Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis

Christian Felt and Ellen Riloff
School of Computing
University of Utah
christianfelt@comcast.net, riloff@cs.utah.edu

Abstract

This paper presents the first research aimed at recognizing euphemistic and dysphemistic phrases with natural language processing. Euphemisms soften references to topics that are sensitive, disagreeable, or taboo. Conversely, dysphemisms refer to sensitive topics in a harsh or rude way. For example, *"passed away"* and *"departed"* are euphemisms for death, while *"croaked"* and *"six feet under"* are dysphemisms for death. Our work explores the use of sentiment analysis to recognize euphemistic and dysphemistic language. First, we identify near-synonym phrases for three topics (FIRING, LYING, and STEALING) using a bootstrapping algorithm for semantic lexicon induction. Next, we classify phrases as euphemistic, dysphemistic, or neutral using lexical sentiment cues and contextual sentiment analysis. We introduce a new gold standard data set and present our experimental results for this task.

1 Introduction

Euphemisms are expressions used to soften references to topics that are sensitive, disagreeable, or taboo with respect to societal norms. Whether as a lubricant for polite discourse, a means to hide disagreeable truths, or a repository for cultural anxieties, veiled by idioms so familiar we no longer think about what they literally mean, euphemisms are an essential part of human linguistic competence. Conversely, **dysphemisms** make references more harsh or rude, often using language that is direct or blunt, less formal or polite, and sometimes offensive. For example, *"passed away"* and *"departed"* are common euphemisms for death, while *"croaked"* and *"six feet under"* are dysphemisms for death. Table 1 shows examples of euphemisms and dysphemisms across a variety of topics.

Following terminology from linguistics (e.g., (Allan, 2009; Rababah, 2014)), we use the term **x-phemism** to refer to the general phenomenon of euphemisms and dysphemisms. Recognizing x-phemisms could be valuable for many NLP tasks. Euphemisms are related to politeness, which plays a role in applications involving dialogue and social interactions (e.g., (Danescu-Niculescu-Mizil et al., 2013)). Dysphemisms can include pejorative and offensive language, which relates to cyberbullying (Xu et al., 2012; Van Hee et al., 2015), hate speech (Magu and Luo, 2014), and abusive language (Park et al., 2018; Wiegand et al., 2018). Recognizing euphemisms and dysphemisms for controversial topics could be valuable for stance detection and argumentation in political discourse or debates (Somasundaran and Wiebe, 2010; Walker et al., 2012; Habernal and Gurevych, 2015). In medicine, researchers found that medical professionals use x-phemisms when talking to patients about serious conditions, and have emphasized the importance of preserving x-phemisms across translations when treating non-English speakers (Rababah, 2014).

An area of NLP that relates to x-phemisms is sentiment analysis, although the relationship is complex. A key feature of x-phemisms is that their directionality (euphemism vs. dysphemism) is relative to an underlying topic, which itself often has affective polarity. X-phemisms are usually associated with negative topics that are culturally disagreeable or have a negative connotation, such as death, intoxication, prostitution, old age, mental illness, and defecation. However x-phemisms also occur with topics that are sensitive but not inherently negative, such as pregnancy (e.g., *"in a family way"* is a euphemism, while *"knocked up"* is a dysphemism). In general, dysphemistic language increases the degree of sensitivity, intensifying negative polarity or shifting polarity from neutral to negative. Conversely, euphemistic language generally decreases sensitivity. But euphemisms for inherently negative topics may still have negative polarity (e.g.,

Topic	Euphemisms	Dysphemisms
DEATH	passed away, eternal rest, put to sleep	croaked, six feet under, bit the dust
INTOXICATION	tipsy, inebriated, under the influence	hammered, plastered, sloshed, wasted
LYING	falsehood, misrepresent facts, untruth	bullshit, rubbish, whopper, quackery
PROSTITUTE	lady of the night, working girl, sex worker	whore, tart, harlot, floozy
DEFECATION	bowel movement, number two, pass stool	take a dump, crap, drop a load
VOMITING	be sick, regurgitate, heave	blow chunks, puke, upchuck

Table 1: Examples of Euphemisms and Dysphemisms

vomiting is unpleasant no matter how gently it is referred to).

