Uncovering and Mitigating the Hidden Chasm: A Study on the Text-Text Domain Gap in Euphemism Identification

Yuxue Hu[1,2,3,4], **Junsong Li**[1], **Mingmin Wu**[1], **Zhongqiang Huang**[1], **Gang Chen**[5], **Ying Sha**[1,2,3,4*]

[1] College of Informatics, Huazhong Agricultural University, Wuhan, China
[2]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China
[4] Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
[5] Jointown Healthcare Technology Group, Wuhan, China
{hyx, shaying}@mail.hzau.edu.cn, {kikuss, wmm_nlp, hzq}@webmail.hzau.edu.cn, chengang@jk998jt.wecom.work

Abstract

Euphemisms are commonly used on social media and darknet marketplaces to evade platform regulations by masking their true meanings with innocent ones. For instance, "weed" is used instead of "marijuana" for illicit transactions. Thus, euphemism identification, i.e., mapping a given euphemism ("weed") to its specific target word ("marijuana"), is essential for improving content moderation and combating underground markets. Existing methods employ self-supervised schemes to automatically construct labeled training datasets for euphemism identification. However, they overlook the text-text domain gap caused by the discrepancy between the constructed training data and the test data, leading to performance deterioration. In this paper, we present the text-text domain gap and explain how it forms in terms of the data distribution and the cone effect. Moreover, to bridge this gap, we introduce a feature alignment network (FA-Net), which can both align the in-domain and cross-domain features, thus mitigating the domain gap from training data to test data and improving the performance of the base models for euphemism identification. We apply this FA-Net to the base models, obtaining markedly better results, and creating a state-of-the-art model which beats the large language models.

Introduction

Euphemism, a significant form of linguistic communication, refers to the use of gentle and implicit expressions to convey indirect information. Now, euphemisms are widely employed on social media (Zhu et al. 2021; Yuan et al. 2018) and darknet marketplaces (Li et al. 2021; Foye et al. 2021) to mask the true meaning of undesirable information, such as discrimination, hatred, and illicit transactions. Table 1 illustrates the use of euphemisms to circumvent monitoring for illegal activities. For instance, the euphemism "ice" in Table 1 was used as a substitute for the target keyword "methamphetamine". These euphemisms can seem innocent and vague, making it challenging to trace illicit transactions. Therefore, euphemism identification, which identifies the target keyword of a given euphemism, is essential for improving content moderation and combating underground markets. However, euphemisms are continually evolving like a 'treadmill' (Pinker 2003), making it difficult to maintain an up-to-date corpus for training the euphemism identification task. Furthermore, the use of euphemisms in either a literal or figurative sense adds complexity to the task.

Example sentences (euphemisms are in bold)
1. For all vendors of **ice**, it seems pretty obvious that it is not as pure as they market it.
2. Back up before I pull my **nine** on you.
3. I'm looking for the **girlfriend experience**, without having to deal with an actual girlfriend.

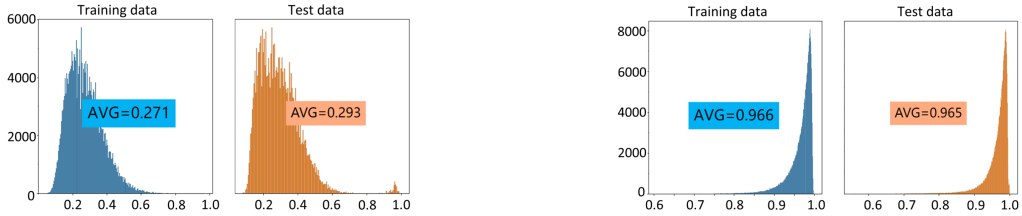
Table 1: Examples of sentences containing euphemisms.

(a) t-SNE visualization (b) UMAP visualization

Figure 1: Representation distribution of training and test data using selfEDI on Drug by (a) t-SNE visualization and (b) UMAP visualization.

Due to the challenges above, there are only a few works on euphemism identification. The work by Yuan et al. (2018) pioneered the task of euphemism identification. However, they focused on identifying the hypernyms of euphemisms rather than directly identifying the specific meanings of them. The most relevant work was reported by Zhu et al. (2021), who explicitly proposed euphemism identification task for the first time. They developed a self-supervised framework to construct labeled training datasets while using the sentences with target keywords masked out as samples and the corresponding target keywords as labels. This ingenious idea provides a new way to solve this similar problem.

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) SelfEDI (b) Bert-base-uncased

Figure 2: The cosine similarities of the training and test data on (a) SelfEDI model and (b) Bert-base-uncased model. Blue is for the training data and orange is for the test data.

