X}_i^t$ is calculated as follows:

$$H(\hat{X}_i^t) = -\sum_{c=1}^{C} p^{(c)}(\hat{X}_i^t) \log_2 p^{(c)}(\hat{X}_i^t) \tag{10}$$

where $p_c(\hat{X}_i^t)$ is the model's predicted probability for class c, and C is the total number of classes. Entropy quantifies the "uncertainty" or "disorder" in the prediction. If the prediction is concentrated, where one class probability is close to 1 and others are close to 0, the entropy is low, indicating high confidence. Therefore, if the prediction distribution is more uniform, the entropy is high, signaling uncertainty in the model's decision.

Next, we define the weight w_i of a sample based on its entropy $H(\hat{X}_i^t)$ as follows:

$$w_i = \exp(-H(\hat{X}_i^t)) \tag{11}$$

Using the negative exponential ensures that samples with lower entropy (more confident predictions) receive higher weights, while samples with higher entropy (less certain predictions) are given lower weights. Furthermore, the exponential function helps avoid excessively penalizing samples near the decision boundary, where predictions may show less consistency.

Finally, the Weighted Classification Loss function (WCL) is defined as:

$$\mathcal{L}_{WCL} = -\frac{1}{\hat{n}_t} \sum_{i=1}^{\hat{n}_t} w_i [\hat{y}_i^t \log(p(\hat{X}_i^t)) + (1 - \hat{y}_i^t) \log(1 - p(\hat{X}_i^t))] \tag{12}$$

By adjusting the weights based on the entropy of each sample, the model can focus more on confident samples during training, while minimizing the influence of uncertain ones.

Additionally, in order to effectively address the issue of posterior collapse in linguistic steganalysis, we introduce the Prediction Diversity Loss function (PDL). This loss function is defined as follows:

$$\mathcal{L}_{PDL} = -\frac{1}{\hat{n}_t} \sum_{i=1}^{\hat{n}_t} \sum_{c=1}^{C} p^{(c)}(\hat{X}_i^t) \log(p^{(c)}(\hat{X}_i^t)) \tag{13}$$

The prediction diversity loss encourages the model to maximize the entropy of the predicted class distribution, thereby promoting more balanced class predictions and preventing the model from becoming overly biased toward either cover text or stego text during training. By minimizing $\mathcal{L}_{PDL}$, the model learns a more comprehensive representation of the target domain data, thereby significantly improving the detection performance of stego texts.

Finally, we construct the domain adaptation objective function for the target domain data by jointly optimizing the weighted classification loss and the prediction diversity loss:

$$\mathcal{L}_t = \lambda \mathcal{L}_{WCL} + \mathcal{L}_{PDL} \tag{14}$$

Table 1: Comparative evaluation of different domain adaptation linguistic steganalysis of cross-datasets for different steganographic.

