A Report on the Euphemisms Detection Shared Task

Patrick Lee and Anna Feldman and Jing Peng
Montclair State University
New Jersey, USA
{leep6,feldmana,pengj}@montclair.edu

Abstract

This paper presents The Shared Task on Euphemism Detection for the Third Workshop on Figurative Language Processing (FigLang 2022) held in conjunction with EMNLP 2022. Participants were invited to investigate the euphemism detection task: given input text, identify whether it contains a euphemism. The input data is a corpus of sentences containing potentially euphemistic terms (PETs) collected from the GloWbE corpus (Davies and Fuchs, 2015), and are human-annotated as containing either a euphemistic or literal usage of a PET. In this paper, we present the results and analyze the common themes, methods and findings of the participating teams.

1   Introduction

Euphemisms are mild or indirect expressions that are used in place of other ones when discussing potentially offensive or sensitive topics. Their linguistic functions include (politeness, concealment, and neutralization of unappealing words/phrases). Despite being an important element of language use, their figurative nature poses a challenge for natural language processing (NLP).

There are numerous challenges to working with euphemisms. One is the phenomenon of the "euphemism treadmill", by which words/phrases gain or lose euphemistic meanings over time (Pinker, 2003). Another is that researchers may not agree on what euphemisms are. For example, Zhu and Bhat (2021); Zhu et al. (2021) treat code words as euphemisms, but our working definition does not. Even when restricted to our working definition, however, annotators were found to sometimes disagree in the task of labeling example sentences as euphemistic (Gavidia et al., 2022). For all these reasons, the words/phrases in this shared task are referred to as potentially euphemistic terms (PETs). The main challenge, which is the focus of this shared task, is the ambiguity of PETs: the same words/phrases that may be euphemistic in one context may be literal in another. For example,

Asked to choose between jobs and the environment, a majority – at least in our warped, first-past-the-post system – will pick jobs. (non-euphemistic)

This summer, the budding talent agent was between jobs and free to babysit pretty much any time. (euphemistic)

We propose the Shared Task on Euphemism Detection: given input text, identify whether it contains a euphemism; i.e., distinguish between euphemistic and literal usages of the same PETs in different contexts. The data used is a corpus of texts containing PETs, collected by Gavidia et al. (2022), which contains parallel euphemistic and literal examples for a range of PETs. 46 participants spanning 13 teams attempted the task, and we received 9 system descriptions.

Due to a lack of extensive research in this area, it is unclear how NLP techniques, such as language models (LMs), will perform on euphemism detection. The purpose of this shared task is to (1) explore the ability of NLP techniques for this task and (2) investigate what methods could further improve upon their performance.

2   Related Work

There is not much work on automatic detection of euphemisms. The most directly related work is by Magu and Luo (2018), Felt and Riloff (2020), Kapron-King and Xu (2021), Zhu et al. (2021) and Zhu and Bhat (2021). Felt and Riloff (2020) present the first effort to recognize euphemisms and dysphemisms (derogatory terms) using NLP. The authors use the term x-phemisms to refer to both. They first identify three sensitive topics (lying, stealing, and firing). They use a weakly supervised algorithm for semantic lexicon induction (Thelen

and Riloff, 2002) to generate lists of near-synonym phrases for each topic semi-automatically. Felt and Riloff (2020) experiment with two methods to classify phrases as euphemistic, dysphemistic, and neutral: 1) dictionary-based method addressing affect, connotation, intensity, arousal, and dominance; 2) contextual sentiment analysis to classify x-phemisms. The important product of this work is a gold-standard dataset of human x-phemism judgements. The important lesson here is that Felt and Riloff (2020) show that sentiment connotation and affective polarity are useful for identifying x-phemisms, but not sufficient. While the performance of Felt and Riloff (2020)'s system is relatively low and the range of topics is very narrow, this work certainly inspires further investigations.

Zhu et al. (2021) define two tasks: 1) euphemism detection (based on the input keywords, produce a list of candidate euphemisms) 2) euphemism identification (take the list of candidate euphemisms produced in (1) and output an interpretation). Zhu et al. (2021) select sentences matched by a list of keywords, create masked sentences (mask the keywords in the sentences) and apply the masked language model proposed in BERT (Devlin et al., 2018) to filter out generic (uninformative) sentences and then generate expressions to fill in the blank. These expressions are ranked by relevance to the target topic.