However, they ignored the text-text domain gap between the constructed training data and the test data. Liang et al. (2022) present the modality gap in multi-modal models and provide insight into the cone effect to explain the modality gap. The cone effect means that the effective embedding space is restricted to a narrow cone. Inspired by them, we observe that the single-modal gap also exists in the euphemism identification task, which we call the text-text domain gap. As shown in Figure 1, the constructed training data (source) embeddings and test data (target) embeddings are located in two completely separate regions of the embedding space (Figure 1(a)) and each of the source and target text forms a distinctively different cone (Figure 1(b)). Meanwhile, we find that the accuracy of the test data is far below that of the training and validation data during the experiment. We give a two-part explanation for the text-text domain gap: 1) different datasets lead to different data distributions; 2) different data distributions create different embedding cones.

To bridge the text-text domain gap in euphemism identification, we propose a feature alignment network (FA-Net), which can mitigate the gap and enhance the generalization ability of the base models. Specifically, FA-Net consists of three sub-networks: an in-domain contrastive learning network (IC-net), a hard cross-domain contrastive learning network (HC-net), and an entropy minimization network (EM-net). IC-net employs an instance-wise contrastive learning objective to coalesce the same class and delineate different classes of samples, making representation in the source domain more discriminative. Simultaneously, HC-net uses hard cross-domain contrastive learning to align the nearest positive samples from the source and target data, reducing the discrepancy between them. EM-net jointly minimizes the entropy of the prediction, disambiguating the different class instances over the unlabeled data. By combining the above three sub-networks, FA-Net can ensure the in-domain category alignment and cross-domain feature alignment, thereby effectively reducing the text-text domain gap and improving the accuracy of euphemism identification. Our proposed model yields top1 accuracies that are 30.3%-75% higher than the SOTA baselines.

The main contributions are summarized as follows:

- To the best of our knowledge, we are the first to demonstrate a text-text domain gap phenomenon for NLP on euphemism. We explain this domain gap in terms of the data distribution and the cone effect.
- To bridge the text-text domain gap, we propose a feature alignment network, named FA-Net. The FA-Net jointly introduces instance-wise contrastive learning for in-domain category alignment, uses hard cross-domain contrastive learning for cross-domain feature alignment, and employs entropy minimization to obtain more confident predictions.
- Extensive results show that our FA-Net can improve the prediction accuracy of the base models and establish a new state of the art in euphemism identification. Notably, our approach outperforms the current best large language models (LLMs) on three large-scale benchmark datasets.

Related Work

Euphemism Identification. Existing models mainly focused on exploring whether the given word is a euphemism, using conventional NLP techniques (Magu and Luo 2018; Felt and Riloff 2020), deep learning (Yuan et al. 2018; Gavidia et al. 2022), and pre-trained models (Zhu et al. 2021; Zhu and Bhat 2021; Ke, Chen, and Wang 2022). Yuan et al. (2018) focused on identifying the hypernyms of euphemisms while not directly identifying the specific meanings of euphemisms. They identify "horse" as an illicit drug rather than heroin. Zhu et al. (2021) first proposed euphemism identification task. They developed a self-supervised scheme constructing labeled training datasets and used the bag-of-words model to analyze euphemisms in their sentence-level context, identifying each euphemism to the corresponding target keyword. Nevertheless, they overlook the text-text domain gap from the constructed training data to test data, leading to poor performance.

Domain Gap. Domain gap has been extensively studied in the field of machine learning and domain adaptation. When the distributions of training data and test data are different, the performance of models often deteriorates due to domain gaps (Candela et al. 2009). Existing work shows that modal gap is common in multi-modal models (Radford et al. 2021; Xu et al. 2021; Zhang et al. 2022). Liang et al. (2022) exhibit the generality of the modality gap phenomenon across a wide spectrum of multi-modal models and explain the cause of the modality gap. Meanwhile, work on domain gap mainly appear in computer vision field which uses Maximum Mean Discrepancy (MMD) approaches (Long et al. 2015; Zhang et al. 2019; Lu et al. 2021) or adversarial optimization objective (Ganin and Lempitsky 2015; Ma, Zhang,

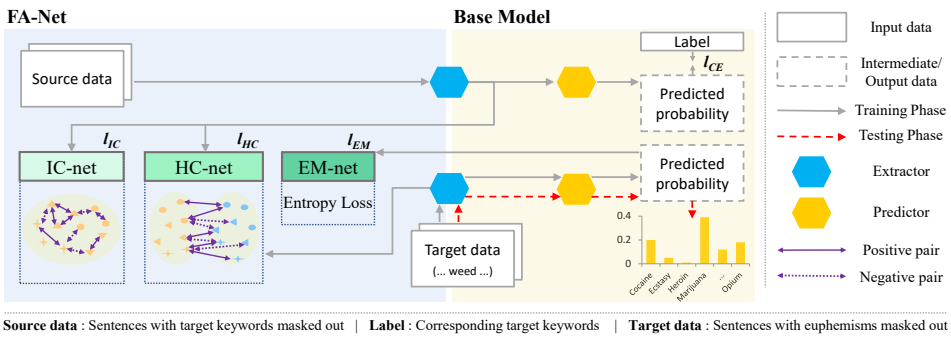


Figure 3: The architecture of FA-Net and base model for euphemism identification. The left block shows our FA-Net and right block shows the base model for euphemism identification. Different blocks, lines and nets are marked in the rightmost. The illustration of the inputs is in the bottom.

