An Exploration of Linguistically-Driven and Transfer Learning Methods for Euphemism Detection

Devika Tiwari and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois Chicago
{dtiwari, parde}@uic.edu

Abstract

Euphemisms are often used to drive rhetoric, but their automated recognition and interpretation are under-explored. We investigate four methods for detecting euphemisms in sentences containing potentially euphemistic terms. The first three linguistically-motivated methods rest on an understanding of (1) euphemism's role to attenuate the harsh connotations of a taboo topic and (2) euphemism's metaphorical underpinnings. In contrast, the fourth method follows recent innovations in other tasks and employs transfer learning from a general-domain pre-trained language model. While the latter method ultimately (and perhaps surprisingly) performed best (F_1 = 0.74), we comprehensively evaluate all four methods to derive additional useful insights from the negative results.

1 Introduction

Euphemism is a ubiquitous figurative language tool, wherein the speaker refers to taboo topics in indirect, metaphorical terms to convey politeness or formality. Identifying euphemism can reveal tacit facts about the speaker's intention and the context of the utterance (Gómez, 2009), but there has been minimal work exploring how this might be done computationally (Felt and Riloff, 2020; Gavidia et al., 2022). In this paper, we compare the performance of four methods for automated euphemism detection. The first two methods identify euphemism based on expected sentiment differences between euphemisms and their automatically generated non-euphemistic paraphrases. The third method exploits the metaphorical underpinnings of euphemism, following the hypothesis that the euphemism's root word will have more possible senses than its single-word literal paraphrase. In contrast to these linguistically-driven methods, the last method fine-tunes a popular pre-trained language model (Devlin et al., 2019, BERT) for the task of euphemism detection. We find strong, and perhaps surprising, evidence that the last method outperforms the alternatives.

Our contribution in this paper is twofold. First, we demonstrate the utility of pre-trained language models for novel figurative language processing tasks. Second, we demonstrate our process of translating linguistic theory of euphemism into empirical models. Although our results show that those methods need to be refined, it is our hope that this transparency will minimize redundancy in future research. Thus, our work is well-aligned with Nissim et al. (2017)'s position that reporting negative results in shared tasks can produce useful insights.

2 Related Work

2.1 Linguistic Theories of Euphemism

We frame our study of euphemism detection through the lens of established linguistic theory. Gómez (2009) explains euphemism from a cognitive and pragmatic perspective, emphasizing that euphemism suspends the negative connotations of taboo concepts to serve a discursive purpose within a given context. It is not merely a lexical substitution at the linguistic level; rather, it is a socially-motivated cognitive strategy that has the effect of signaling politeness to the interlocutor. Euphemism is thus characterized by both the speaker's intentional indirectness and the hearer's recognition of their attempt to veil the concept's offensiveness.

Fernández (2008) highlights that euphemism is almost always predicated on a metaphor. Using metaphor to express a taboo concept makes discussion of the taboo more permissible in public discourse. Hence, the function of euphemism is to neutralize a topic by speaking of it in vague terms. The ambiguity of the individual words in a euphemistic expression masks the overtly unacceptable features of the concept for which it stands.

These two theories delineate two hallmarks of euphemism: It produces a change in perceived sentiment, and it relies on an abstract metaphor to stand for a concrete concept. We use these linguistic facts as the foundation for our sentiment- and word sense-based solutions.

2.2 Euphemism Detection in NLP

Zhu and Bhat (2021) presented the first attempt at euphemistic phrase detection. From a raw text corpus of online posts, they mine euphemistic phrase candidates that represent target keywords and then apply a masked language model (MLM) based on SpanBERT (Joshi et al., 2019) to rank the candidate phrases in order of confidence. Their work was limited to euphemisms in the drug domain, with downstream applications in content moderation. In contrast, we designed our models to generalize to euphemism at large, independent of topic. We also employ MLMs in two of our methods, but for the purpose of generating single-word paraphrases, not to compute model confidence.