Euphemisms are also related to the language of politeness (e.g., Danescu-Niculescu-Mizil et al. (2013); Rababah (2014)), which plays a role in applications involving dialogue and social interactions in different contexts.

Other shared tasks have proposed a similar classification task on other types of figurative language. Ghosh et al. (2020) report on a sarcasm detection task run on conversation data from Twitter and Reddit, while Madabushi et al. (2022) report on an idiom detection and embedding task.

3   Task Setting

Participants were given a dataset of PET-containing texts created by Gavidia et al. (2022). In this section, we describe the dataset and the classification task.

3.1   Dataset Description

The corpus of PETs was formed by taking a list of PETs (single and multi-word expressions, collected from a variety of sources) and extracting texts from the GloWBe corpus (Davies and Fuchs, 2015) (only the US-English portion) which contained them. Each text sample comprised up to 3 sentences: the sentence that the PET appeared in, as well as the preceding and following sentences, if available. In total, the dataset contains 1,965 text samples spanning 129 different PETs and 7 topics/categories. Of these, 1,382 were annotated to contain a euphemistic usage of a PET, and the remaining 583 a literal usage. Thus, the dataset is imbalanced (an aspect which multiple teams explicitly considered in their approaches). The full details of the data, including the distribution amongst the PETs and topics, can be found in the original paper (Gavidia et al., 2022).

The training and test sets were created using an 80-20 split. The range of PETs which appeared in each split was balanced as much as possible, given that several PETs only appeared once as a euphemistic or literal example. Details of the split are summarized in Table 1.

Split	Rows	Euphemistic		Literal	
		Rows	PETs	Rows	PETs
Train	1572	1106	122	466	54
Test	393	276	121	117	55

Table 1: Train-Test Split Details

A simplified version of the dataset was created for the participants, where each row contained only (1) the text sample with the PET denoted in brackets, and (2) its label (a '1' for euphemistic, and a '0' for literal). This version omitted information about each row, such as meta-information about the PET: the specific morphological variant present in the row, the topic category (e.g., "death", "politics", etc.), and whether it always appeared in the dataset as a euphemism ("always-euphs") or only sometimes ("sometimes-euphs"). This information, however, was available to the participants via a Github link to the full dataset (which several teams chose to leverage).

3.2   Task Description

The shared task was set up as a competition on Codalab[1]. Participants were invited to develop systems trained on the training data (see Table 2 for some examples) and submit answer labels on the test set, which would be compared to the labels in

the original dataset for evaluation. The evaluation metric used to rank submissions was the macro-F1 score.

Text	Label
More likely it'll harm them. With less products to make, Foxxcom will have to <lay off> workers. With more workers seeking jobs those other factories will be able to resist demands for higher wages.	1
We do NOT need some self-imposed book cop telling us what to read or not to read. <Lay off>. Get over it.	0
After about 30 minutes of waiting, a fight broke out between an older African American man and an African American woman of <a certain age>. After making a lot of noise and landing a few blows to their respective bodies, the armed security guards escorted them out of the terminal.	1

Table 2: Example Rows from the Dataset

4   Participants and Results

8 teams that participated in the task also submitted descriptions of their systems, with one author additionally exploring a zero-shot/few-shot variant of the task. A summary of their performances is shown in Table 3. In this section, we describe the methods used by the best-performing teams, and analyze the common themes between the approaches and motivations of all the submissions.

4.1   Best Submissions

The best-performing team (Keh et al., 2022) (macro-F1 0.881) explores a variety of data and modeling modifications, and combine the best-performing ones into an ensemble of three models to improve upon a baseline RoBERTa-large model (Liu et al., 2019). On the data side, they explore two methods of data augmentation, and find that adding examples containing similar/opposite word senses to PETs (for positive and negative examples) works best; this highlights the potential significance of sense-based approaches for this task. They also explore identifying and correcting 25 potentially mislabeled rows from the dataset, reporting an improvement of 0.0036 points over the base model and a final score 0.007 higher when using

their "cleaned" dataset. While it is unclear how this cleaning would affect other systems in general, their investigation hints at the potential for not only human disagreement but also human error in labeling figurative language. On the modeling side, they find that classifying on the tokens of the PETs, rather than the [CLS] token, yields significant a performance increase. They also experiment with two methods of incorporating extra context and find that k-Nearest Neighbor (kNN) (5NN in this case) augmentation yields slight improvements. They report the best improvement by combining the following three models: (1) a RoBERTa-large model classifying on the PET token(s), (2) the same, but using their sense-augmented dataset, and (3) the same as (1), but interpolating the classification probabilities of its base model and the 5NN classifier.