and Xu 2019; Cui et al. 2020) to mitigate the distribution divergence or aligns each category distribution between the source and target domains. In this paper, we find that the model gap also appears in a single-modal, and we use a contrastive learning and entropy minimization based method to mitigate the text-text domain gap.

The Text-Text Domain Gap on Euphemism

As euphemisms evolve like a "treadmill" (Pinker 2003), there is no publicly available curated dataset sufficient to cover the growing list of mappings between euphemisms and their target keywords. Existing methods address the challenges identified above via a self-supervised learning scheme. In the training and validation phase, they take the sentences masking the target keywords (e.g., cocaine and heroin) as training samples, using the corresponding target keywords as labels for training. During the testing phase, they feed the sentences with the euphemisms masked into the trained model and finally specify the masked euphemism into the corresponding target keyword.

Sentences containing euphemisms tend to use vague or figurative vocabulary and complex syntactic structures, while sentences containing target keywords use simple, direct syntactic structures to convey information. They can be considered to belong to different discourse systems, and exist in different representation spaces. Each of them has unique word formation methods, syntactic structures, and writing styles, but they are approximately isomorphic, conveying similar topics in similar contexts. In this manner, the source text (training data) is completely different from the target text (test data) in euphemism identification, while different data will lead to different data distributions, as shown in Figure 1(a).

As mentioned above, the source text and target text have different data distributions. For the text-text domain gap to exist, embeddings from either the source text or target text must be centered around a subregion, or the embeddings will overlap. (Liang et al. 2022) claimed that the modality gap has appeared in random model initialization due to the cone effect. Inspired by them, we randomly extract 512 embeddings from the last layer of the pre-trained models SelfEDI (Zhu et al. 2021) and bert-based-uncased[1], respectively. We then calculate the cosine similarity between all possible embedding pairs between the training data for each model and between the test data respectively (Figure 2). We found that the average cosine similarity of the source text and target text on the two models is 0.271, 0.293, 0.966, 0.965, and the minimum cosine similarity is 0.045, 0.081, 0.606, 0.553, all positive values. In 2D space, these results imply that all embeddings are restricted to a fan-shaped surface of less than 90°, in fact, these embeddings are very high-dimensional (863 for SelfEDI, 768 for Bert-base-uncased).

Taking the cosine similarity of 0.271 as an example, in 2D, arccos(0.271)=74.28°, indicating that the cosine similarity of 0.271 can "occupy" 74.28°/360°=20.63% of the 2D unit circle. In 3D, it can "occupy" $\frac{2\pi r^2(1-cos\frac{74.28^\circ}{2})}{4\pi r^2} = 10.14\%$. In 863D, the cosine similarity of 0.271 can "occupy" a surface area of less than $\frac{1}{2^{863}}$ in the unit 863D hypersphere, suggesting that the effective embedding space is restricted to an extremely narrow cone. In this way, the source and target text of different distributions each form a narrow cone, forming a gap between them, as shown in Figure 1(b). It shows that the domain gap also occurs in a single-modal, which is not captured in previous work.

Feature Alignment Network

The overall framework of our proposed FA-Net and the base model for euphemism identification is shown in Figure 3, where we take the Bert-based method as the base model. The FA-net consists of three parts: 1) the in-domain contrastive learning network (IC-net), 2) the hard cross-domain contrastive learning network (HC-net), and 3) the entropy minimization network (EM-net). By combining the above three sub-networks, the FA-Net can ensure the in-domain feature alignment of the same category and cross-domain feature alignment, thereby effectively reducing the cross-domain gap. Next, we use E_Bert, which was pre-trained on the benchmark dataset using the bert-base-uncased as the basis[1], as an example feature extractor to detail each component of the proposed FA-Net.

[1]https://huggingface.co/bert-base-uncased/

IC-net Module In the source domain, there are different categories y^s of target keywords with similar semantics (e.g., cocaine, and heroin), and the same category of target words contains multiple subcategories (e.g., mdma and molly share the same category). To better mitigate the domain gap, first of all, we need to align the in-domain features by clustering the sentence features of the same category and separating the features of different categories, making representation in the source domain more discriminative.