This paper presents the first effort to identify euphemistic and dysphemistic language in text. Since affective polarity clearly plays a role in this phenomenon, our research explores whether sentiment analysis can be useful for recognizing x-phemisms. We deconstructed the problem into two subtasks. First, we identify phrases that refer to three sensitive topics: LYING, STEALING, and FIRING (job termination). We use a weakly supervised algorithm for semantic lexicon induction (Thelen and Riloff, 2002) to semi-automatically generate lists of *near-synonym phrases* for each topic. Second, we investigate two methods to classify phrases as *euphemistic*, *dysphemistic*, or *neutral*[1]. (1) We use dictionary-based methods to explore the value of several types of information found in sentiment lexicons: affective polarity, connotation, intensity, arousal, and dominance. (2) We use contextual sentiment analysis to classify x-phemism phrases. We collect sentence contexts around instances of each candidate phrase in a large corpus, and assign each phrase to an x-phemism category based on the polarity of its contexts. Finally, we introduce a gold standard data set of human x-phemism judgments and evaluate our models for this task. We hope that this new data set will encourage more work on x-phemisms. Our experiments show that sentiment connotation and affective polarity can be useful for identifying euphemistic and dysphemistic phrases, although this problem remains challenging.

2 Related Work

Euphemisms and dysphemisms have been studied in linguistics and related disciplines (e.g., (Allan and Burridge, 1991; Pfaff et al., 1997; Rawson, 2003; Allan, 2009; Rababah, 2014)), but they have received little attention in the NLP community. Magu and Luo (2014) recognized code words in "euphemistic hate speech" by measuring cosine distance between word embeddings. But their code words conceal references to hate speech rather than soften them (e.g., the code word *"skypes"* covertly referred to Jews), which is different from the traditional definition of euphemisms that is addressed in our work.

The NLP community has explored several linguistic phenomena related to x-phemisms, such as metaphor (e.g., (Shutova, 2010; Wallington et al., 2011; Shutova et al., 2010; Kesarwani et al., 2017)), politeness (e.g., (Danescu-Niculescu-Mizil et al., 2013; Aubakirova and Bansal, 2016)), and formality (e.g., (Pavlick and Tetreault, 2016)). Pfaff et al. (1997) found that people comprehend metaphorical euphemisms or dysphemisms more quickly when they share the same underlying conceptual metaphor. For example, people are likely to use the euphemism *"parted ways"* to describe ending a relationship in the context of the conceptual metaphor A RELATIONSHIP IS A JOURNEY, but more likely to use the euphemism *"cut their losses"* in the context of the metaphor A RELATIONSHIP IS AN INVESTMENT.

Our research focuses on the relationship between x-phemisms and sentiment analysis. We take advantage of several existing sentiment resources, including the NRC EmoLex, VAD, and Affective Intensity Lexicons (Mohammad and Turney, 2013; Mohammad, 2018a,b) and Connotation WordNet (Feng et al., 2013; Kang et al., 2014). We also re-implementated the NRC-Canada sentiment classifier (Mohammad et al., 2013) to use in our work.

Allan (2009) examined the connotation of color terms according to how often they appear in dysphemistic, euphemistic, or neutral contexts. For instance, *"blue"* is often used as a euphemism for *"sad"*, while *"yellow"* can be dysphemistically used to mean *"cowardly"*. Our paper takes the reverse approach, recognizing x-phemisms by means of

[1]Direct ("straight-talking") references to a topic are called *orthophemisms*, but for simplicity we refer to them as neutral.

connotation.

Rababah (2014) studied how medical professionals use x-phemisms when talking to patients and found that serious conditions tend to inspire more euphemism. Rababah argued that translating x-phemisms appropriately is important when providing medical care to non-English speakers. It follows that it is important for machine translation systems to preserve euphemistic language across translations in medical applications. More generally, machine translation systems should be concerned not only with preserving the intended semantics but also preserving the intended discourse pragmatics, which includes translating euphemisms into euphemisms and translating dysphemisms into dysphemisms. When a speaker chooses to use a euphemistic or dysphemistic expression, that choice usually reflects a viewpoint or bias that is a significant property of the discourse. Consequently, it is important for NLP systems to recognize x-phemisms and their polarity, both for applications where views and biases are central (e.g., medicine, argumentation and debate, or stance detection in political discourse) and for comprehensive natural language understanding in general.

3 Overview of Technical Approach

X-phemisms are so pervasive in language that euphemism dictionaries have been published containing manually compiled lists (Bertram, 1998; Holder, 2002; Rawson, 2003). However these dictionaries are far from complete because new x-phemisms are constantly entering language, both for long-standing sensitive topics and new ones. For example, every generation of youth invents new ways of referring to defecation, and political trends can trigger heightened sensitivity to controversial topics (e.g., *"enhanced interrogation"* is a recently introduced euphemism for torture). Euphemistic terms can even become offensive with time and replaced by new euphemisms, a phenomenon known as "the euphemism treadmill." For instance, the phrase *"mentally retarded"* began its life as a euphemism. Now, even *"special needs"* is sometimes viewed as offensive. The goal of our research is to develop methods to automatically curate lists of euphemistic and dysphemistic phrases for a topic from a text corpus, which would enable emerging x-phemisms to be continually discovered.