The second-best performing team (Kesen et al., 2022) achieved a macro-F1 of 0.872 using additional supervision and, interestingly, incorporating visual imagery into their approach. Using DeBERTa-v3 (He et al., 2021) as their baseline (the "large" version of which performs best), they incor-porate additional supervision by including PETs themselves, as well as their (manually collected) literal descriptions, in their inputs. The authors found this to greatly improve performance, reason-ing that such direct supervision could help mitigate ambiguity inherent to the task. This is a similar result to (Keh et al., 2022), where extra attention on the PETs themselves is effective. They then obtain imageries of both the PETs and their literal descriptions using a text-to-image model, and ob-tain image embeddings using a pretrained visual encoder. These are incorporated into training, and yields statistically significant performance improve-ments. Qualitative analysis of the images for each PET reveals insights into how LMs might under-stand figurative expressions.

4.2   Analysis of Methods

Below, we describe approaches that we observed in multiple submissions. Since the objective was to explore different aspects of this task, we find these insights to be valuable, even if they did not score high.

4.2.1   Using PET Embeddings Directly

Multiple teams found that explicitly involving the tokens of the PET for classification led to significant improvement. Kesen et al. (2022) and Wang et al. (2022) include the PET in their inputs by con-

Rank	Username	Macro-F1	Title of Paper
1	vgangal	0.88	EUREKA: Euphemism Recognition Enhanced Through KNN-based Methods and Augmentation (Keh et al., 2022)
2	ilker	0.87	Detecting Euphemisms with Literal Descriptions and Visual Imagery (Kesen et al., 2022)
3	Wanderer	0.85	A Prompt Based Approach for Euphemism Detection (Maimaitituoheti, 2022)
4	liuyiyi	0.84	Euphemism Detection by Transformers and Relational Graph Attention Network (Wang et al., 2022)
5	peratham.bkk	0.82	TEDB System Description to a Shared Task on Euphemism Detection 2022 (Wiriyathammabhum, 2022)
6	PaulTrust	0.79	Bayes at FigLang 2022 Euphemism Detection shared task: Cost-Sensitive Bayesian Fine-tuning and Venn-Abers [...] (Trust and Kadusa, 2022)
7	devika	0.74	An Exploration of Linguistically-Driven and Transfer Learning Methods for Euphemism Detection (Tiwari and Parde, 2022)
8	gunetsk99	0.72	Adversarial Perturbations Augmented Language Models for Euphemism Identification (Kohli et al., 2022)

Table 3: Results of submitted systems to the Shared Task on Euphemism Detection

catenating it to each input text prior to learning. Keh et al. (2022) run their final classification on the PET embedding, rather than that of the [CLS] token. These changes were a significant feature of these teams' best approaches. It appears that providing direct supervision/focusing the modeling procedure on the PET helps, perhaps because the PET is the semantic focus of the task (rather than other words in the data, which are not always important).

4.2.2 Using the Literal Meanings of PETs

Each PET is detailed in Gavidia et al. (2022) to have a literal meaning or paraphrase, which is the more offensive or unpleasant "real meaning" that the PET substitutes. Several teams chose to integrate literal meanings of each input PET into their methods, though they opted to generate their own literal meanings, rather than use the ones from the original paper. This seemed to be effective, as the two best-performing teams found it to improve performance — Kesen et al. (2022) appended literal meanings to their inputs and used them to generate image embeddings, while Keh et al. (2022) used literal meanings to select examples for data augmentation — in conjunction with direct supervision on the PET (4.2.1). Tiwari and Parde (2022) paraphrased PETs with their literal meanings in attempt to obtain sentiment shifts, but this did not work well for classification, likely because the paraphras-

ing mechanism did not produce quality paraphrases that could serve as literal meanings for the PETs.

4.2.3 Addressing Data Imbalance/Inadequacy

Multiple teams addressed the fact that the dataset was imbalanced (see 1). Maimaitituoheti (2022) found that their approach scored an F1 of 0.914 versus 0.789 on the euphemistic and literal examples, respectively, suggesting that the model performance could indeed be skewed towards the higher-volume euphemistic examples.