Inspired by He et al. (2020), we implement an in-domain contrastive learning network by using instance-wise contrastive learning. Given a batch of samples containing target keywords from the source domain D^s, we first obtain the representation for these samples according to the extractor of the base model. To make the representation more discriminative, we then employ an instance-wise contrastive learning objective to cluster the same class and separate samples of different classes, the objective function is calculated as:

$$h_i = \texttt{E\_Bert}(x_i) \quad x_i \in D^s/D^t \tag{1}$$

$$L^s_{IC} = -\frac{1}{|B^s|}\sum_{i=1}^{|B^s|}\sum_{j=1}^{|B^s|}\mathbb{I}(y^s_i = y^s_j)\log\frac{e^{sim(h^s_i,h^s_j)/\tau}}{\sum_{k=1}^{|B^s|}e^{sim(h^s_i,h^s_k)/\tau}} \tag{2}$$

where D^t is the target domain, $|B^s|$ is the batch size of source domain, $\mathbb{I}$ is an indicator, $sim(h_i, h_j)$ is the cosine similarity $\frac{h_i^T \cdot h_j}{|h_i||h_j|}$ and τ is the temperature hyper-parameter.

HC-net Module As mentioned earlier, the domain gap occurs between the source and target data due to the data distribution and cone effect. Although the source data and target data belong to different discourse systems and exist in different representation spaces, they are approximately isomorphic (Barone 2016), conveying similar topics in similar contexts. Furthermore, the source and target domains share the same set of labels. We hypothesize that samples in the same category are close to each other while samples from different categories lie far apart, regardless of which domain they come from. Thus, we introduce a hard cross-domain contrast learning module to learn the isomorphism information between the source domain and target domain, so as to align the cross-domain features and reduce the domain gap.

Specifically, given an anchor sample in the target domain, it forms a positive pair with a sample from the source domain. Here, we determine the nearest samples as the positive pairs measured by cosine similarity. We set the similarity threshold γ and regard the sample from the source domain whose cosine similarity with the anchor is greater than the threshold γ as a positive sample. We formulate the hard cross-domain contrastive loss as:

$$L^{t\to s}_{HC} = -\frac{1}{|B^t|}\sum_{i=1}^{|B^t|}\sum_{j=1}^{|B^s|}\mathbb{I}(sim(h^t_i,h^s_j)>\gamma) \log\frac{e^{sim(h^t_i,h^s_j)/\tau}}{\sum_{k=1}^{|B^s|}e^{sim(h^t_i,h^s_k)/\tau}} \tag{3}$$

Datasets	Entries	Num	Key_Entries
Drug	1271907	33	drug:4566
Weapon	3108988	9	weapon:19003
Sexuality	2894869	12	sexuality:7215

Table 2: Overview of the datasets. Num means categories of target keywords. Key_Entries means entries containing the target keywords.

The cross-domain loss forces intra-class distance to be smaller than inter-class distance from different domains to reduce the domain gap. Alternatively, we can also use source samples as anchors and compute $L^{s\to t}_{HC}$ loss as:

$$L^{s\to t}_{HC} = -\frac{1}{|B^s|}\sum_{i=1}^{|B^s|}\sum_{j=1}^{|B^t|}\mathbb{I}(sim(h^s_i,h^t_j)>\gamma) \log\frac{e^{sim(h^s_i,h^t_j)/\tau}}{\sum_{k=1}^{|B^t|}e^{sim(h^s_i,h^t_k)/\tau}} \tag{4}$$

where $|B^t|$ is the batch size of the target domain, the hard cross-domain contrastive loss can be denoted as follows,

$$L_{HC} = \alpha L^{t\to s}_{HC} + (1-\alpha)L^{s\to t}_{HC} \tag{5}$$

α is a learnable parameter to control the weight of $L^{t\to s}_{HC}$ and $L^{s\to t}_{HC}$.

EM-net Module For the source domain, since it is automatically constructed with labels by a self-supervised learning scheme, it's easy to train an optimal predictor. While for the target domain without labels, we can still align the optimal predictor by minimizing the entropy of the model's prediction (Ahuja et al. 2020; Wang et al. 2020). We minimize the entropy loss on unlabeled target domain instances, which widens the margin between positive and negative clusters, enhancing the likelihood that the optimal decision boundary for the source domain falls within it. In this way, EM-net can promote the convinced positive pairs while suppressing the unconvinced positive pairs in the HC-net module and also disambiguate the positive and negative instances in the predictor for euphemism identification.