Stego	Methods	SF	$M \rightarrow N$		$M \rightarrow T$		$N \rightarrow M$		$N \rightarrow T$		$T \rightarrow M$		$T \rightarrow N$		Average	
			ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
VLC	MDA	*N*	84.62	84.51	83.33	83.25	86.38	86.63	82.38	82.03	85.42	85.15	84.38	84.56	84.41	84.36
	SANet	*N*	**86.58**	**86.55**	83.40	83.31	87.46	87.44	83.72	83.70	86.78	86.73	86.53	86.41	85.75	85.69
	Base	*Y*	80.75	80.53	81.65	81.65	87.13	87.22	82.58	82.52	86.50	86.44	84.55	84.52	83.86	83.81
	PDTS	*Y*	86.31	86.30	83.05	82.94	88.12	88.10	84.52	84.50	87.57	87.48	88.25	88.21	86.30	86.26
	CPSLS	*Y*	86.06	86.05	**83.95**	**83.91**	**89.45**	**89.45**	**84.71**	**84.62**	**90.95**	**90.91**	**89.11**	**89.06**	**87.54**	**87.50**
AC	MDA	*N*	84.38	84.23	78.12	78.38	88.54	88.30	82.29	82.33	88.54	88.53	84.38	84.00	84.38	84.30
	SANet	*N*	90.58	90.57	85.03	84.99	89.23	89.23	**85.42**	**85.41**	89.42	89.39	88.57	88.57	88.04	88.03
	Base	*Y*	85.85	85.81	82.40	82.39	88.44	88.21	83.02	83.00	87.25	87.25	86.20	86.20	85.53	85.48
	PDTS	*Y*	90.00	89.92	84.15	84.14	89.45	89.43	84.85	84.82	90.97	91.01	90.05	90.04	88.25	88.23
	CPSLS	*Y*	**91.03**	**91.02**	**85.21**	**85.17**	**91.19**	**91.14**	84.70	84.69	**91.84**	**91.77**	**91.05**	**91.03**	**89.17**	**89.14**
ADG	MDA	*N*	48.35	37.44	45.83	29.04	45.29	36.12	48.96	33.61	50.20	43.78	50.92	41.54	48.26	36.92
	SANet	*N*	50.43	36.60	50.68	45.35	50.43	44.51	50.67	45.23	51.17	45.49	51.40	49.21	50.80	44.40
	Base	*Y*	52.57	48.55	52.85	51.34	50.35	45.66	52.23	50.34	51.67	48.55	51.87	43.86	51.92	48.05
	PDTS	*Y*	53.05	52.88	**53.24**	53.12	51.80	51.64	52.26	52.19	52.95	52.82	54.04	53.97	52.89	52.77
	CPSLS	*Y*	**53.60**	**53.59**	53.18	**53.15**	**52.46**	**52.31**	**52.70**	**52.57**	**53.77**	**53.77**	**54.83**	**54.41**	**53.42**	**53.30**

where λ is a hyperparameter that balances the two loss terms. Through this joint optimization strategy, the model can fully leverage the clustering structure of the target domain data while effectively mitigating the posterior collapse issue during the optimization process, thereby enhancing the ability to distinguish between cover text and stego text in the target domain.

To further validate the effectiveness of the clustering structure in model training, we visualize the target domain feature space before and after domain adaptation, as shown in Figure 4. Compared to the feature distribution before adaptation shown in Figure 2, the clustering structure of stego and cover texts becomes more compact after applying the CPSLS method, with significantly enhanced intra-class cohesion and clearer decision boundaries between classes. This observation indicates that the clustering-based pseudo-labeling mechanism effectively guides the model to distinguish between stego and cover samples more accurately during training. The improved clustering structure encourages same-class samples to be more tightly grouped in the feature space, significantly reducing the overlap between different classes. As a result, the model is able to construct clearer decision boundaries, thereby improving detection performance in the target domain.

4 EXPERIMENTS

4.1 Experimental Setup

To comprehensively evaluate the proposed CPSLS method, we selected three widely used natural language corpora: Movie (M) [11], Twitter (T) [3], and News (N) [13]. During training, we generated cover text datasets using a pretrained statistical language model and created corresponding stego text datasets using three distinct steganographic algorithms: Variable Length Coding (VLC) [27], Arithmetic Coding (AC) [30], and Adaptive Dynamic Grouping (ADG) [29]. The embedding rates for VLC, AC, and ADG are 0.954, 0.939, and 0.965; 0.601, 0.572, and 0.545; and 1.765, 1.753, and 1.778, respectively, for the Movie, Twitter, and News corpora. For each corpus, we randomly selected 12,000 cover texts, which were split into training, validation, and test sets at a 10:1:1 ratio. Similarly, we generated 12,000 stego texts for each dataset and divided them into training, validation, and test sets with the same ratio. It is important to note that there is no overlap between the cover and stego texts across these datasets.

