Detecting Euphemisms: Analysis of Potentially Euphemistic Terms Through Op-Eds

Yasmine Chim
yc947@cornell.edu

Abstract

The study of euphemisms from a natural language processing (NLP) framework has generally been application dependent– how we can detect them, how we can use them, and how accurate models are. Here, I hope to draw on the automation of detection efforts with an eye towards recognizing euphemisms and dysphemisms (derogatory terms) using NLP. I utilized Gavidia et. al's potentially euphemistic terms (PETs) corpus, a list of 284 PETs contextualized, as well as their example text corpus. From there, I built my own (one corpus utilizing 674 sentences from 371 articles from 2023 and another utilizing 321 sentences from 181 articles from 1982-1983). What was found was that this detector was apt at finding euphemisms falling in the categories: physical/mental attributes, politics, death, and employment. We see this track for both recent and archival op-eds, however, what's interesting is the potentially euphemistic term that is more common according to time period. In the 80s, usage of the pet "african americans" was less likely, while usage of the pet "elderly" was more commonplace. The PETs correctly identified in the archival data has largely been related to age, whereas PETs correctly identified in the recent data can be more associated with "political correctness": "low-income", "people of color", "aging", and "african americans" are all ranked among the most frequent PETs.

Introduction:

As literary devices, euphemisms replace negative terms with more positive expressions, altering the context favorably. Usage attempts politeness, substituting one phrase for another in order not to possibly offend the reader or listener. It's in this vein that we can see euphemisms as culturally situative; those who use them are familiar with the places where they're used, are at the behest of not offending the listening party, are privy to potential blowback from using an insult.

Euphemisms are altogether not a good or bad thing– they're necessary to function in a society and reveal much about what and who we value. The power of euphemisms exists in how they function– to inflate and magnify or to deflate and diminish. Hugh Rawson's Dictionary of Euphemisms and Other Doubletalk highlights this and provides a unique way to calculate the magnitude of euphemisms. As it pertains to class– pay, type of occupation, and employer– euphemisms, as articulated by Rawson, can imply prevalence according to circumstance.

As a decade, the 80s can be remembered as one class obsessive. Following Vietnam and welcoming Reagan, we can look towards a popular genre at the time; satirical reference guides. Bestsellers in the early 80s included the *Official Preppy Handbook* by Lisa Birnbach and *Class: A Guide Through the American Status System* by Paul Fussell; the former inspiring the founding of J. Crew. In these books, wealthy is derided in favor of rich, people of color are magnanimously ignored, and readers of the New York Times are deemed middle class.

In this paper, I hope to borrow from scant euphemism research compiled by Gavidia et. al– namely their PET corpus and their sentence corpus. I web scraped NYT op-eds, extracted the PETs from the aforementioned corpus into sentences, and built my corpus. The rest of the paper is as follows: Section 2 reviews previous work done on euphemisms and metaphors– how it pertains to language and use cases. Section 3 gives details on how I assembled my corpus and utilized Gavidia et. al's corpus. In Section 4, I describe my own corpus and how the model performed with non-euphemisms vs euphemisms. Section 5 discusses experimental results of classification with DistilBertForSequenceClassification– what insights I derived and what they could say about the NYT's op-ed section from then to now. Section 6 would introduce potentials for future research into euphemisms.

Section 2:

The study of euphemisms from a natural language processing (NLP) framework has generally been application dependent– how we can detect them, how we can use them, and how accurate models are. Here, I hope to draw on the automation of detection efforts with an eye towards recognizing euphemisms using NLP.

Much of this research draws on Gavidia et. al's *CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms*, where they found that the introduction of PETs to sentences would decrease offensiveness and negative sentiment and that it was difficult for computers and human annotators to disambiguate between commonly accepted PETs. In their sentiment analysis effort, Gavidia et. al engaged RoBERTa to perform sentence masking (swapping in euphemisms for potentially offensive terms, eg. pass away for dying).