In the predictor module of the base model for euphemism identification, each candidate target word is scored and the probability of the selected target word in the context of the given mask is calculated and denoted as $P(T_i)$, T refers to the target keyword candidate set. The loss of EM-net on unlabeled target domain is as follows:

$$L^t_{EM} = -\frac{1}{|B^t|}\sum_{i=1}^{|B^t|}P(T_i)\log P(T_i) \tag{6}$$

However, our model performs not well when applying the entropy loss starting from the first epoch. This is probably because the model needs to get a general picture of the source and target domains with contrastive learning in the first place and draw the decision boundary for the source domain. Employing the entropy loss too precipitously requires

	Drug			Weapon			Sexuality		
Method	**Acc@1**	**Acc@2**	**Acc@3**	**Acc@1**	**Acc@2**	**Acc@3**	**Acc@1**	**Acc@2**	**Acc@3**
Word2Vec	0.07	0.14	0.21	0.10	0.27	0.40	0.17	0.22	0.42
SelfEDI	0.20	0.31	0.38	0.33	0.51	0.67	0.32	0.55	0.64
LSTM	0.11	0.21	0.30	0.20	0.43	0.61	0.24	0.40	0.48
LSTM+FA	0.13	0.22	0.30	0.24	0.43	0.60	0.28	0.48	0.56
CNN	0.12	0.21	0.32	0.23	0.46	0.63	0.25	0.45	0.55
CNN+FA	0.15	0.24	0.33	0.27	0.46	0.66	0.31	0.50	0.56
Bert	0.24	0.31	0.40	0.38	0.55	0.73	0.38	0.50	0.69
Bert+FA	**0.35**	**0.39**	**0.42**	**0.47**	**0.60**	**0.77**	**0.50**	**0.62**	**0.75**

Table 3: Experimental results of our FA-Net against baselines. X+FA is a method obtained by our FA-Net using X as the base model for euphemism identification. The Word2Vec method uses static embeddings to predict the word with the highest similarity as the target keyword, and the SelfEDI method uses the bag of words model, which needs to build different dictionaries for different datasets. Neither of the two methods can be directly extended with FA-Net.

the model to early decide labels for uncertain instances before the model has learned good representations and the classification boundary on the source domain. Therefore, we apply the entropy loss from the latter epoch, which is detailed in the hyper-parameter analysis part in Section Experiment.

Experiment

Experimental Setup

Datasets We empirically validated our proposed model on three separate datasets: Drug, Weapon, and Sexuality (Zhu et al. 2021). These datasets are sourced from the Reddit website[2], Gab social networking services[3], Online Slang Dictionary[4], etc., and contain three types of sentences: sentences containing target words, sentences containing euphemisms, and sentences without euphemisms and target words. An overview of each dataset is shown in Table 2. There are 33, 9, and 12 categories of target keywords corresponding to datasets Drug, Weapon, and Sexuality respectively (target keywords under each category hold the same meaning).

As with existing methods, we employ a self-supervised learning framework to construct the training data. When training the model, the training and validation data must mask out the target keywords. When testing, the test data must mask out the euphemisms. Therefore, we require two kinds of inputs: 1) sentences from the original text corpus that mask out the target keywords (for training/validation) and sentences that mask out the euphemisms (for training/testing), and 2) a list of target keywords (e.g., heroin, cocaine, etc.). To evaluate the results, we need to rely on a ground truth list (Zhu et al. 2021) of euphemisms and the corresponding target keywords, which should contain a one-to-one mapping from each euphemism to its true meaning. The ground truth list of Drug was compiled by the Drug Enforcement Administrator, which is intended as a practical reference for law enforcement personnel (Administration et al. 2018). The ground truth list of Weapon was obtained from the Online Slang Dictionary[5], Slangpedia, and the Urban Thesaurus[6]. The ground truth list of Sexuality was obtained from The Online Slang Dictionary. Due to the rapidly evolving language used on the social network, it cannot be comprehensive or error-free, but it is the most reliable ground truth available to us.

Note that no additional resources or supervision is required during training, and the ground truth lists do not participate in the whole training process but are only used to help evaluate the accuracy of euphemism identification.

Baselines Since our goal is to perform feature alignment, to verify the validity of the proposed FA-Net model, we compare the results with and without FA-Net using the following base models, including the method proposed by Zhu et al. (2021) (the current best model, denoted as "SelfEDI"), the Word2vec baseline they created, and CNN, LSTM and Bert based models established by us.

- **Word2vec**: Use cosine similarity to select the word closest to its target by obtaining Word2Vec embeddings (100 dimensions) for all words.
- **SelfEDI**: Use the bag of words model to extract features from sentence-level contexts, and train a multinomial logistic regression classifier to identify euphemisms.
- **LSTM**: Use the long short-term memory network to extract features, followed by a classifier to identify euphemisms.
- **CNN**: Use convolution neural networks to extract features, followed by a classifier to identify euphemisms.
- **Bert**: Fine-tune the E_Bert which is obtained on a specific corpus to identify euphemisms.