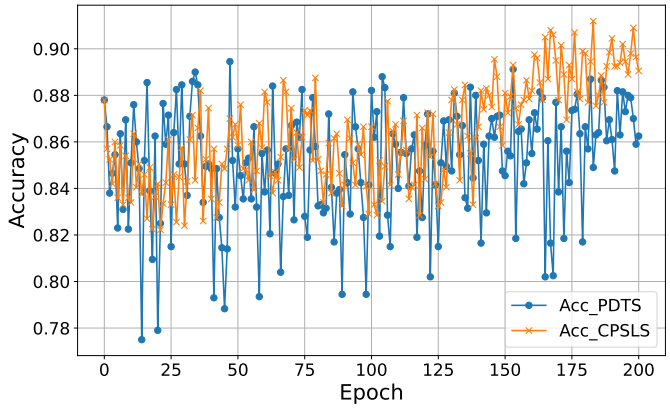
The CPSLS method utilizes a pre-trained BERT model [1] for word embeddings, consisting of 12 transformer encoder layers, each with a 768-dimensional hidden layer. To maintain BERT's feature representation capabilities, its parameters remain fixed during training. The Bi-LSTM [4] used in the model has 500 hidden units and a dropout rate of 0.5. Training is carried out using the Adam optimizer with an initial learning rate of 1×10^{-4}, and the exponential decay rates for beta1 and beta2 are set to 0.5 and 0.9, respectively. To select the optimal hyperparameters for the loss function, we tested the value of λ with 1, 2, 3, 4, 5 and found that $\lambda = 5$ gave the best performance, so we adopted $\lambda = 5$ as the final choice. For the pseudo-label selection ratio p, we tested 0.05, 0.10, 0.15, 0.20, 0.25 and found that $p = 0.10$ performed best, so we used this value in all experiments.

4.2 Comparison with Domain Adaptation in Linguistic Steganalysis

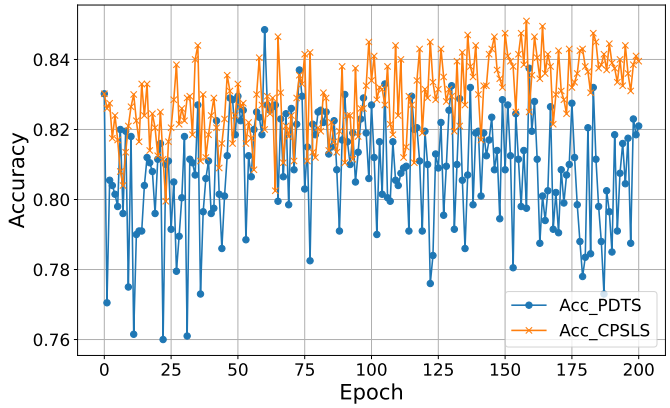
We compare the proposed CPSLS with current state-of-the-art cross-domain steganalysis models, including MDA [22], SANet [21], and

Table 2: Comparative evaluation of different source-free domain adaptation methods of cross-datasets for AC steganographic.

Stego	Methods	SF	$M \rightarrow N$		$M \rightarrow T$		$N \rightarrow M$		$N \rightarrow T$		$T \rightarrow M$		$T \rightarrow N$		Average	
			ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
AC	SHOT	*Y*	86.66	86.43	82.85	82.82	89.21	89.13	83.93	83.91	90.94	90.62	89.50	89.21	87.18	87.02
	G-SFDA	*Y*	88.95	88.73	84.29	84.13	89.75	89.42	84.46	84.45	90.67	90.58	90.04	90.02	88.03	87.89
	CPSLS	*Y*	**91.03**	**91.02**	**85.21**	**85.17**	**91.19**	**91.14**	**84.70**	**84.69**	**91.84**	**91.77**	**91.05**	**91.03**	**89.17**	**89.14**



(a) N→M



(b) N→T

Figure 5: Accuracy of Two Source-Free Domain Adaptation Methods for Linguistic Steganalysis in Two Cross-Domain Tasks under the AC Steganographic Algorithm. "N->M" represents a cross-domain task with small domain differences, while "N->T" represents a cross-domain task with significant domain differences.