Recently, Gavidia et al. (2022) created the first corpus of sentences containing potentially euphemistic terms (PETs). To do so, they compiled a list of 184 PETs on a variety of taboo topics such as death, sexual activity, and substances. Then, they extracted sentences from the U.S.-dialect subsection of the Corpus of Global Web-Based English (Davies and Fuchs, 2015, GloWbE) that contained an instance of one of those PETs. PETs either did or did not function as a euphemism, dependent on context. They used RoBERTa-based sentiment analysis (Liu et al., 2019) to show that PETs function as euphemisms when replacing them with literal paraphrases causes an increase in negative and offensive sentiment. This work informed the sentiment-based technique in two of our methods. Subsequently, Lee et al. (2022) expanded on their work by developing a method that mines single and multi-word expressions, filters them based on similarity to sensitive topics, and identifies the euphemistic PETs based on the phrases that caused the greatest sentiment shift when paraphrased.

3 Dataset and Task Description

Our work was conducted as part of a shared task with the goal of creating a system that determines whether or not a given sentence containing a PET is euphemistic. The data was sourced from Gavidia et al. (2022)'s corpus of PETs. The training dataset consisted of 1572 utterances, with PETs demarcated within angled brackets. An *utterance* was defined as the sentence containing the PET along with the preceding and following sentences to provide additional context. Utterances were assigned labels of 1 or 0, with 1 indicating that the PET was euphemistic and 0 indicating that it was not. A condensed example of an entry in the training dataset is shown in Table 1. The test dataset consisted of 393 unlabeled utterances. Similar to the training data, each utterance included three sentences, with the PET denoted within angle brackets.

Index	Utterance	Label
81	...locked up in a military <detention camp> on vague charges of being a Terrorist sympathizer...	1

Table 1: Sample entry from the training dataset.

4 Methods

We explored four methods for euphemism detection, broadly categorized by their reliance on engineered, linguistically-driven features or transfer learning. In the first two methods, we expected that if the original sentence contained a euphemism, then substituting the PET with a synonymous non-euphemistic term should produce a difference in the sentiment between the original and the generated, paraphrased sentence. The third approach relies on the premise that euphemisms are metaphorical extensions of the head of the phrase, while their non-euphemistic paraphrases have more specific semantic scope. The fourth approach employs BERT, a popular transformer-based model that we fine-tuned to detect euphemism. We provide further details regarding the intuition and implementation guiding each of these approaches in §4.1-4.3.

4.1 Sentiment-based Methods

Consider the PET *armed conflict* for which the non-euphemistic paraphrase is *war*. *Armed conflict*, more indirect and ambiguous, evokes less negative and offensive sentiment than its synonym *war*, which is more direct and richer in emotional content. On the other hand, consider the sentence *Her ideas were <underdeveloped>*. In this context, the PET *underdeveloped* is not functioning as a euphemism. Substituting it with a non-euphemistic paraphrase such as *weak* has little effect on the sentence's sentiment. Following this, the underlying

Feature	Description
Negative_Diff	$\textsc{SentDiff}(o, p)$ when measuring $\mathrm{S}_d(\cdot)$ along the *negative* dimension (*d=negative*).
Neutral_Diff	$\textsc{SentDiff}(o, p)$ when *d=neutral*.
Positive_Diff	$\textsc{SentDiff}(o, p)$ when *d=positive*.
Offensive_Diff	$\textsc{SentDiff}(o, p)$ when *d=offensive*.

Table 2: Sentiment-based features computed based on measured differences in negative, neutral, positive, and offensive sentiment between the original and paraphrased versions of the sentence.

intuition guiding our sentiment-based methods was that there may be a greater difference in sentiment between the original sentence and the paraphrase when the original PET was euphemistic.

