A Unified Generative Framework for Multilingual Euphemism Detection and Identification

Anonymous ACL submission

Abstract

Currently, various euphemisms are emerging in social networks, attracting widespread attention from the natural language processing community. However, existing euphemism datasets are only domain-specific or language-specific. In addition, existing approaches to the study of euphemisms are one-sided. Either only the euphemism detection task or only the euphemism identification task is accomplished, lacking a unified framework. To this end, we construct a large-scale **M**ulti-lingual **M**ulti-category dataset of **E**uphemisms named **MME**, which covers a total of 12 categories for two languages i.e., English and Chinese. Then, we first propose a unified generative model to **Joint**ly conduct the tasks of multilingual **E**uphemism **D**etection and **I**dentification named **JointEDI**[1]. By comparing with LLMs and human evaluation, we demonstrate the effectiveness of the proposed JointEDI and the feasibility of unifying euphemism detection and identification tasks. Moreover, the MME dataset also provides a new reference standard for euphemism detection and identification.

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers.

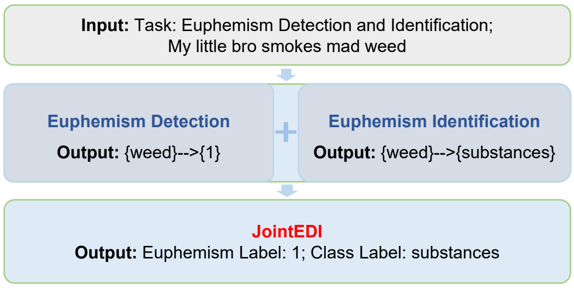


Figure 1: Comparison of JointEDI with Euphemism Detection Methods and Euphemism Identification Methods.

1 Introduction

Euphemisms are forms of language that express ideas or convey information through the use of indirect or cryptic language. The original intention of using euphemisms is to avoid direct, blunt or potentially offensive expressions (Pinker, 2003). However, to avoid explicitly expressing unfriendly views or statements, some users choose to use euphemisms to cover up discriminatory, insulting or unfair remarks (Chilton, 1987). For instance, lawbreakers use euphemisms (eg: “weed” means “rank grass” in literal English, but “drugs” in euphemisms) (Zhu et al., 2021), to distract the attention of the cyber police and complete the transaction of drugs, guns and other illegal goods. People discriminate or insult others using euphemisms (eg: “同志” in Chinese means people who strive for a common ideal or cause, but in euphemisms it means “同性恋” (homophobic)) (Lee et al., 2023). Therefore, it is important to study the detection and identification of euphemisms to detect and intervene in the transmission of euphemisms promptly.

Many euphemisms in English and Chinese are frequently used, but current research is limited to a single language, such as (Gavidia et al., 2022) (Gavidia et al., 2022) (Keh et al., 2022). As the world globalizes and communicates more intensively, some euphemism expressions combine Chinese and English to convey implicit meanings (eg: “OMG, 你这是<发福>了吗？”(OMG, are you <happy>?). It’s a sign that someone’s getting fat.), where shows that it is not enough to study euphemisms from a monolingual perspective (Romaine, 2012). Therefore, euphemism datasets covering multiple languages and domains are urgently needed, which is important for the study of euphemisms, especially for euphemism detection and identification tasks.

As shown in Figure 1, existing euphemism tasks can be divided into two categories according to their purpose (Zhu et al., 2021): (1) Euphemism

[1]Our data and code will be open source once acceptance.

Detection: the main purpose of the task is to determine whether a text contains euphemisms so that they can be further analyzed or processed. (2) Euphemism Identification: this task focuses more on identifying specific euphemistic expressions in the text and aims to understand and analyze the use of euphemisms in the text in more details. However, detecting and identifying euphemisms in practice is an ongoing process. To the best of our knowledge, there is not yet a methodology to unify the two tasks into a single framework.

To solve the above challenges, by integrating an existing dataset of euphemisms and additionally collecting some other data from websites. We construct a large-scale multi-lingual multi-category dataset of euphemisms, named MME, which includes two popular languages of the world, English and Chinese. This dataset was filtered in detail and manually labeled, and finally, the Chinese euphemisms were classified into 9 categories and the English euphemisms into 8 categories. Furthermore, we propose a novel unified framework for the joint implementation of the euphemism detection and identification tasks, a generative model named JointEDI, which adopts two auxiliary tasks. We conducted extensive experiments on the MME dataset comparing existing large language models and human evaluation, verifying the superiority of our proposed method and providing new insights for future work.

Our contributions are as follows:

- We construct a large-scale **M**ulti-language **M**ulti-category **E**uphemism dataset named MME, including 2 languages and covering 12 categories in total, which provides a new benchmark in the field of euphemism detection and identification. We also provide an in-depth statistical analysis.
- We propose a unified generative framework to jointly conduct the tasks of euphemism detection and identification named **JointEDI**, employing two auxiliary tasks. To the best of our knowledge, this is the first framework to unify the task of euphemism detection and identification.
- Experimental results on MME datasets show that 1) the proposed JointEDI outperforms other model baselines and the LLMs, demonstrating the validity of our approach, and 2) our results are much higher than common human evaluation results, but lower than those evaluated by professional human, demonstrating the challenging nature of our dataset and the unified task.

2 Related work

2.1 Datasets

For computers, euphemisms often involve complex contexts and emotions, and even for large language models, accurately understanding and processing these linguistic expressions is still a challenging task (Gibbs, 1999). Many domain-specific euphemism datasets have been proposed. We summarize and analyze the most representative datasets in recent years in Table 1.