(Trust and Kadusa, 2022) experimented with multiple modeling enhancements, such as Bayesian modeling, exclusively to address the imbalance issue and found that they improved performance. Kohli et al. (2022) sought to address the imbalance by using adversarial perturbations to augment the '0' label, and found it to increase performance slightly. As aforementioned, Keh et al. (2022) augment the dataset by strategically selecting sentences from other corpora (albeit for general augmentation, rather than to achieve a balance) and similarly report slight performance increases, but, like Kohli et al. (2022), they do not exclusively use the augmented data in their final approach. These results generally support the intuition that addressing data inadequacies helps models learn, but only partially for this shared task.

4.2.4 Incorporating Additional Context

While this task necessitates making use of contextual differences between input texts, several teams attempted to incorporate additional information from the context. Wang et al. (2022) model syntactic connections between the PET and other words in the text as a relational graph, finding that using this as an input to BERT is effective (though no baseline or example parse is provided). As aforementioned in Section 4.1, Keh et al. (2022) slightly improve performance by using the k-nearest neighbors for each input as additional context. Kohli et al. (2022) finds that simply using longer input sequence lengths than standard BERT allows for (512) generally improves performance, although it is not clear at what sequence length this would cease to be the case. All these methods to incorporate more input context seemed to generally improve performance.

4.2.5 Linguistic Transparency

Some teams attempted solutions that would promote model explainability. Kesen et al. (2022) claim that images of PETs and their literal meanings are a way to gain insight into how models interpret PETs, but while this is an interesting way to probe models' current understanding of figurative expressions, it is unclear how models might use these images to enhance classification in an understandable way. The two-model ensemble used by Wang et al. (2022) attempt to incorporate linguistic (semantic and syntactic) information of the PET and its context, though without compelling examples of how these may help classification, the improvements seem somewhat unexplainable. Finally, Tiwari and Parde (2022) interestingly pursued an approach that is based exclusively on the linguistic intuition that euphemisms produce sentiment shifts, but using these shifts alone for this task was ineffective.

On the other hand, methods which found success using transformers are not very transparent. Wiriyathammabhum (2022) test various transformers and obtain their best result by combining a CNN variant with the highest-performing one, but it is not clear what this network is learning, as is typically the case with neural networks. Furthermore, Keh (2022) find that transformers work decently for this task in the zero-shot setting (see 4.2.6), but admit that it's not clear what BERT is learning in order to do so.

4.2.6 Zero/Few-shot Learning

Maimaitituoheti (2022) train a RoBERTa model using prompt-tuning because it has been shown to work well (better than regular fine-tuning) with few-shot examples. While our task was not formulated as a zero/few-shot task, several PETs appeared only a few times in the data and were effectively few-shot examples. Keh (2022) notably re-formulate the task for the zero/few-shot setting. When PETs were randomly selected to be zero-shot examples, RoBERTa-large achieved a score of 0.740, showing that the model was able to "learn" something about euphemisms (not simply memorizing) and apply it to examples with PETs unseen during training. They also show that few-shot examples benefit the model greatly, with 3-shot performance (0.825) nearly matching the baseline performance (0.836). Furthermore, they found that GPT3, which typically works well in the zero/few-shot setting, worked badly for this task.

5 Discussion and Findings

Here, we describe some common findings of the submitted systems that may be useful for future work.

5.1 More Data is Better

Having more examples of each PET generally led to better performance. Kesen et al. (2022) and Keh et al. (2022) improved performance by augmenting the dataset. Maimaitituoheti (2022) showed that performance on the euphemistic label is better, likely because there were many more examples than the non-euphemistic label. Compellingly, Keh (2022) showed in their zero/few-shot task that 3-shot learning was much better than 1-shot. The results of this task call for larger datasets of euphemisms, perhaps from a variety of different sources.

5.2 BERT Works

All teams experimented with some variation of BERT and reported decent scores using unmodified BERT models, the highest being 0.839 (Kesen et al., 2022) using RoBERTa-large. The zero-shot investigation by Keh (2022), too, shows that RoBERTa picks up something during training that is generalizable to other euphemisms. Overall, pre-trained transformers seem to provide a solid baseline from which to launch euphemism work.

5.3 Linguistic Intuitions

Because euphemisms (as well as other types of figurative language) are often commonly used expressions, it is likely that large language models have some existing "knowledge" about these collocations. One could interpret the success of using PET embeddings directly (Kesen et al., 2022; Wang et al., 2022) as evidence that models can leverage this knowledge for the task.