There they compiled a set of commonly used PETs, one namely being Rawson's Dictionary of Euphemisms and Other Doubletalk. While they express the limitations of their PET corpus in their literature and the "euphemism treadmill"-- "how euphemisms can sometimes become offensive over time and thus lose their euphemism status (Pinker, 2003)", they don't include use cases where that's apparent. I hope to include the NYT's op-eds to provide a before and after snapshot of PET usage. Given the attention paid to sociological lexicography in the 80s, I hope to see if that does become apparent in the literature.

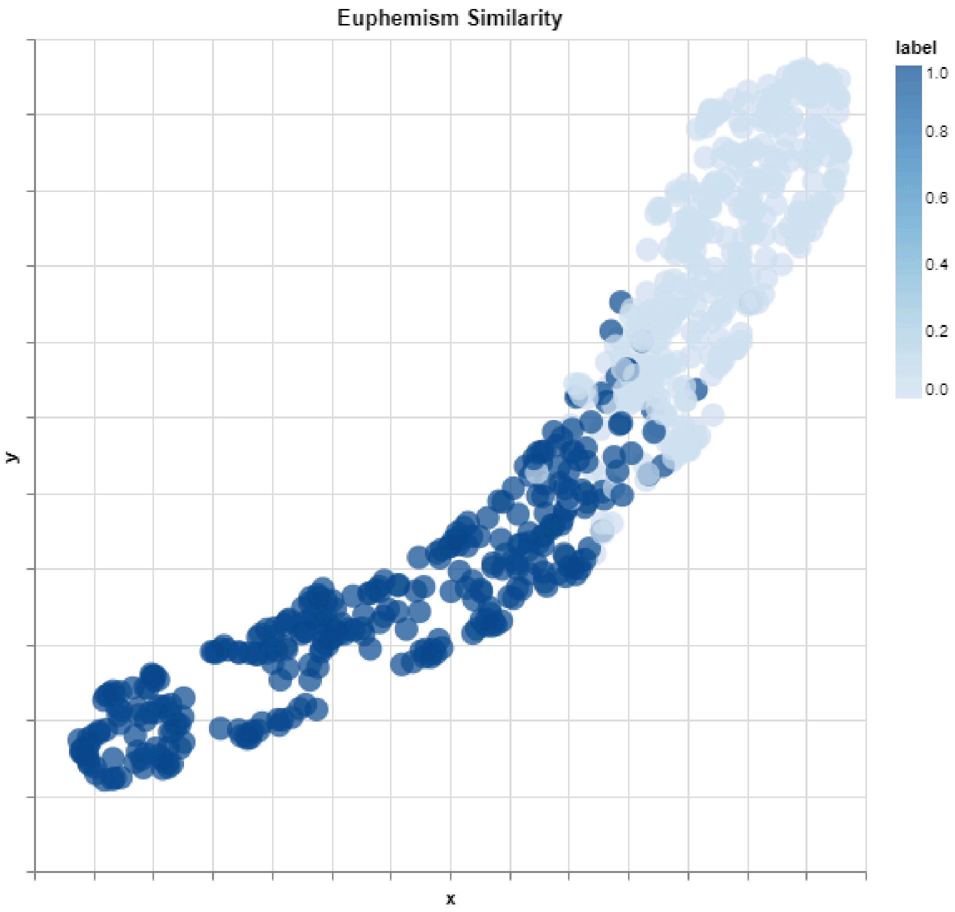
Section 3:

For recent NYTs op-ed, for expediency's sake, I used previously scraped op-eds from another class. For previous NYT's op-eds, given my dating constraint (as close as possible to 1980, without crossing over a threshold at which point each article was $3.95), I collected one-thousand article links and parsed the ones which included both author and title– yielding 490 files.

From there, for both, I found euphemistic sentences in all of these articles, where the list of 284 PETs were included in one sentence and I split them up. From there, I yielded 674 sentences from 371 recent articles and another utilizing 321 sentences from 181 articles from 1982-1983.

Instead of using RoBERTa for sentiment analysis like Gavidia et. al did, I decided on using a binary classifier with DistilBERTforSequenceClassification– for both is_euph and cat_pred. Given this was text data, I decided on using an encoder and decoder transformer model– fine tuning with their euphemism data. For deciding on what was a euphemism, I performed a train/test split on the euphemisms dataset that Gavidia et. al provided, receiving a 72% accuracy score. This was in line with their average observed percent agreement of 71.74%.