PDTS [10]. MDA utilizes the MMD metric for domain feature alignment; SANet combines an adaptive weight selection network with the SDDM method for precise cross-domain feature adjustment; PDTS adopts a two-stage source-free text steganalysis approach. Additionally, we present the CPSLS model after the pre-training phase as "Base" for comparison. To evaluate the model's performance, we design six cross-domain tasks (source domain→target domain): $M \rightarrow N$, $M \rightarrow T$, $N \rightarrow M$, $N \rightarrow T$, $T \rightarrow M$, and $T \rightarrow N$, and use three steganographic algorithms: VLC [27], AC [30], and ADG [29]. The experimental results are summarized in the table, where "SF" indicates whether it is source-free domain adaptation, with "N" representing "No" and "Y" representing "Yes". The results are shown in Table 1.

The experimental results show that CPSLS achieves the highest average detection accuracy among all tested cross-domain steganalysis methods. It not only effectively addresses the challenges in SFDA scenarios but also demonstrates superior detection performance compared to existing advanced steganalysis algorithms. Specifically, for the three steganographic algorithms (VLC, AC, and ADG), CPSLS improves upon the best average detection accuracy of other methods by 1.24, 0.82, and 0.53 percentage points, respectively. Notably, under the AC algorithm, CPSLS achieves an average accuracy close to 90%. Compared to the source-dependent SANet method, CPSLS improves the average detection accuracy by 1.79, 1.13, and 2.62 percentage points on the three algorithms, respectively. This indicates that CPSLS can effectively learn steganographic features in the target domain, demonstrating stronger robustness and generalization capabilities. Furthermore, under the more adversarial ADG algorithm, despite an overall decline in detection performance, CPSLS still maintains a leading position. This suggests that CPSLS is capable of effectively learning steganographic features in the target domain, exhibiting enhanced robustness and generalization ability.

To further demonstrate the performance of CPSLS, we plotted the loss function curves for CPSLS and PDTS during training on the $N \rightarrow T$ (AC) task and the $N \rightarrow M$ (AC) task, as shown in Figure 5. From the figure, we observe that PDTS shows more fluctuation in detection accuracy (ACC), whereas CPSLS exhibits much less fluctuation. This is because PDTS uses samples with different confidence levels with equal weight, making it more susceptible to noise. In contrast, CPSLS leverages clustering-based pseudo-label generation, which enhances the accuracy of pseudo-labels. Additionally, by using a weighted classification loss function, CPSLS assigns lower weights to samples with lower confidence, reducing their impact on the model. This helps CPSLS effectively learn the steganographic features in the target domain, demonstrating stronger robustness and generalization capabilities.

Although CPSLS achieves the best performance in most tasks, it still shows certain performance gaps in a few cases. Specifically, in the $N \rightarrow T$ (AC) task, CPSLS yields an accuracy that is 0.72 percentage points lower than SANet. This gap is mainly attributed to the significant domain shift between the source domain (News) and the target domain (Twitter), particularly in terms of text length

and linguistic style. News texts are typically longer and more formally structured, whereas Twitter texts are shorter, more colloquial, and highly fragmented. Additionally, the perturbations introduced by the AC steganographic algorithm are relatively subtle, further reducing the separability between stego and cover texts in the feature space. Under these conditions, the clustering structure used by CPSLS for pseudo-label generation becomes less reliable, thereby affecting the model's ability to effectively learn steganographic features. In the $M \rightarrow T$ (ADG) task, CPSLS performs 0.06 percentage points lower than PDTS. This is primarily because the ADG steganographic algorithm introduces extremely slight perturbations, resulting in a high degree of overlap between cover and stego texts in the feature space. Moreover, Twitter data inherently exhibits considerable noise and weak semantic coherence, further complicating the pseudo-label generation process. Although CPSLS mitigates the influence of noisy samples through a confidence-based weighting mechanism, under the compounded challenges of domain shift and data noise, even a small number of pseudo-label errors can accumulate and negatively impact the training process, leading to a slight drop in detection performance.