It can be seen from Table 1, (Zhu et al., 2021) and (Ke et al., 2022) proposed English and Chinese datasets for the domain of darknet euphemisms, respectively. (Rahman et al., 2021) and (Yadav et al., 2023) proposed two-classification and five-classification datasets for the domain of hate euphemisms, respectively. (Gavidia et al., 2022) first introduced the concept of PET (Potentially Euphemistic Terms) and proposed a multi-categorical euphemism dataset. (Lee et al., 2023) proposed four different languages to present a novel euphemism corpus which is expanded into four languages based on the dataset proposed by (Gavidia et al., 2022). Although some datasets are quite large (Zhu et al., 2021) (Ke et al., 2022) (Mody et al., 2023) (Yadav et al., 2023). However, more data is not always better, and the extra irrelevant data may affect the model due to pseudo-correlation coincidence (Feng, 2021), we have to ensure the size of the dataset while improving the quality of the dataset, such as data categories and their distribution. Therefore, in order to promote the research of euphemisms and better reflect the diversity of euphemisms in real scenarios, a large-scale multi-lingual and multi-category euphemism dataset is urgently needed.

2.2 Euphemism Detection

The main objective of the euphemism detection task is to detect whether a piece of text contains euphemisms or not. (Magu and Luo, 2018) proposed a method to help identify unknown words to detect hate speech euphemisms using word embedding and network analysis. (Ghosh et al., 2020) proposed a Sarcasm Detection Shared Task that focuses on the detection of hate speech euphemisms using the entire context of a previous conversation, which achieved a high detection accuracy of 0.932 for the first-place team in that competition. It is worth noting that almost all teams used pre-trained transformer-based models. (Zhu

Dataset	Instances		Categoriy	PET	Domain	Language
	English	Chinese				
Zhu et al. (2021)	-	-	3	Yes	Darknet	English
Rahman et al. (2021)	4275	-	1	No	Hateful	English
Gavidia et al. (2022)	1965	-	7	Yes	Hateful	English
Keh et al. (2022)	-	44720	10	Yes	Darknet	Chinese
Mollas et al. (2022)	999	-	6	No	Hateful	English
Mody et al. (2023)	451709	-	1	No	Hateful	English
Yadav et al. (2023)	227836	-	5	No	Hateful	Six languages
Lu et al. (2023)	-	12011	4	No	Hateful/Offensive	Chinese
(Lee et al., 2023)	1952	1552	7	Yes	-	Four languages
MME (Ours)	4512	4495	12	Yes	All	English/Chinese

Table 1: Comparison of existing euphemism datasets. For comparison, we show only the number of English and Chinese sentences in the datasets of (Lee et al., 2023).

et al., 2021) formulated the euphemism detection problem as an unsupervised filler mask problem and solved it by combining self-supervision with a masked language model (MLM). A recent work that has attracted attention is the presentation of the Euphemisms Detection Shared Task (Lee et al., 2022). The purpose of the task is: to give an input text and detect whether it contains euphemisms or not. The competition attracted 13 teams and in that competition (Keh et al., 2022) combined the best-performing models into an ensemble of three models and achieved first place in that competition. (Kesen et al., 2022) used additional supervised information to obtain imageries of both the PETs and their literal descriptions using a text-to-image model, combining textual modalities and visual modalities to achieve good euphemism detection results.

2.3 Euphemism Identification

Once euphemisms are detected, the subsequent identification of euphemisms is very important because different types of euphemisms determine the specific application scenarios of euphemisms. However, there is relatively little work related to the study of euphemism identification tasks. Since a euphemism often contains several different meanings, this task is more challenging than the euphemism detection task (Zhu et al., 2021). (Yuan et al., 2018) proposed Cantreader, which employs a neural network-based embedding technique to analyze the semantics of words, to be used for automatic detection and comprehension of cryptic speech. Instead of directly recognizing the specific meaning of a euphemism, they generate a set of superlatives and use a binary random forest classifier and recursive lookup to categorize a given euphemism into a specific superlative. (Felt and Riloff, 2020) used sentiment analysis to identify euphemisms and dysphemisms, and although the performance of (Felt and Riloff, 2020)'s system was relatively low and the subject matter was narrow, this work certainly has stimulated further research. (Zhu et al., 2021) explicitly defined the task of euphemism identification for the first time, and developed a self-supervised learning algorithm that utilizes a bag-of-words model to classify a given euphemism to a specific superordinate word at the sentence level.

Although both the euphemism detection and euphemism identification tasks have achieved some results, they are two independent tasks. In practical applications, the detection and identification of euphemisms is a continuous process. We not only need to detect euphemisms from a sentence but also identify the meaning of the specific expression of the euphemism. To the best of our knowledge, only (Zhu et al., 2021) have proposed a pipeline that connects these two tasks in tandem, but this is not a unified framework. Moreover, the approach is limited to three specific tasks, namely drugs, weapons and sex, in the darknet. Unlike all previous approaches, we propose a unified framework to unify the tasks of euphemism detection and euphemism identification to fully understand the implicit meaning to be conveyed throughout the sentence. As shown in Figure 1, JointEDI can detect whether a sentence contains a euphemism or not, and at the same time identify to what category the euphemism belongs.