Another linguistic notion is that euphemisms' function may lead to changes in sentiment, which has been found to potentially be useful for identifying euphemisms (Felt and Riloff, 2020; Lee et al., 2022), but it remains somewhat unclear whether it is useful for the disambiguation proposed in this task. Wiriyathammabhum (2022) do find that transformers pre-trained specifically on sentiment were more helpful than those pre-trained on other tasks (e.g., sarcasm or hate speech detection). Tiwari and Parde (2022) try a non-transformer-based approach based on the intuition that PETs should produce higher sentiment shifts in euphemistic sentences when paraphrased with its literal meaning, but found it was difficult to generate such paraphrases. This corroborates our own past experimentation (Lee et al., 2022), and it seems that future approaches based on sentiment shifts have to address the need for better paraphrasing mechanisms, or consider using them to supplement a larger input feature set.

6 Conclusions and Future Work

We present the results of "The Shared Task on Euphemism Detection for the Third Workshop on Figurative Language Processing" and summarize the various systems submitted, as well as common findings. Overall, we find that results are promising: even when dealing with the difficult issue of an especially polite and indirect form of figurative language, current NLP techniques such as transformers and augmentation seem to work quite well. Teams explored a variety of intriguing methods to enhance the baseline performance of these models, some of which were even linguistically transparent. If one considers that labeling euphemisms is subject to human disagreement, the F1-scores achieved by the teams are even more compelling since they may be near, if not already at, the level of human agreement on the task. The results of this shared task establish a baseline for future work on euphemisms and figurative language in general.

Future work on this task could be expanding on the dataset to include more examples and a wider range of PETs, testing further enhancements, and improving performance by ensembling various combinations of the best-performing improvements. Future work for euphemism detection in general could be to expand from the disambiguation task; e.g. identifying where euphemisms are in a text, providing interpretations of a euphemism, or even euphemistic language generation.

Limitations

While the data used denoted where the target PET was in each text sample, this information is not provided in raw text. Identifying the PET in a text sample is a challenge that future approaches, especially those seeking to focus models on the PET, will need to consider. Additionally, this shared task was run on a dataset that could be significantly expanded and balanced. The dataset also contained potentially subjective labels that were only made by two human annotators; this could be made more robust by ensembling more annotators. Finally, this task was based on a dataset comprising only of texts of US English, and it is unclear how these results would transfer cross-lingually to other kinds of euphemisms.

Ethics Statement

When we created the shared task, we tried to be compliant with the ACL Ethics Policy. Euphemisms are expressions that 'hide' prejudices by using softened language. Models capable of recognizing and interpreting euphemisms should be better at detecting biases related to gender, age, race, or socioeconomic background, detrimental to the society.

Acknowledgements

References

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. arXiv preprint arXiv:1306.6078.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). English World-Wide, 36(1):1–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543.

Anna Kapron-King and Yang Xu. 2021. A diachronic evaluation of gender asymmetry in euphemism. arXiv preprint arXiv:2106.02083.

Sedrick Scott Keh. 2022. Exploring Euphemism Detection in Few-Shot and Zero-Shot Settings. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Sedrick Scott Keh, Rohit K. 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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2022. Adversarial Perturbations Augmented Language Models for Euphemism Identification. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. Searching for pets: Using distributional and sentiment-based methods to find potentially euphemistic terms. arXiv preprint arXiv:2205.10451.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding.

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.

Abulimiti Maimaitituoheti. 2022. A Prompt Based Approach for Euphemism Detection. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Steven Pinker. 2003. The Blank Slate: The modern denial of human nature. Penguin.

Hussein Abdo Rababah. 2014. The translatability and use of x-phemism expressions (x-phemization): Euphemisms, dysphemisms and orthophemisms) in the medical discourse. Studies in literature and language, 9(3):229–240.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002), pages 214–221.

Devika Tiwari and Natalie Parde. 2022. An Exploration of Linguistically-Driven and Transfer Learning Methods for Euphemism Detection. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Paul Trust and Provia Kadusa. 2022. Bayes at FigLang 2022 Euphemism Detection shared task: Cost-Sensitive Bayesian Fine-tuning and Venn-Abers Predictors for Robust Training under Class Skewed Distributions. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan, and Jiafeng Guo. 2022. Euphemism Detection by Transformers and Relational Graph Attention Network. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Peratham Wiriyathammabhum. 2022. TEDB System Description to a Shared Task on Euphemism Detection 2022. In Proceedings of the 3rd Workshop on Figurative Language Processing. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model.

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. arXiv preprint arXiv:2103.16808.