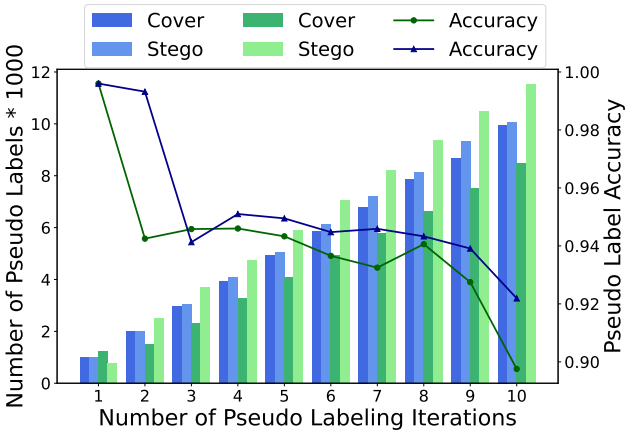
Additionally, in the $M \rightarrow N$ (VLC) task, CPSLS achieves 0.52 percentage points lower accuracy than SANet and 0.25 percentage points lower than PDTS. This is mainly because when the source and target domains share similar text styles and there exists a certain degree of separability between cover and stego texts, the pre-trained model M_s can better adapt to the target domain. However, the prediction diversity loss introduces greater uncertainty in classifying target domain samples, which may interfere with the model's learning process and result in performance degradation.

4.3 Comparison with Others Source-Free Domain Adaptation Methods

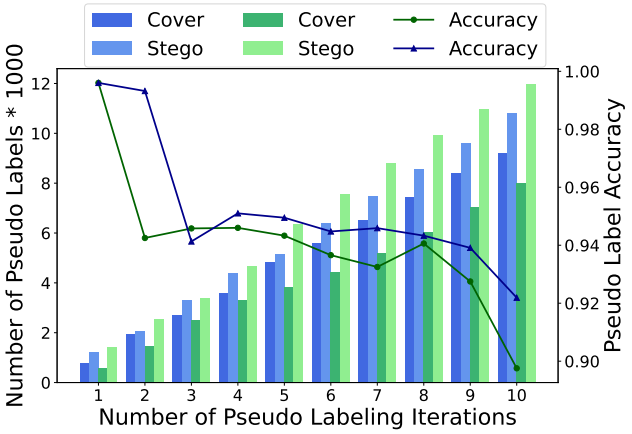
In this section, we compare the proposed CPSLS method with other widely used source-free domain adaptation methods, including G-SFDA [26] and SHOT [9], to evaluate its effectiveness. G-SFDA improves the compactness of the target domain data through local structure clustering, while SHOT employs information maximization and self-supervised pseudo-labeling to train a feature extraction module for the target domain.

To ensure a fair comparison, we use the same pre-trained model M_s as the baseline for the comparison methods. Then, we apply G-SFDA and SHOT as source-free domain adaptation methods to fine-tune M_s and evaluate the performance of these SFDA methods in six cross-domain linguistic steganalysis tasks using the AC steganographic algorithm, as described in the previous section. The experimental results are summarized in the table 2 .

The results demonstrate that CPSLS outperforms other source-free domain adaptation methods across various cross-domain linguistic steganalysis tasks. Specifically, compared to SHOT, our method improves the average detection accuracy and F1 score by 1.99 and 2.12 percentage points, respectively. Although SHOT also utilizes pseudo-labels for supervised training, the substantial distributional differences between the source and target domains in linguistic steganalysis introduce significant noise into the pseudo-labels. This noise hampers the model's ability to effectively learn steganographic features from the target domain and negatively impacts detection performance. In contrast, CPSLS addresses this issue by applying a weighted classification loss based on the uncertainty of the predicted distribution, thereby mitigating the adverse effect of noisy pseudo-labels during optimization.