3 Dataset Construction

3.1 Data Collection

The construction process of the MME dataset is shown in Figure 2. Our goal is to construct a large-scale euphemism dataset covering multi-category and multilingual euphemisms. We first extensively researched and analyzed the purpose euphemism dataset as in Section 2.1. We collected the following potentially usable datasets. These include the dataset proposed by (Lee et al., 2023) (Lu et al., 2023) (Zhu et al., 2021). It is worth stating that although the above studies have proposed datasets of a certain size and category, we found that the above datasets are more or less wrongly labeled with categories. For example, in the dataset proposed by (Lu et al., 2023):

"因为杨笠说男人是<垃圾>,然后教育男性不要对号入座"*(Because Yangli says men are <trash> and then teaches men not to be right), the "垃圾"(trash) here should be labeled with the category of "性别"(gender), but the original dataset is labeled with the category of "种族"(racist).*

To ensure that the dataset covers as many categories as possible while minimizing the problem of inter-class and intra-class imbalance in the MME dataset, We also crawled extensive data from glosbe[2] and sogou[3].

3.2 Data Cleansing and Filtering

For the three collected datasets mentioned above, we mainly used the Chinese dictionary and English dictionary proposed by (Lu et al., 2023) to annotate the Chinese data and English data we collected with keywords, and filter out sentences in the dataset that do not contain the keyword. At the same time, incomplete data were eliminated and data with obvious labeling errors were selected for manual secondary labeling. To construct a high-quality euphemism dataset, we mitigate the problem of inter-class imbalance in the dataset by filtering categories with less than 50 data. For example, the category "misc.".

3.3 Data Annotation

The data we collected contains an assortment of types, including daily polite phrases, discriminatory, sarcastic, and phrases from domains such as the darknet. Since euphemisms are related to the social and cultural aspects of language use, they are an important research area in the field of sociolinguistics in linguistics. Therefore, in order to ensure the quality and authority of the collected data on euphemisms, we hired five linguistic professionals to manually label data, including three professors and two PhD candidates. We provide systematic training for the annotators before data labeling begins. See Appendix A for training programs.

In the process of labeling, if we encountered controversial categories of euphemisms, we used the voting method to select the category with the highest number of votes to get the final category of euphemisms. We followed the basis of the classification of (Lu et al., 2023). We also finalized the MME English dataset to get a total of 8 specific categories, and the MME Chinese dataset to get a total of 9 specific categories.

[2]https://glosbe.com
[3]https://wap.sogou.com/

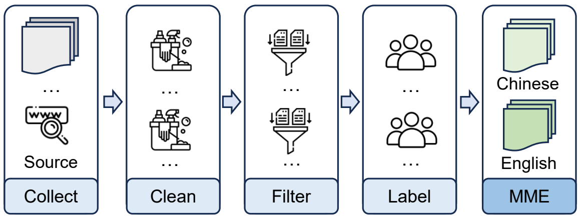


Figure 2: Flowchart of Dataset Construction. It consists of four main processes, which are data collection, data cleaning, data filtering and data labeling

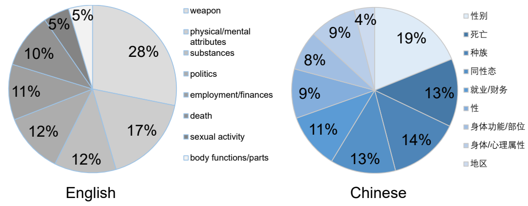


Figure 3: The pie chart on the left shows the percentage of English data by category, and the one on the right shows Chinese. 2.

3.4 Data Analysis

Through the collection and cleaning of data and detailed labeling by professionals, we finally collected a total of Chinese data and English data as shown in Table 2. We ended up with the number of data for each category. The accounting for each category is shown in Figure 3. In order to illustrate what kind of keyword euphemisms are available for each category, we did keyword data analysis on the English and Chinese datasets of MME respectively. Detailed data analysis can be found in the Appendix B.

Category	Sentence	
	English	Chinese
body functions/parts	209	450
death	479	580
employment/finances	484	477
physical/mental attributes	781	401
sexual activity	225	421
politics	525	-
substances	538	-
weapon	1271	-
gender	-	827
racist	-	607
homophobic	-	557
region	-	175
In total	4512	4495

Table 2: Data categories and quantities of English euphemisms and Chinese euphemisms in the MME dataset. The MME English data has 2658 sentences with label 1 and 1854 sentences with label 0. The MME Chinese data has 3100 sentences with label 1 and 1395 sentences with label 0.

4 Task Definition

Unlike previous euphemism detection tasks and euphemism identification tasks, the main objective of our task is to unify euphemism detection and euphemism identification into a single modeling framework, with the core goal of creating a unified framework that is able to automatically identify euphemisms in text and classify them into the correct categories. It is known that a sentence containing a euphemism, $s = [w_1, ..., PET, ..., w_i, ..., w_m]$ (where PET is known to be a potential euphemism). Our goal is to determine whether PET represents a euphemism in a sentence and, if PET is a euphemism, to identify the category to which PET belongs.

As shown in Figure 1, the inputs and outputs of JointEDI are as follows:

Input: "Task: Euphemism Detection and Identification; My little bro smokes mad weed."

Output: "Target: Euphemism Label: 1, Class Label: substances)"

The euphemism label of the model's output is "1", which indicates that "weed" represents a euphemism in the sentence, and the class label of the model's output is "substances", which indicates that "weed" is categorized into "substances".

5 Methodology

5.1 Model Overview

Since our proposed method JointEDI aims to unify the task of euphemism detection and identification for multiple languages, we use a multi-lingual BART (mBART) (Liu et al., 2020), which is an extended version of a transformer-based pre-trained BART (Lewis et al., 2020) for multiple languages, as our Seq2Seq framework. The following Figure 4. illustrates the overall architecture of our proposed JointEDI, which mainly consists of the mBART encoder and mBART decoder.