Implementation Details To exclude other factors from affecting the comparison with the baselines, we also trained the models separately on each dataset and split the training set and validation set in an 8:2 ratio of sentences that mask

[2]https://www.reddit.com/

[3]https://gab.com/

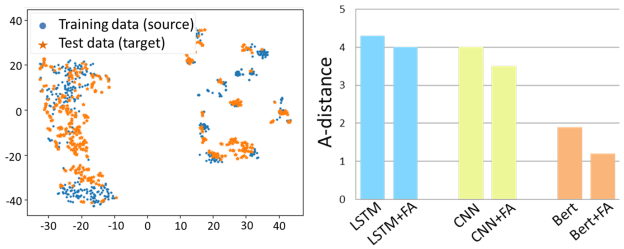
[4]https://slangpedia.org/

[5]http://onlineslangdictionary.com/

[6]http://urbanthesaurus.org/

Model	Drug	Weapon	Sexuality	Cost/S
StableLM[7]	0.02	0.03	0.12	2.08S/0.00475$
mPLUG[8]	0.02	0.13	0.15	2.35S/0.00541$
Llama2[9]	0.17	-	-	18.23S/0.05833$
GPT3.5[10]	0.33	0.17	0.42	1.12S/0.00035$
Bert+FA	**0.35**	**0.47**	**0.50**	**0.29S/0.00003$**

Table 4: Experimental results of our FA-Net against large language models (LLMs). Cost/S represents the average time and cost per sentence. "-" means that the models refuse to answer such questions involving inappropriate content.



(a) Representation distribution (b) Domain discrepancy

Figure 4: (a) Representation distribution of training and test data using Bert+FA on Drug; (b) the domain discrepancy ($\mathcal{A}$-distance) between the training and test data.

out the target keywords, while the test set comprised all sentences that mask out the euphemisms. All experiments were conducted on a Linux server of Ubuntu 18.0.4 LTS version with a Tesla-V100 32G GPU. For each feature extractor of the base models, we have the following configuration:

1) For the LSTM baseline, we use a two-layer LSTM, and the hidden state of each layer is set to 256.

2) For the CNN baseline, to obtain more fine-grained features, we use filter sizes of [1, 3, 5] with 100 feature maps.

3) For the Bert baseline, we fine-tuned the E_Bert parameters, which was pre-trained on bert-base-uncased[1] for MLM task only. During pre-training, the maximum length of the input sequence was set as 512, the batch size as 64, and the number of iterations as 3. For model training, the maximum length of the input sequence was 128, and the batch size was 32. The initial learning rate was 5e-5, the warm-up step was 1000, and the optimizer AdamW (Loshchilov and Hutter 2018) was based on a warm-up linear schedule.

Evaluation Metrics Similar euphemisms that refer to target keywords with similar semantics make it difficult to pinpoint the target keywords. For each euphemism, we generate a probability distribution over all target keywords. To this end, we evaluate the output using the metric precision at Acc@k, that is, the accuracy of whether the first k values of the sorted list generated by us include the true label. To be consistent with the current best model (Zhu et al. 2021), we use Acc@1, Acc@2, and Acc@3 to measure the results.

Experimental Results

Comparison with Baselines Table 3 summarizes the euphemism identification results (the top two rows are taken directly from Zhu et al. (2021)). To be fair, the results of all models are taken from the parameters that make the results the best. Our proposed Bert+FA model achieves the best performance. Specifically, Bert+FA outperforms the state-of-the-art model (SelfEDI) by 15%, 14%, and 18% in top1 accuracy values on three datasets, respectively. From Table 3, we can make the following observations.

1) Compared to Word2vec, SelfEDI and Bert performed better, both extracting sentence semantic information relatively well. Compared to SelfEDI, Bert uses the encoder from Transformer (Vaswani et al. 2017), which considers the semantic connections between words, obtaining features with richer semantics. Moreover, Bert uses a fine-tuning approach, which is superior to the feature-based approach.

2) Our FA-Net model consistently improves the results of the baseline models (i.e., LSTM, CNN, and Bert) using the proposed feature alignment method. This verifies the effectiveness of the proposed feature alignment method of aligning the feature vectors in and cross domains.

3) Compared with traditional CNN, the CNN+FA model increases the result by 3%, 4%, 6% in top1 accuracy on three datasets. The improvement of LSTM+FA is less, increased by 2%, 4%, 4%. This shows that the way that CNN extracts features is quite effective in extracting complete semantic information, which leads to more features being used. That is why the alignment of the CNN features brings more improvements compared to the LSTM model.