(a) N→M



(b) N→T

Figure 6: Pseudo-label selection in two source-free domain adaptation methods under the AC steganographic algorithm for two cross-domain tasks. "N->M" indicates a task with small domain differences, and "N->T" represents a task with larger domain differences. Blue represents CPSLS, and green represents PDTS.

Compared to G-SFDA, CPSLS improves the average detection accuracy and F1 score by 1.14 and 1.25 percentage points, respectively. While G-SFDA enhances the compactness of the target domain distribution through local structure clustering, it does not utilize pseudo-labels to facilitate supervised training. This limitation restricts the model's capacity to fully capture steganographic features in the target domain. In contrast, CPSLS leverages the clustering structure of the target domain to generate more reliable pseudo-labels, enabling better learning of domain-specific features and significantly enhancing detection performance.

Table 3: Ablation experiment results across datasets for the AC steganographic algorithm.

Stego	Methods	$M \rightarrow N$		$M \rightarrow T$		$N \rightarrow M$		$N \rightarrow T$		$T \rightarrow M$		$T \rightarrow N$		Average	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
AC	w-DA	85.85	85.81	82.40	82.39	88.44	88.21	83.02	83.00	87.25	87.25	86.20	86.20	85.53	85.48
	w-CP	91.05	91.05	**85.55**	**85.55**	88.70	88.60	84.05	83.82	88.75	88.65	88.60	88.58	87.78	87.71
	w-WCL	90.70	90.70	84.45	84.44	90.87	90.84	**85.30**	**85.30**	**91.89**	**91.88**	90.50	90.49	88.95	88.94
	w-PDL	**91.85**	**91.85**	83.61	83.25	90.75	90.73	84.54	84.54	90.75	90.72	89.40	89.38	88.48	88.41
	CPSLS	91.03	91.02	85.21	85.17	**91.19**	**91.14**	84.70	84.69	91.84	91.77	**91.05**	**91.03**	**89.17**	**89.14**

4.4 Ablation Study

In this section, we present an ablation study of the proposed CPSLS method, evaluated on cross-domain linguistic steganalysis tasks using the AC steganographic algorithm, as described in the previous section. The CPSLS method consists of two key phases: pre-training with labeled source domain data and domain adaptation using the target domain's clustering structure. In the domain adaptation phase, we incorporate clustering-driven pseudo-label generation, the weighted classification loss function, and the prediction diversity loss function. We examine the contribution of each component individually, with the results summarized in the table 3.

Specifically, "w-DA" refers to the case where Domain Adaptation phase is not performed, "w-CP" refers to the case where clustering-based Pseudo-label generation is not used and an incremental method similar to PDTS is employed, "w-WCL" indicates that the Weighted Classification Loss function is not used and a standard cross-entropy loss function is applied instead, and "w-PDL" denotes the removal of the Prediction Diversity Loss function. The results clearly demonstrate that each component plays a crucial role in the overall effectiveness of CPSLS.

We then analyze the performance of the three main methods in the domain adaptation phase. w-CP involves generating pseudo-labels based on the target domain's clustering structure. Removing this component resulted in a 1.39 percentage point decrease in average accuracy and a 1.33 percentage point drop in F1 score, highlighting the importance of the clustering-based pseudo-label generation method. Although in the $M \rightarrow T$ task, the "w-CP" variant achieved 0.34 percentage points higher accuracy than CPSLS, this suggests that while clustering-based pseudo-labels may improve accuracy, they may encode less discriminative information. Nevertheless, more reliable pseudo-labels enable the model to better learn the steganographic features of the target domain, thereby enhancing overall detection performance. Our pseudo-label generation method, compared to the incremental approach used by PDTS, demonstrates greater stability, with higher reliability in each round of selection, as shown by the line chart in the figure 6.