As we discussed in the previous section, our task can be represented as taking X = [Task: Euphemism Detection and Identification; s=(x0,x1, ..., xt)] as input and outputting a target sequence Y = [Euphemism Label: y; Class Label: c], where s stands for the sentence to be detected and identified. Thus, our euphemism detection and identification task can be formulated as:

$$\widehat{Y} = mBART(X), \qquad (1)$$

where X is the input sequence and $\widehat{Y}$ is the output sequence generated by the model.

To assist the JointEDI unified euphemism detection and identification task to achieve better results, we have included two auxiliary tasks in our model, namely, the euphemism detection and the euphemism identification task. Next, we will introduce each structure of the framework and present our proposed framework separately.

5.2 mBART Encoder

The sentence to be encoded is taken as input and passed to the Encoder of mBART. The sentence is first processed by tokenization to decompose it into sub-words or word fragments. The disambiguated content is passed through a word embedding layer, which maps each word fragment to a high-dimensional word embedding space. Then positional encoding is added to preserve the order information of the words in the input sentence. For the input sequence X, the encoder of mBART can be expressed as:

$$H_{encoder} = Encoder(X), \qquad (2)$$

where $H_{encoder}$ is the output of the encoder and contains the encoded information of the input sequence.

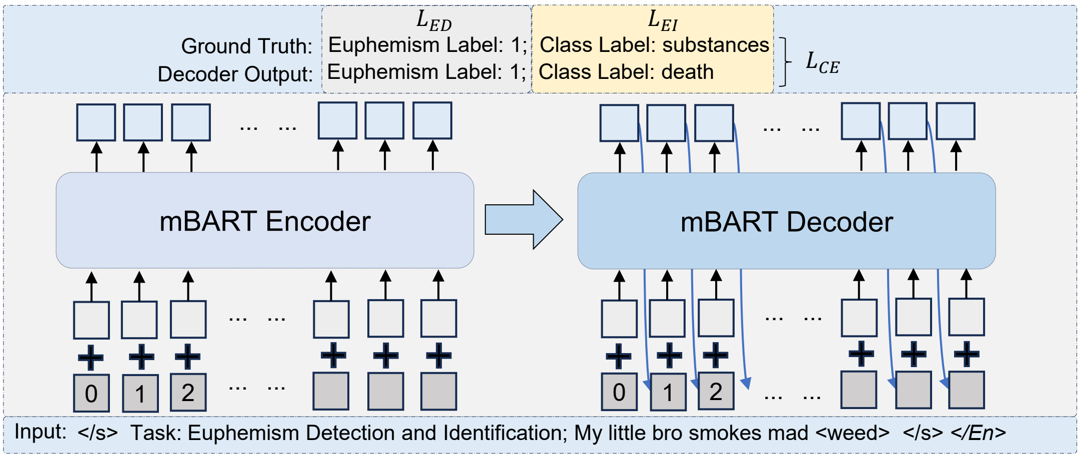


Figure 4: Overall network framework for JointEDI.

5.3 mBART Decoder

The decoder of mBART also uses the same auto-attention mechanism of the transformer model but uses an auto-regressive process for training, it utilizes the portion that has already been generated when generating each output position.

For the output Y_t at the current position t, mBART's decoder can be represented as:

$$Y_t = Decoder(H_{encoder}, Y_{<t}), \tag{3}$$

where $Y_{<t}$ denotes the output sequence that has been generated before position t. The decoder utilizes the previously generated partial sequence and the encoder's output when generating the output for each position.

5.4 Loss Function

The mBART receives an input sequence and outputs a sequence. The output sequence is a probability distribution for each word in the vocabulary, and as each token is generated, the model converts the output score to a probability distribution using the Softmax function, which ensures that the generated sequence is a legitimate probability distribution.

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T}\log(P(Y_{i,j}|\widehat{Y}_{i,j});\theta), \tag{4}$$

where N represents the number of samples, T represents the length of the sequence, $P(Y_{i,j})$ represents the labeling of the j_{th} position of the i_{th} sample in the real sequence, and $\widehat{Y}_{i,j}$ represents the probability distribution of the predicted labeling of the j_{th} position of the i_{th} sample of the sequence species generated by the model.

To enhance JointEDI's ability to achieve a unified euphemism detection and identification task, as well as to measure the ability in the euphemism detection subtask. We include the euphemism detection loss in the output of the encoder as follows:

$$\mathcal{L}_{ED} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right), \tag{5}$$

where N represents the number of samples, y_i represents the true label of the i_{th} sample, and p_i represents the probability predicted by the model for the i_{th} sample.

Furthermore, we enhance the ability of JointEDI to unify the euphemism detection and identification tasks by incorporating a euphemism identification loss at the decoder output as follows:

$$\mathcal{L}_{EI} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(p_{i,j}), \tag{6}$$

where N represents the number of samples, C represents the number of categories, and $y_{i,j}$ is the label of the j_{th} category in the true label of sample i. $p_{i,j}$ is the probability of the j_{th} category in the model prediction of sample i.

The total loss function of the final model is as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{ED} + \gamma\mathcal{L}_{EI}, \tag{7}$$

where α, β and γ represent the weights of the three loss functions, respectively. The sum of α, β and γ is 1.