4.1.1 Paraphrasing Using Back-Translation

We used back-translation between English and German to generate the paraphrase for each utterance, implemented using Ma (2019)'s NLP augmentation (`nlpaug`) library. We anticipated that the original sentence would lose many figurative elements through the process of back-translation, leading the PET to be replaced by a semantically consistent but literal paraphrase. We then computed the difference in sentiment between the original sentence o and back-translated paraphrase p, where $\mathrm{S}_d(\cdot)$ is a measure of sentiment for a given input along a specific dimension d:

$$\textsc{SentDiff}(o, p) = \mathrm{S}_d(o) - \mathrm{S}_d(p) \qquad (1)$$

Sentiment was measured along five dimensions (Lee et al., 2022, negative, neutral, positive, non-offensive, and offensive) using the RoBERTa (Liu et al., 2019) sentiment and offensiveness models. We used differences in negative, neutral, positive, and offensive sentiment as features (Table 2) for a logistic regression model to classify sentences as euphemistic or non-euphemistic. We used Python's `scikit-learn` library[1] to implement our classifier, leaving all hyperparameters at their default values.

4.1.2 Paraphrasing Using MLM

As an alternative to back-translation, we also generated paraphrases using MLM and masking out PETs. Because MLM accounts for sentence context, we expected that the tokens replacing the PET would be influenced by the overall sentiment of the sentence. Thus, if the context was indicative of taboo or sensitive content, then the MLM's suggestions should reflect that sentiment. From the set of suggested replacements for each PET, we selected the token that was most similar in meaning to the original PET. To do this, we generated an embedding for the original PET and each of the token suggestions using the Sentence Transformers framework (Reimers and Gurevych, 2019). We ignored MLM tokens that were either stopwords or identical to the original PET.

We selected the MLM token that had the highest cosine similarity to the PET, with the expectation that this token would be a non-euphemistic paraphrase of it. The selected token was substituted for the PET in the original sentence. We then calculated negative, neutral, positive, and offensive sentiment differences between the original sentence and the paraphrase as explained in §4.1.1, using those shifts as features for classification.

4.2 Word Sense-based Method

In the third approach, rather than analyzing sentiment differences, we compared the number of word senses between the syntactic head of the PET and its single-word non-euphemistic paraphrase. Consider the euphemism *expecting* used instead of *pregnant*. *Expect*, the lemma of the euphemism, has much wider semantic scope than *pregnant*. In replacing a very specific term with a more vague, metaphorical one, euphemism functions to reduce the explicitly taboo undertones of the target concept (Fernández, 2008). We captured this apparent ambiguity of the euphemistic term compared to the concreteness of its non-euphemistic paraphrase through measured polysemy. The euphemism is expected to be built on a word with more senses than the non-euphemistic word it replaces.

The non-euphemistic paraphrase of the PET in each utterance was determined using the same MLM technique described in §4.1.2. Because the PET can be a multi-word expression, and senses are counted for individual words, we extracted the syntactic head of each PET. If the PET was a single word, then the head was the word itself. Otherwise,

[1] `https://scikit-learn.org/stable/`

the head was identified as the root token of the PET's dependency parse (predicted using Python's spaCy[2] library). For example, if the euphemism *lay off* was used in the context of firing employees, then the head of the PET would be the verb *lay*.

We used WordNet (Fellbaum, 1998) to find the number of word senses for the lemmas of both the chosen MLM token and the head of the original PET. If a lemma did not appear in the WordNet dictionary, then its number of senses was set to one. The number of word senses of the head of the PET and of the chosen MLM token were used as features for a logistic regression model to classify the test utterances as euphemistic or non-euphemistic. Similarly to our first approach, we used Python's scikit-learn library[3] to implement our logistic regression classifier, with default hyperparameters.

4.3 Transfer Learning Method

Our final method was a fine-tuned BERT (Devlin et al., 2019) model. Specifically, we fine-tuned the bert-base-cased pre-trained model from Hugging Face[4] for euphemism detection using the Trainer API. The model was pretrained on data from BookCorpus (Zhu et al., 2015) and English Wikipedia.[5] We anticipated that this model would offer a strong baseline to which the other models could be compared, while also facilitating study into the extent that general-domain pre-training data can be leveraged for this task. Input was tokenized using AutoTokenizer, also from the Hugging Face library. We set the model to pad shorter input sequences to the maximum sequence length, and truncate longer input sequences to the maximum acceptable input length for the model (512 tokens).