4) By comparing the results of the LSTM and CNN-based models, we can see that our FA-Net can improve the feature representation capabilities of the Bert-based models the best. The FA+Bert can increase the accuracy by 11%, 9%, and 12% on three datasets. That is to say, our in-domain and cross-domain contrastive learning method can make the representation of Bert-based obtain a higher discriminative power for euphemism identification.

Comparison with LLMs Table 4 summarizes the top1 euphemism identification results of our proposed Bert+FA against the LLMs. Although LLMs perform well in a series of tasks in natural language processing, there is still room for improvement in a specific task such as euphemism identification. From Table 4, we observe: 1) Our Bert+FA model beats all the four LLMs; 2) In the four LLMs, GPT3.5 is the best and most stable for euphemism identification; 3) Compared with our Bert+FA model, the time-consuming of the LLMs is about 3-7 times that of ours, and the cost is about 10-200 times that of our Bert+FA. For the LLMs in Table 4, the introductions, API instructions, content templates and some cases are detailed in the supplementary material.

Visualization Analysis To substantiate the soundness of the FA-Net, taking the Bert+FA for example, we map the distribution of the training data (source data) and test data (target data) embeddings to a two-dimensional coordinate

[7]https://replicate.com/stability-ai/stablelm-tuned-alpha-7b
[8]https://modelscope.cn/studios/damo/mPLUG-Owl/summary
[9]https://huggingface.co/models?other=llama-2
[10]https://platform.openai.com/docs/api-reference/introduction

Model	Drug	Weapon	Sexuality
Bert	0.24	0.38	0.38
Bert+EM	0.27	0.38	0.40
Bert+IC	0.27	0.39	0.41
Bert+HC	0.29	0.42	0.43
Bert+IC+EM	0.30	0.40	0.42
Bert+HC+EM	0.33	0.43	0.45
Bert+IC+HC	0.34	0.45	0.47
Bert+IC+HC+EM	**0.35**	**0.47**	**0.50**

Table 5: Ablation experiments. Bert+X means adding the X module of FA-Net to Bert-based models.

space by t-SNE, as shown in Figure 4(a). Obviously, the gap between the training data and the test data has been alleviated, and the two are fused together. Furthermore, we compute the domain discrepancy ($\mathcal{A}$-distance) (Long et al. 2015) between the source domain and the target domain. From Figure 4(b), it is easily observed that our FA-Net can effectively reduce the domain gap between the source and target data.

Ablation Study

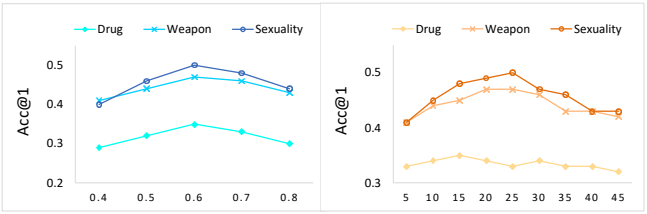
In order to investigate the efficacy of each component of FA-Net, we conducted experiments using or removing the in-domain contrastive learning network (IC-net), hard cross-domain contrastive learning network (HC-net), or entropy minimization network (EM-net) to verify the model's effectiveness on the three datasets. Here, we use the Bert-based method as the base model.

Experimental results of the ablation test of each component of FA-Net are presented in Table 3, where Bert+X (e.g., IC, or HC) refers to adding the IC-net (or HC-net) to Bert-based models. As FA-Net mainly uses contrastive learning and entropy minimization with no additional parameters, the parameters of all the experiments compared in Table 5 are the same. From the results, we can observe:

1) Whether IC, HC, or EM is added, or any two of them, or all three are added to the Bert model at the same time, the accuracy will increase significantly compared with the Bert model. For example, for the Drug dataset, when we add the IC-net (i.e., Bert+IC), the top1 accuracy on Drug increases by 3%. When we add both IC-net and HC-net, the top1 accuracy on Drug increases by 10%, etc. These results show that each component in FA-Net is important, and the absence of any one component will lead to a decline in accuracy.

2) In the Bert+HC, we add only the HC-net module to Bert model. Compared to Bert+IC and Bert+EM, the accuracy of Bert+HC increases the most by 5% on Drug, 4% on Weapon, and 5% on Sexuality, which means that the HC-net contributes more to FA-Net than other modules do.

3) In the experiment of Bert+IC+HC, we add both the IC-net and HC-net. As stated in the FA-Net Section, the IC-net clusters the same category and separates different categories of samples while the HC-net aligns the cross-domain features of the same category, the two modules jointly align the in-domain and cross-domain features. The results showed that the accuracy of Bert+IC+HC increases by 10% on Drug, 7% on Weapon, and 9% on Sexuality, which increase more than the addition of Bert+IC and Bert+HC, revealing that the IC-net and HC-net play a complementary role and promote each other.