Model	Pair-F1 (%)		Pair-R (%)		Pair-P (%)		F1 (%)		R (%)		P (%)	
	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN	EN	CN
mBART-large-cc25	85.71	87.14	82.95	85.65	88.67	86.26	86.47	89.44	83.16	86.21	90.06	92.92
mBART-large-50	85.18	73.53	89.68	94.44	81.10	60.20	87.16	83.05	90.05	95.40	84.45	73.52
mT5-base	63.32	54.55	80.07	72.97	52.36	43.55	79.81	79.12	85.20	82.49	75.06	76.01
mT5-large	79.90	83.77	88.27	91.39	72.98	77.33	84.85	88.54	89.29	92.12	80.83	85.22
JointEDI	**93.11**	88.81	**92.51**	91.96	**93.72**	85.87	93.80	92.97	92.60	92.56	95.03	93.38
Falcon	0	0	0	0	0	0	10.53	0	5.71	0	66.67	0
LLaMA2-70b-chat	36.67	17.39	31.43	20.00	44.00	15.38	66.67	72.73	57.14	60.00	80.00	92.31
mPLUG-Owl	0	0	0	0	0	0	0	0	0	0	0	0
Stability-AI	10.53	0	5.71	0	66.67	0	10.53	0	5.71	0	66.67	0
GPT-3.5	66.67	43.59	67.50	37.78	65.85	51.12	88.89	74.36	90.00	64.44	87.80	87.88
GPT-4.0	73.17	53.66	75.00	48.89	71.43	59.46	90.24	82.93	92.50	75.56	88.10	91.89
Human-com	68.20	80.03	71.15	79.49	66.08	80.64	84.29	88.65	77.50	82.22	92.93	96.25
Human-pro	92.75	**94.33**	92.17	**95.40**	93.40	**93.28**	**95.44**	**96.64**	**93.75**	**96.67**	**97.39**	**96.64**

Table 3: Comparison of JointEDI with Baseline and Large Language Model on MME Chinese dataset and English dataset respectively, where EN stands for English and CN stands for Chinese. Human-com represents the average metrics of test results for non-professional people, and Human-pro represents the average metrics of test results for professional people.

6 Experiments

6.1 Evaluation Setup

Datasets: We evaluate our method on the MME dataset constructed in Section 3. There are 4512 sentences in the English dataset and 4495 sentences in the Chinese dataset. We divided the two datasets according to the ratio of training, validation, and testing 7:1.5:1.5, and when dividing the datasets, we tried to ensure the balance of inter-class and intra-class data. The final results of the division of the two datasets into each class are shown in Appendix C.

Implementation Details: During the fine-tuning process, the maximum length of the input sequence was set to 128, and the initial learning rate to 1e-5. We trained the model for 20 epochs on a 40GB Tesla A100 GPU with the batch size set to 32. We used the Adam optimizer and the model employed a cosine annealing learning rate schedule.

Baselines: We compared four baselines and six LLMs. The setup of the large language model and the details of the comparison models are described in Appendix C.

Accuracy metrics: We set up six evaluation metrics, where P, R and F1 represent the metrics for euphemism detection, Pair-Recall, Pair-Precision and Pair-F1 represent the metrics for the task of unifying euphemism detection and identification. The values of Pair-F1 and F1 are used as the main evaluation metrics.

6.2 Results and Analysis

Comparison with Baselines: As can be seen in Table 3, on the unified euphemism detection and identification task, We find that JointEDI is better at both Pair-P and Pair-F1 in English language, but does not perform as well as the other methods in Chinese data, which suggests that JointEDI can effectively detect euphemisms in sentences and identify the categories of euphemisms. Even though mT5 is pre-trained in more languages and has more parameters, it is less effective than JointEDI. Since the two models have different pre-trained methods, if the input is “A-B-E”, the output of mBART is labeled “ABCDE”, but the output of mT5 is labeled “CD”. It seems that mBART is performing a more difficult task, and is more effective in detecting and identifying euphemisms.

On the euphemism detection task, JointEDI’s P and F1 are superior in both languages, but still not as good as the other baselines in terms of R for Chinese data. We analyze that this is due to the fact that due to the higher number of parameters in the compared baselines leads to a model that may tend to predict as containing euphemisms more frequently, and thus have a higher R value.

Comparison with LLMs: Table 3 summarizes the results of the comparison between JointEDI and LLMs. We note that GPT-4.0 has the best results among all the LLMs, which is analyzed since GPT-4.0 has the largest number of parameters compared to the other LLMs. Despite this, the performance of our proposed JointEDI is still able to outperform LLMs, which suggests that unifying the euphemism detection task with the euphemism identification task is still challenging for LLMs.

Comparison with Human Evaluation: Although JointEDI achieved the best results, it was still not

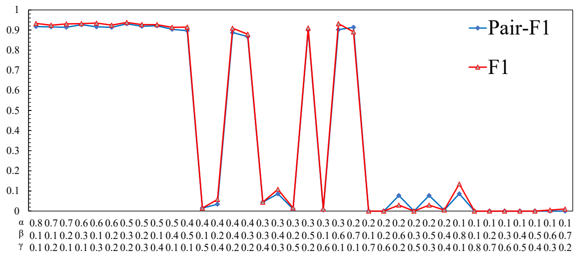


Figure 5: For hyperparameter analysis in the MME dataset, the blue color represents the value of Pair-F1 and the red color represents the value of F1.

1) *Correct detection and Incorrect identification:*
Input: White boy, buys <coke> from me.
Output: Euphemism Label: 1; Class Label: death
2) *Incorrect detection and Correct identification:*
Input: ... kinetic impact devices , and chemical control substances, tear gas and pepper <spray>…
Output: Euphemism Label: NULL; Class Label: weapon
3) *Incorrect detection and Incorrect identification:*
Input: …relax in stressful situations, <neutralize> interoffice conflict, add zest to dull relationships…
Output: Euphemism Label: 1; Class Label: politics

Figure 6: Analysis of different error types. (1) The true label of "coke" is 1, but it is categorized into the wrong category "death"; (2) The true label of "spray" is 1, the model does not predict the result, the output is "NULL" and it is categorized into the right category "weapon"; (3) The true label of "neutralize" is 0, the model incorrectly predicts 1, and it is divided into the wrong category "politics".

as good as the human-professional assessment on the MME Chinese dataset, we analyze that this is because mBART uses a larger English corpus than a Chinese one for pre-training, and thus will perform better on English data. At the same time, euphemisms are usually informal and subjective expressions whose form and meaning may change in different contexts. The same sentence may be regarded as a euphemism or as a direct expression in different contexts. This subjectivity and variety increase the difficulty of understanding and identifying euphemisms in JointEDI.