Next, we examine the effect of sample weighting. After removing the weighted classification loss function, "w-WCL" exhibits an overall decrease of 0.22 percentage points in accuracy and 0.20 percentage points in F1 score. This decline is particularly pronounced in tasks with greater distributional differences between the source and target domains—for example, in the $M \rightarrow T$ task, where the detection accuracy drops by 0.76 percentage points. This indicates that the weighted classification loss effectively mitigates the negative impact of incorrect pseudo-labels on the model's ability to learn the target domain's steganographic features. In the $N \rightarrow T$ and $\rightarrow M$ tasks, "w-WCL" outperforms CPSLS, suggesting that although the weighted classification loss reduces the negative impact of uncertain samples during model optimization, these samples also provide additional information to the model, helping it avoid overfitting to simpler samples and thereby better learn the target domain's features. Lastly, removing the prediction diversity loss function resulted in a 0.69 percentage point decrease in accuracy and a 0.73 percentage point drop in F1 score. This decline is attributed to the subtle differences between stego and cover texts, which can lead to posterior collapse during optimization, preventing the model from learning the full range of target domain features.

The use of the diversity loss function effectively prevents posterior collapse, resulting in a more balanced selection of pseudo-labels, as shown in figure 6. This balanced pseudo-label set ensures that the model is exposed to a wider variety of samples during training, helping to avoid overfitting to one category or neglecting the features of the other category. As a result, the model learns more comprehensive data features and improves its ability to distinguish between the two types of texts. In the $M \rightarrow N$ task, the detection rate of "w-PDL" is 0.82% higher than CPSLS. This is because the distributional difference between the Movie and News corpora is relatively small, and the pre-trained model is already able to classify target domain samples accurately. However, the prediction diversity loss function increases the uncertainty of the model's classification of target domain samples, which affects the effective learning from those samples.

Finally, by comparing results with and without domain adaptation, we observe a 3.64% improvement in average accuracy and a 3.66% increase in F1 score after domain adaptation. These results clearly demonstrate that the CPSLS method effectively mitigates the distributional differences between source and target domains in source-free domain adaptation scenarios for linguistic steganalysis, significantly enhancing the model's detection capabilities.

5 CONCLUSION

To improve the detection performance of the model in source-free domain adaptation scenarios, this paper proposes the Clustering-driven Pseudo-labeling method for Source-free Domain Adaptation in Linguistic Steganalysis (CPSLS). The method consists of two

main phases: pre-training with labeled source domain data and domain adaptation using the target domain's clustering structure. In the pre-training phase, the model is trained using labeled source domain data to provide an initial understanding of the target domain data. During the domain adaptation phase, we proposed three key techniques: clustering-driven pseudo-label generation, a weighted classification loss function, and a prediction diversity loss function. These techniques help the model effectively learn the steganographic features of the target domain without relying on source domain data, mitigating the impact of distribution differences between the source and target domains, and improving detection performance in the target domain. Extensive experimental results demonstrate that CPSLS not only has stronger practical applicability by functioning in source-free domain adaptation scenarios, but also outperforms existing cross-domain linguistic steganalysis methods in terms of detection performance.

In current cross-domain linguistic steganalysis research, adaptive model training typically relies on rich labeled source domain data, including cover and stego texts. However, obtaining such a large amount of labeled source domain data is often difficult in real-world scenarios. Therefore, future research will focus on a more practical scenario where only a small amount of labeled source domain data is available in the pre-training phase. We aim to explore effective strategies for unsupervised domain adaptation under these constraints, to improve the model's adaptability and discriminative ability for new domain data, while also addressing data privacy concerns.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U21B2020).

References

[1] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Liting Gao, Juan Wen, Ziwei Zhang, Guangying Fan, and Yinghan Zhou. 2024. Linguistic Steganalysis via Probabilistic Weighted Contrastive Learning. *IEEE Signal Processing Letters* (2024).

[3] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[4] Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

[5] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[6] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. 2020. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4544–4553.

[7] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. 2020. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12376–12385.

[8] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. 2020. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9641–9650.

[9] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*. PMLR, 6028–6039.

[10] Yufei Luo, Zhen Yang, Ru Zhang, and Jianyi Liu. 2024. Pseudo-label Based Domain Adaptation for Zero-Shot Text Steganalysis. *arXiv preprint arXiv:2406.18565* (2024).