5 Evaluation

We compared the performance of all methods using precision, recall, and F_1-measure, following task guidelines. The sentiment-based methods described in §4.1 were excluded from our shared task submission and thus not evaluated on the test data, due to their observed under-performance during validation experiments. Our validation experiments were evaluated using a withheld subset of 20% of the training data. In Table 3, we report all models' performance on the the validation set, enabling comparison between all techniques described in §4. In Table 4, we report the performance of the two top-performing methods, *Word Sense* (§4.2) and *BERT* (§4.3), on the test dataset as evaluated by the shared task submission portal.

Method	**P**	**R**	**F_1**
Sentiment-BT	0.34	0.50	0.41
Sentiment-MLM	0.35	0.50	0.41
Word Sense	0.59	0.51	0.44
BERT	**0.83**	**0.92**	**0.87**

Table 3: Performance comparison among all models on a held-out validation subset of the training data.

Method	**P**	**R**	**F_1**
Word Sense	0.50	0.55	0.43
BERT	**0.74**	**0.75**	**0.74**

Table 4: Performance comparison among shared task submissions on the test data.

6 Discussion

The results show that *BERT* unquestionably outperforms the sentiment- and word sense-based methods. This illustrates that a fine-tuned model pretrained on general-domain data can be successfully leveraged for euphemism detection. Close inspection of the predictions from the three linguistically-driven methods revealed that they overwhelmingly classified sentences as euphemistic. We suspect that they learned to reliably detect the presence of figurative language but require further refinement to discriminate between euphemism and other figurative language phenomena (e.g., metaphor).

Sentiment-BT likely under-performed because we found that PETs remained surprisingly intact through the process of back-translation. Hence, there were few sentiment differences between the original and paraphrased sentences. Similarly, the tokens selected in *Sentiment-MLM* may have fit the sentence context but were not literal paraphrases of the PET. Beyond *Sentiment-MLM*, this may also explain the failure of *Word Sense* relative to *BERT*. If the paraphrases themselves are unreliable, then it entails that subsequent downstream comparisons of sentiment or polysemy between the original and paraphrased sentences will also be inaccurate.

[2] https://spacy.io/
[3] https://scikit-learn.org/stable/
[4] https://huggingface.co/bert-base-cased
[5] https://en.wikipedia.org/

7 Conclusion

In this paper, we explored linguistically-driven and transfer learning methods to detect euphemism. Our linguistically-driven methods drew upon differences in sentiment and word sense frequency between euphemisms and their paraphrases. Our transfer learning method fine-tuned BERT for euphemism detection and proved to be the most successful. We motivate our sentiment- and word sense-based methods using linguistic theory and report their results despite under-performance to highlight the scope for future improvement. In our next steps, we aim to devise techniques for more accurately paraphrasing euphemisms (simultaneously driving the dial forward towards *euphemism understanding*), allowing us to further investigate linguistically-driven approaches. We will also study whether fine-tuning source models intended for metaphor detection or sentiment analysis will further improve upon our transfer learning results.

Limitations

We acknowledge that the linguistically-driven models in this paper are only applicable to data where the PET has been explicitly demarcated. To deploy these models in a real-world setting, we would have to create a system that is capable of not only detecting the presence of a euphemism but can identify it from data that has not been annotated.

Furthermore, in addition be being limited to euphemisms in English, our proposed models are trained only on American dialectal data. This calls into question the cross-cultural validity of our models. Specifically, the target concepts that necessitate euphemism and the metaphors that those euphemisms are built upon are culturally-dependent constructs, posing a challenge for building generalizable euphemism detection models.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback, and the shared task organizers for encouraging research on this exciting and challenging problem. This work was completed as part of an Honors College capstone project at the University of Illinois Chicago.

References

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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, Online. Association for Computational Linguistics.

Eliecer Crespo Fernández. 2008. Sex-related euphemism and dysphemism: An analysis in terms of conceptual metaphor theory. *Atlantis*, 30(2):95–110.

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 Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Miguel Gómez. 2009. Towards a new approach to the linguistic definition of euphemism. *Language Sciences*, 31:725–739.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Edward Ma. 2019. NLP augmentation. https://github.com/makcedward/nlpaug.

Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Last words: Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.