Hyperparametric Analysis: We have analyzed the ablation of different combinations of α, β and γ. The experimental results are shown in Figure 5. When 0.5, 0.3 and 0.2 are selected for α, β and γ respectively, and the results of the model are the most optimal. We also find that when the sum of β and γ is greater than 0.5, the performance of JointEDI starts to be general and becomes unstable. For example, when α, β and γ take values of 0.4, 0.1 and 0.5, respectively, the results of the model plummet to near 0. It shows that the loss of uniform euphemism detection and identification plays a dominant role in the task of unifying euphemism detection and identification, followed by the loss of euphemism detection, and finally the loss of euphemism identification.

Error Analysis: We have selected three instances in the results of error detection and error identification, respectively. As shown in Figure 6. The first case is correctly detected and incorrectly identified, the second case is incorrectly detected and correctly identified, and the third case is both detected incorrectly and identified incorrectly. This shows that JointEDI is still challenging in domain-specific, context-specific or type-specific euphemism detection and identification tasks. To visualize the LLMs error output results, we also show some cases where LLMs fail in the task of euphemism detection and identification in Appendix C.

7 Practical Implications

This paper provides a new benchmark to unify the euphemism detection task with the euphemism identification task, which has high practical application value. Firstly, The method can be directly applied to social media to assist platforms in filtering offensive, inappropriate or controversial content in a timely manner, and reduce the auditing cost. Second, the method can be integrated into a large language model to deepen contextual understanding, detect euphemisms more accurately, and provide users with more accurate and sensitive responses by learning from large-scale corpora. Finally, this technology can facilitate the quality of cultural interactions on social media.

8 Conclusion

In this paper, we construct a multi-lingual multi-category euphemism dataset named MME, which contains two languages, English and Chinese, and covers more than a dozen categories, which provides a new benchmark for the research of euphemism detection and identification tasks. Meanwhile, we also propose a novel generative approach to unify the euphemism detection task and the euphemism identification task, which proves the effectiveness of our proposed JointEDI and the difficulty of this task by comparing it with LLMs and human evaluation. New insights are provided for the research of euphemism task.

Limitations

Although the work in this paper achieves certain results. However, the following limitations still exist: (1) The MME dataset still covers a relatively limited number of languages, and future efforts are needed to expand the language scope in order to achieve true multilingual euphemism detection and identification. The current dataset is relatively monolingual, while there exists a rich linguistic diversity globally, including about 7000 active languages. (2) The proposed JointDEI still has a performance gap when compared to a professional human evaluation. This suggests that the current method still needs to be further improved and optimized for more accurate euphemism detection and identification.

Ethics Statement

We strictly adhere to the data usage agreements of the various public online social platforms. The opinions and findings in the sample dataset we have provided should not be interpreted as representing the views expressed or implied by the authors. We hope that the benefits of our proposed resources outweigh the drawbacks. All resources are intended for scientific research only.

References

Paul Chilton. 1987. Metaphor, euphemism and the militarization of language. *Current research on peace and violence*, 10(1):7–19.

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

He Feng. 2021. *Detection and Data Mining of Diabetic Retinopathy Using Classic Machine Learning*. Ph.D. thesis, National University of Singapore (Singapore).

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11.

Raymond W Gibbs. 1999. Taking metaphor out of our heads and putting it into the cultural world. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 145–166.

Liang Ke, Xinyu Chen, and Haizhou Wang. 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*, pages 458–466.

Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. Eureka: Euphemism recognition enhanced through knn-based methods and augmentation. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 111–117.

Ilker Kesen, Aykut Erdem, Erkut Erdem, and Iacer Calixto. 2022. Detecting euphemisms with literal descriptions and visual imagery. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 61–67.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190.

Patrick Lee, Iyanuoluwa Shode, Alain Chirino Trujillo, Yuan Zhao, Olumide Ebenezer Ojo, Diana Cuervas Plancarte, Anna Feldman, and Jing Peng. 2023. Feed pets: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. *arXiv preprint arXiv:2306.00217*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating finegrained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint arXiv:2305.04446*.

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

Devansh Mody, YiDong Huang, and Thiago Eustaquio Alves de Oliveira. 2023. A curated dataset for hate speech detection on social media text. *Data in Brief*, 46:108832.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

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

Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. 2021. An information retrieval approach to building datasets for hate speech detection. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Suzanne Romaine. 2012. The bilingual and multilingual community. *The handbook of bilingualism and multilingualism*, pages 443–465.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*.

Kan Yuan, Haoran Lu, Xiaojing Liao, and XiaoFeng Wang. 2018. Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1027–1041.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 229–246. IEEE.

Appendix

A Data Labeling Training Guide

Your task is to determine, given a sentence containing a potential euphemism, whether PET behaves as a euphemism in the sentence, and if so, to label it as "1" and indicate to which category of euphemisms it belongs. If not, label it as "0". There are 9 categories in English and 8 categories in Chinese. Please make sure that the entire labeling process is free from outside interference, and pay attention to the context of the text when labeling to ensure that the euphemisms are accurately captured. In case of uncertainty or ambiguity, please mark according to your best judgment.