[11] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.

[12] Wanli Peng, Jinyu Zhang, Yiming Xue, and Zhenghong Yang. 2021. Real-time text steganalysis based on multi-stage transfer learning. *IEEE Signal Processing Letters* 28 (2021), 1510–1514.

[13] A. Thompson. [n. d.]. Kaggle. https://www.kaggle.com/snapcrack/all-the-news/data. 2017.

[14] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726* (2020).

[15] Yihao Wang, Ru Zhang, and Jianyi Liu. 2023. RLS-DTS: Reinforcement-learning linguistic steganalysis in distribution-transformed scenario. *IEEE Signal Processing Letters* (2023).

[16] Juan Wen, Liting Gao, Guangying Fan, Ziwei Zhang, Jianghao Jia, and Yiming Xue. 2023. SCL-Stega: Exploring Advanced Objective in Linguistic Steganalysis using Contrastive Learning. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*. 97–102.

[17] Juan Wen, Xuejing Zhou, Ping Zhong, and Yiming Xue. 2019. Convolutional neural network based text steganalysis. *IEEE Signal Processing Letters* 26, 3 (2019), 460–464.

[18] Haifeng Xia, Handong Zhao, and Zhengming Ding. 2021. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9010–9019.

[19] Qiong Xu, Ru Zhang, and Jianyi Liu. 2023. Linguistic steganalysis by enhancing and integrating local and global features. *IEEE Signal Processing Letters* 30 (2023), 16–20.

[20] Yiming Xue, Lingzhi Kong, Wanli Peng, Ping Zhong, and Juan Wen. 2022. An effective linguistic steganalysis framework based on hierarchical mutual learning. *Information Sciences* 586 (2022), 140–154.

[21] Yiming Xue, Jiaxuan Wu, Ronghua Ji, Ping Zhong, Juan Wen, and Wanli Peng. 2023. Adaptive domain-invariant feature extraction for cross-domain linguistic steganalysis. *IEEE Transactions on Information Forensics and Security* (2023).

[22] Yiming Xue, Boya Yang, Yaqian Deng, Wanli Peng, and Juan Wen. 2022. Domain adaptational text steganalysis based on transductive learning. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security*. 91–96.

[23] Hao Yang, YongJian Bao, Zhongliang Yang, Sheng Liu, Yongfeng Huang, and Saimei Jiao. 2020. Linguistic steganalysis via densely connected LSTM with feature pyramid. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*. 5–10.

[24] Jinshuai Yang, Zhongliang Yang, Jiajun Zou, Haoqin Tu, and Yongfeng Huang. 2022. Linguistic steganalysis toward social network. *IEEE Transactions on Information Forensics and Security* 18 (2022), 859–871.

[25] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. 2020. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427* 1, 2 (2020), 5.

[26] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. 2021. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8978–8987.

[27] Zhong-Liang Yang, Xiao-Qing Guo, Zi-Ming Chen, Yong-Feng Huang, and Yu-Jin Zhang. 2018. RNN-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security* 14, 5 (2018), 1280–1295.

[28] Shaokai Ye, Kailu Wu, Mu Zhou, Yunfei Yang, Sia Huat Tan, Kaidi Xu, Jiebo Song, Chenglong Bao, and Kaisheng Ma. 2020. Light-weight calibrator: a separable component for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13736–13745.

[29] Siyu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2021. Provably secure generative linguistic steganography. *arXiv preprint arXiv:2106.02011* (2021).

[30] Zachary M Ziegler, Yuntian Deng, and Alexander M Rush. 2019. Neural linguistic steganography. *arXiv preprint arXiv:1909.01496* (2019).

[31] Jiajun Zou, Zhongliang Yang, Siyu Zhang, Sadaqat ur Rehman, and Yongfeng Huang. 2020. High-performance linguistic steganalysis, capacity estimation and steganographic positioning. In *International Workshop on Digital Watermarking*. Springer, 80–93.