(a) Similarity threshold γ (b) Epoch η

Figure 5: Performance sensitivity of the hyper-parameters.

Effects of Hyper-parameters

In the whole FA-Net module, except for the hyper-parameters similarity threshold γ in HC-net and the number of the beginning epoch of entropy minimization η, the remaining parameters can be learned through the model training. We test the sensitivity of Bert+FA on three datasets, as shown in Figure 5. From Figure 5(a), we observe that the best results on the three datasets are obtained when γ takes 0.6. When $\gamma < 0.6$, the obtained cross-domain positive sample may be mixed into the negative sample, resulting in training bias and bad results. When $\gamma > 0.6$, the obtained cross-domain positive sample is less, which is insufficient to guide the model training. From 5(b), we observe that FA-Net has different sensitivity in different datasets on η. Employing the entropy loss too precipitously requires the model to early decide labels for uncertain instances before the model has learned good representations and the classification boundary on the source domain. Meanwhile, each dataset has different source data and distribution, so the epochs when models learn good representations and classification boundaries of the source domain are different. For instance, for the Drug data, when $\eta = 15$, Bert+FA achieves the best performance.Whereas for the other datasets, the optimal hyper-parameters are at different settings.

Conclusion

In this paper, we present a text-text domain gap in the training data and test data in euphemism identification and explain the domain gap in terms of data distribution and cone effect. Moreover, to reduce the domain gap, we proposed a novel feature alignment network (FA-Net) based on contrastive learning and entropy minimization with no additional parameters. The proposed FA-Net consists of three sub-networks: one for aligning in-domain features, another for aligning cross-domain features, and the last for finally increasing the model's confidence. Experiments demonstrate the effectiveness of the FA-Net and the state-of-the-art performance of our proposed Bert+FA model. In addition, although the LLMs achieve good results in natural language processing, they do not perform well in the domain-specific task of euphemism identification and are much more time-consuming and expensive than our model, which further proves the significance of our research.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (No. 62272188), the Inner Mongolia Autonomous Region Major Science and Technology Project (No. 2021ZD0046), and the Fundamental Research Funds for the Central Universities (No. 2662021JC008). We would like to thank the Experimental Teaching Center of the College of Informatics, Huazhong Agricultural University for providing the computing resources. Thanks to all reviewers for their constructive comments.

References

Administration, D. E.; et al. 2018. Slang terms and code words: A reference for law enforcement personnel. *DEA Intelligence Report DEAHOU-DIR-022*, 18(2018): 2018–07.

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*, 145–155. PMLR.

Barone, A. V. M. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*.

Candela, J. Q.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. Dataset shift in machine learning. *The MIT Press*, 1: 5.

Cui, S.; Wang, S.; Zhuo, J.; Su, C.; Huang, Q.; and Tian, Q. 2020. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12455–12464.

Felt, C.; and Riloff, E. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, 136–145.

Foye, J.; Ball, M.; Jiang, C.; and Broadhurst, R. 2021. Illicit firearms and other weapons on darknet markets. *Trends and Issues in Crime and Criminal Justice [electronic resource]*, 1(622): 1–20.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gavidia, M.; Lee, P.; Feldman, A.; and Peng, J. 2022. CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms. *arXiv preprint arXiv:2205.02728*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Ke, L.; Chen, X.; and Wang, H. 2022. An Unsupervised Detection Framework for Chinese Jargons in the Darknet. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 458–466.

Li, Z.; Du, X.; Liao, X.; Jiang, X.; Champagne-Langabeer, T.; et al. 2021. Demystifying the dark web opioid trade: content analysis on anonymous market listings and forum posts. *Journal of Medical Internet Research*, 23(2): e24486.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lu, Y.; Li, D.; Wang, W.; Lai, Z.; Zhou, J.; and Li, X. 2021. Discriminative invariant alignment for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 24: 1871–1882.

Ma, X.; Zhang, T.; and Xu, C. 2019. Deep multi-modality adversarial networks for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 21(9): 2419–2431.

Magu, R.; and Luo, J. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 93–100.

Pinker, S. 2003. *The blank slate: The modern denial of human nature*. Penguin.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Yuan, K.; Lu, H.; Liao, X.; and Wang, X. 2018. Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces. In *27th USENIX Security Symposium (USENIX Security 18)*, 1027–1041.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25. PMLR.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, 7404–7413. PMLR.

Zhu, W.; and Bhat, S. 2021. Euphemistic Phrase Detection by Masked Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 163–168.

Zhu, W.; Gong, H.; Bansal, R.; Weinberg, Z.; Christin, N.; Fanti, G.; and Bhat, S. 2021. Self-Supervised Euphemism Detection and Identification for Content Moderation. In *Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP)*.