Example 1:

"My little bro smokes mad <weed> " Could you please indicate whether weed is a euphemism in the sentence, and if so, which of the following categories does the euphemism belong to?

"body functions/parts", "politics", "sexual activity", "physical/mental attributes", "death", "substances", "weapon", "employment/finances".

Labeling result: "1", "substances".

Example 2:

"…供桌供案主要应用于纪念<仙逝>长辈和敬供先人…", could you please indicate whether "仙逝" is a euphemism in the sentence, and If so, which of the following categories does the euphemism belong to?

"同性恋", "地区", "就业/财务", "性", "性别", "死亡", "种族", "身体/心理属性", "身体功能/部位".

Labeling result: "1", "死亡".

B Data Analysis

The MME English dataset is divided into a total of eight categories:

"body functions/parts", "politics", "sexual activity", "physical/mental attributes", "death", "substances", "weapon", "employment/finances".

The MME Chinese dataset is divided into a total of nine categories:

"同性恋", "地区", "就业/财务", "性", "性别", "死亡", "种族", "身体/心理属性", "身体功能/部位".

The top 10 keywords for each category in the MME dataset. Table 4 shows the MME English dataset and Table 5 shows the MME Chinese dataset.

C Experiments

The final results of the two datasets by dividing the training set, the validation set and the test set are shown in Figure 7 and Figure 8.

Baselines: The configurations of the four baseline models are as follows:

- **mBART-large-cc25**: Pre-trained mBART using 25 languages.
- **mBART-large-50**: Pre-trained mBART using 50 languages.
- **mT5-base**: Pre-trained mT5 base model for 101 languages with 580 million parameters (Xue et al., 2021).
- **mT5-large**: Pre-trained mT5-large model in 101 languages with 1.2 billion parameters.

Details of LLMs setup:The configurations of the 6 LLMs are as follows:

Category	Top 10 keywords
body functions/parts	accident, rear end, time of the month, accidents, droppings, chest, tinkle, lavatory, pass gas, latrine
politics	fishing, trick, underdeveloped, pro-life, inner city, wolf pack, armed conflict, undocumented immigrants, freedom, Global South
sexual activity	go all the way, work, sleep around, sex workers, birds and the bees, same-sex, sex worker, porn, slept with, girls
physical/mental attributes	special needs, disabled, plump, aging, slim, expecting, overweight, over the hill, troubled, mixed up
death	late, demise, collateral damage, put to sleep, perish, pass on, long sleep, deceased, pass away, fatality
substances	weed, coke, intoxicated, speed, pure , weeds, sober, substance abuse, inebriated, Coke
weapon	machine, shot, shoot, car, heavy, German, bear, saw, police, spray
employment/finances	disadvantaged, let go of, sanitation workers, economical, dismissed, deprived, well off, income inequality, homemaker, indigent

Table 4: Top 10 keywords for each category in the MME English dataset.

Category	Top 10 keywords
同性恋	基佬, 反同, 男同, txl, 通讯录, 同志, gay, 男童, 恐同, 撑同
地区	棒子, 弯弯, 南满, 飞舟, 蛮, 小日子, 南蛮, 九头鸟, 飞周, 蛮夷
就业/财务	调动, 环卫工人, 结构优化, 滑铁卢, 下岗, 低收入, 辞退, 负增长, 拮据, 裁员,
性	小姐, 失足, 上床, 亲热, 性侵, 慰安妇, 房事, 夫妻生活, 同房, 第三者
性别	女拳, 仙女, 打拳, 普信, 拳师, 幕刃, 亩, 牧人, 圣母, eg
死亡	牺牲, 没了, 走了, 解脱, 不在了, 挂了, 过世, 逝世, 遇难, 去世
种族	默, 猩猩, 虫类, 黑猩猩, 黑女, 嘿嘿, 类人猿, 媚黑, 小黑, 三非
身体/心理属性	丰满, 有喜, 发福, 年长, 年迈, 苗条, 高龄, 失明, 特殊人群, 长寿,
身体功能/部位	姨妈, 方便, 胸部, 卫生间, 洗手间, 大号, 下身, 私处, 生理期, 如厕

Table 5: Top 10 keywords for each category in the MME Chinese dataset.

- **Falcon**: A new series of large-scale language models created by the Technology Innovation Institute in Abu Dhabi, with 40 billion parameters.
- **StableLM**[4]: A Stable Diffusion startup, Stability AI, released and open-sourced a large language model trained by the team with 7 billion parameters
- **mPLUG-Owl**[5]: A large multimodal model based on a modular implementation with 7.2 billion parameters.
- **LLaMA2-70b-chat**[6]: A Meta AI official release of the latest generation of open source big models with 70 billion parameters.
- **GPT-3.5-turbo**[7]: A fourth in a series of NLP models designed by OpenAI, with 20 billion parameters.
- **GPT-4.0**: A large-scale, multimodal artificial intelligence model developed by OpenAI.
- **Human Evaluation**: We invited four professionals in the field of linguistics and four non-professional people to evaluate 200 English texts and Chinese texts, respectively, in order to detect and identify euphemisms in sentences. To ensure the fairness of the experiment, the participants in the test do not parameterize the annotation process.