When performing sentence embeddings for my model's tokenizer, there seemed to be some type of clear delineation between euphemistic sentences and non euphemistic sentences.



Afterwards, I fine tuned eight DistilBERTforSequenceClassification models for the potentially offensive categories: death, employment, bodily functions, physical/mental attributes, politics, sexual activity, substances, and miscellaneous. In doing so, it revealed a poor model-wide categorical detection effort:

```
Accuracy for category bodily functions: 0.9898
Accuracy for category death: 0.9135
Accuracy for category employment: 0.8601
Accuracy for category misc.: 0.9822
Accuracy for category physical/mental attributes: 0.9466
Accuracy for category politics: 0.9644
Accuracy for category sexual activity: 0.9440
Accuracy for category substances: 0.9746
Precision: 0.5652
Recall: 0.4285
F1 Score: 0.4809
```

While the model performed well for each category, when it came to disambiguating between categories, the model fell flat.

Section 4:

When deployed on my assembled corpus, I made sure to hand label which PETs were euphemistic or not and I selected which categories they belonged to. In this, when comparing the NYT's corpus to Gavidia et. al's corpus, my suspicions regarding accuracy were confirmed. I was not surprised that recent articles were more incorrectly classified, due to the assemblage of the PET corpus. The archival dataset used more euphemistic terms that fell more in line with PET terms that were labeled as "always_euph" like "elderly". Nonetheless, I was surprised to find that the sentences performed just as well on my archival dataset as the euphemistic corpus, because that was assembled from "spaCy's PhraseMatcher (Honnibal and Montani, 2017) to identify rows from our raw text data which contain terms" from the 284 PET list. The category evaluation for both corpuses were the same. I'd assume it has to do with the PETs selected by Gavidia et. al, then selecting late as a PET to substitute for dead, would likely lead to many false positives.

NYT Euphemism Evaluation

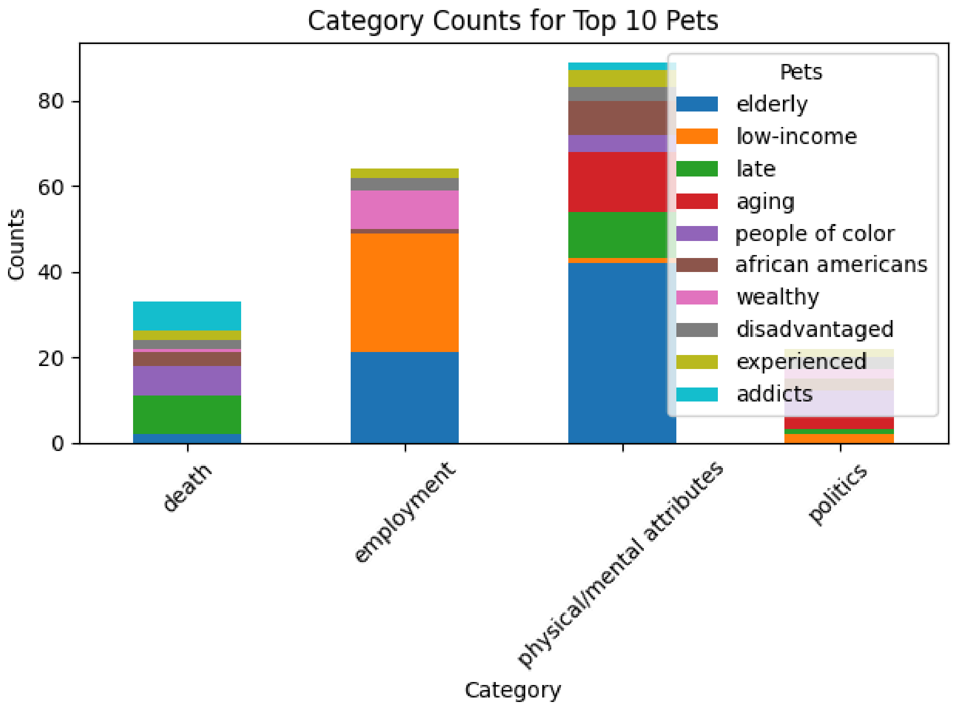
```
Accuracy: 0.6443452380952381
Classification Report:
              precision    recall  f1-score   support

         0.0       0.51      0.70      0.59       248
         1.0       0.78      0.61      0.68       424

    accuracy                           0.64       672
   macro avg       0.65      0.66      0.64       672
weighted avg       0.68      0.64      0.65       672

Confusion Matrix:
 [[174  74]
 [165 259]]

NYT Category Evaluation
Accuracy: 0.5342261904761905
Classification Report:
                            precision    recall  f1-score   support

          bodily functions       0.00      0.00      0.00         8
                     death       0.12      0.69      0.20        36
                employment       0.65      0.70      0.67        80
                      misc       0.00      0.00      0.00        92
physical/mental attributes       0.73      0.72      0.72       225
                  politics       0.76      0.58      0.66       206
           sexual activity       0.00      0.00      0.00        12
                substances       0.00      0.00      0.00        14

                  accuracy                           0.53       672
                 macro avg       0.25      0.30      0.25       672
              weighted avg       0.55      0.53      0.53       672
```

NYT Archive Euphemism Evaluation

```
Accuracy: 0.7366771159874608
Classification Report:
              precision    recall  f1-score   support

         0.0       0.55      0.90      0.68        99
         1.0       0.94      0.66      0.78       220

    accuracy                           0.74       319
   macro avg       0.74      0.78      0.73       319
weighted avg       0.81      0.74      0.75       319

Confusion Matrix:
 [[ 89  10]
 [ 74 146]]
```

```
NYT Category Evaluation
Accuracy: 0.525
Classification Report:
                             precision    recall  f1-score   support

          bodily functions       0.00      0.00      0.00         2
                     death       0.22      0.72      0.34        29
                employment       0.62      0.72      0.67        69
                      misc       0.00      0.00      0.00        24
physical/mental attributes       0.68      0.43      0.53       106
                  politics       0.67      0.62      0.65        82
            sexual activity       0.00      0.00      0.00         3
                substances       0.00      0.00      0.00         5

                  accuracy                           0.53       320
                 macro avg       0.27      0.31      0.27       320
              weighted avg       0.55      0.53      0.52       320
```

Section 5:



Ultimately, some categories were weighted more heavily towards archival versus non-archival op-eds, namely employment. Given the smaller sample size from the archival set, I was surprised to see more euphemisms being used in an employment context. Upon further evaluation, much of the terms for employment were euphemisms like “elderly”. Archival op-eds from the early 80s talked much about an aging workforce– most likely using elderly to speak about worker conditions more politely, while recent op-eds talk a lot more about race and income inequality. The contexts in which older people were being spoken of in recent op-eds happened largely for topics such as COVID-19 and Joe Biden.

Moreover, I think there was a lot of imbalance when the model determined which terms were euphemisms and which were not. Terms like “african american”, “elderly”, “low income” clearly have a sociological, politically correct skew which may seem dated if considered euphemistic. Moreover, the sum of actual euphemistic terms per sentences is interesting, because despite the drastically different number of sentences: 674 sentences from 371 articles from 2023 and another utilizing 321 from the archive, the numbers of actual euphemisms, on average, is higher from the archive. I suspect this may be due to a dated PET corpus.

```
Recent    259.0
Archive    146.0
```

Conclusion:

In this paper, I fine tuned a sequence classification model and applied it across categories of PETs and to determine the euphemistic-ness of said PETs within the context of sentences. Drawbacks included the small PET corpus size, the limitations of using only NYT op-eds, and small corpus size. Nonetheless, I think adding a larger and labeled sentence corpus provides additional avenues for research down the line.

Citations:

Gavidia, M., Lee, P., Feldman, A., & Peng, J. (2022). CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. arXiv. https://arxiv.org/abs/2205.02728

Yandell, James. “Things Too Fierce To Mention.” *The San Francisco Jung Institute Library Journal*, vol. 9, no. 3, 1990, pp. 35–46. *JSTOR*, https://doi.org/10.1525/jung.1.1990.9.3.35. Accessed 20 May 2024.