[4] https://replicate.com/stability-ai/stablelm-tuned-alpha-7b

[5] https://modelscope.cn/studios/damo/mPLUG-Owl/summary

[6] https://huggingface.co/models?other=llama-2

[7] https://platform.openai.com/docs/api-reference/introduction

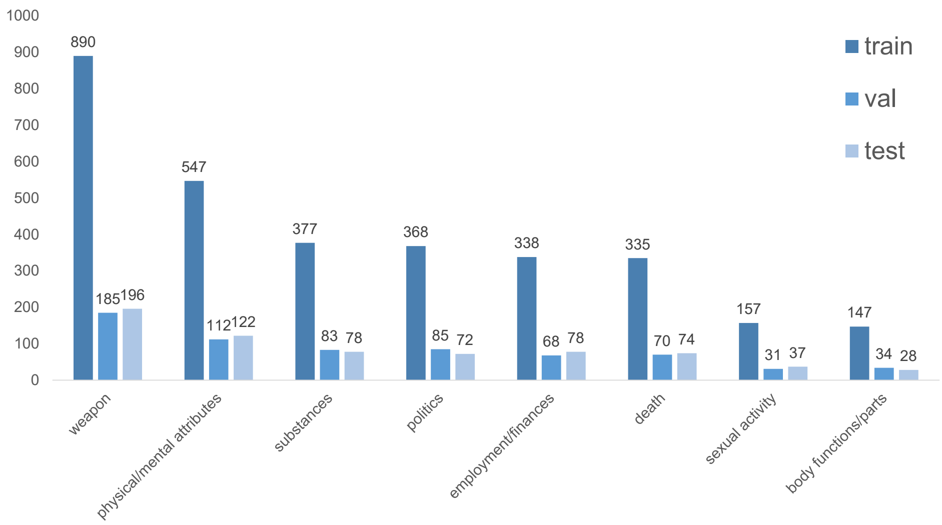


Figure 7: The final results of the MME English dataset by dividing the training set, validation set, and test set

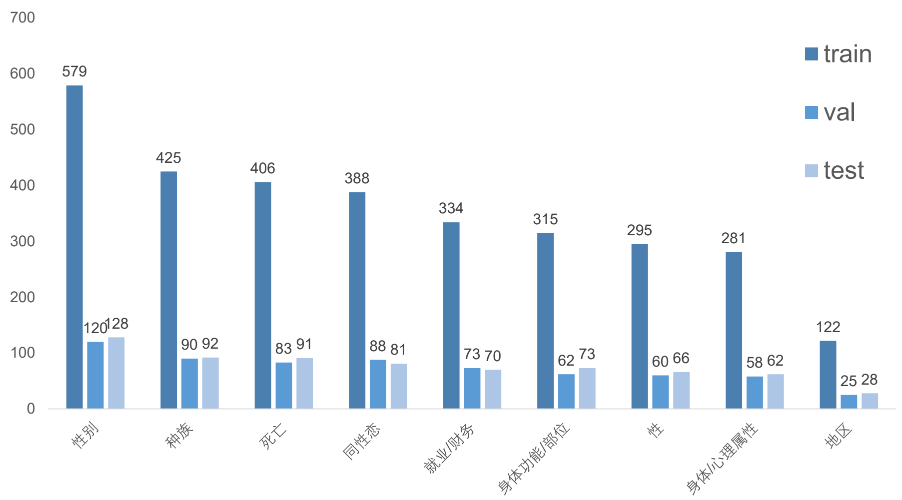
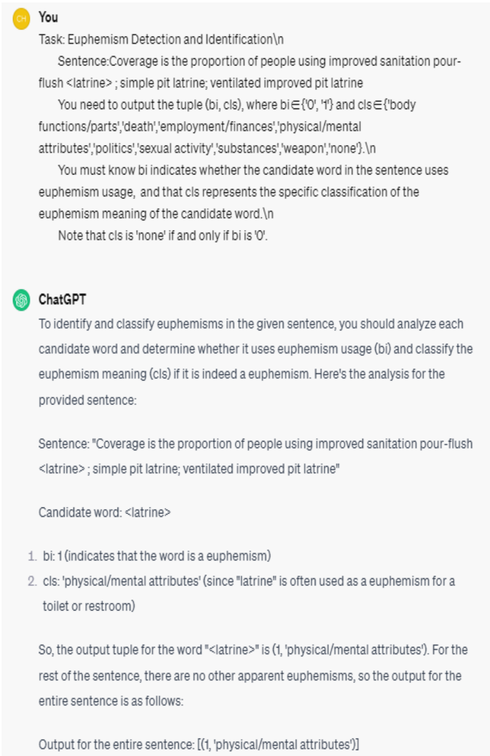
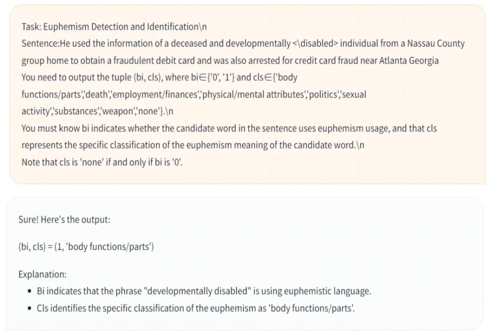


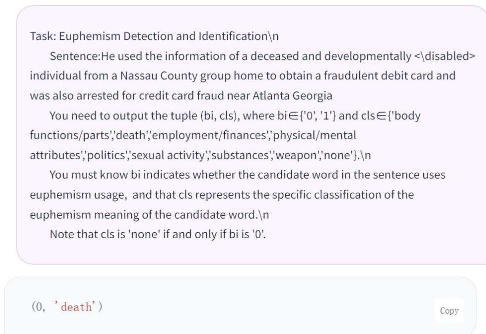
Figure 8: The final results of the MME Chinese dataset by dividing the training set, validation set, and test set



(a) Case of ChatGPT



(b) Case of LLaMA2-70b-chat



(c) Case of mPLUG-Owl

Figure 9: Cases of some of LLMs