We tackled this problem by decomposing the task into two steps: (1) identifying near-synonym phrases for a topic, and (2) classifying each phrase as *euphemistic*, *dysphemistic*, or *neutral*. For the first step, we considered using existing thesauri (e.g., WordNet, Roget's thesaurus, Wiktionary, etc.) but their synonym lists were relatively small.[2] Roget's thesaurus was among the best resources, but included only a few dozen entries for most topics. Furthermore, x-phemisms can stretch meaning to soften or harden a sensitive subject, so we wanted to include *near-synonyms* that have a similar (but not identical) meaning. For example, *laid off*, *resigned*, and *downsized* are not strictly synonymous with FIRING, but broadly construed they all refer to job termination.

Ultimately, we decided to use the Basilisk bootstrapping algorithm for weakly supervised semantic lexicon induction (Thelen and Riloff, 2002). Basilisk begins with a small set of seed terms for a desired category and iteratively learns more terms that consistently occur in the same contexts as the seeds. While there are other methods for near-synonym generation (e.g., (Gupta et al., 2015)), we chose Basilisk because it can learn phrases corresponding to syntactic constituents (e.g., NPs and VPs) and can use lexico-syntactic contextual patterns. For the bootstrapping process, we used the English Gigaword corpus because it contains a large and diverse collection of news articles. We focused on three sensitive topics that are common in news and rich in x-phemisms: LYING, STEALING, and FIRING (job termination).

4 Generating Near-Synonym Phrases with Semantic Lexicon Induction

The Basilisk algorithm learns new phrases for a category using a small list of "seed" terms and a text corpus. In an iterative bootstrapping framework, Basilisk extracts contextual patterns surrounding the seed terms, identifies new phrases that *consistently* occur in the same contexts as the seeds, adds the learned phrases to the seed list, and restarts the process. Our categories of interest (LYING, STEALING, FIRING) are actions, so we wanted to learn verb phrases as well as noun phrases (e.g., event nominals). Consequently, we provided Basilisk with two seed lists for each topic, one list of verb phrases (VPs) and one list of noun phrases

[2]We considered using the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) as well, but many of its paraphrases are syntactic variations (e.g., active vs. passive) which are not useful for our purpose, and many entries are noisy as they were automatically generated.

FIRE		LIE		STEAL	
NPs	**VPs**	**NPs**	**VPs**	**NPs**	**VPs**
dismissal	dismiss	exaggeration	deceive	larceny	defraud
downsizing	fire	fabrication	distort	misappropriation	embezzle
firing	force resignation	falsehood	exaggerate	pickpocketing	extort
forced retirement	furlough	fib	fabricate	pilfering	loot
layoff	lay off	lie	falsify	purloining	mug
redundancy	leave company	mendacity	lie	robbery	pilfer
reorganization	oust	misrepresentation	misinform	shoplifting	plunder
sacking	resign	misstatement	mislead	stealing	rob
suspension	sack	prevarication	misrepresent	theft	steal
termination	step down	untruth	misstate	theiving	swindle

Table 2: Seed Phrases per Topic

(NPs). To collect seed terms, we identified common phrases for each topic that had high frequency in the Gigaword corpus. The seed lists are shown in Table 2. We included both active and passive voice verb phrase forms for the verbs shown in Table 2, except we excluded *resign* in passive voice because *"was resigned to"* is a common expression with a different meaning.

Most previous applications of Basilisk have used lexico-syntactic patterns to represent the contexts around seed terms (e.g., (Riloff et al., 2003; Qadir and Riloff, 2012)). For example, a pattern may indicate that a phrase occurs as the syntactic subject or direct object of a specific verb. So we used the dependency relations produced by the SpaCy parser (https://spacy.io/)[3] for contextual patterns. For generality, we used word lemmas both for the learned phrases and the patterns.

4.1 Representing Contextual Patterns and Verb Phrases

We defined a contextual pattern as a dependency relation linked to/from a seed term, coupled with the head of the governing/dependent phrase. For example, consider the sentence *"The lie spread quickly"*. The contextual pattern for the noun *"lie"* would be ←**NSUBJ(spread)**, indicating that the NP with head *"lie"* occurred as the syntactic subject of a governing VP with head *"spread"*. We treated *"have"*, *"do"*, and *"be"* as special cases because of their generality and paired them with the head of their complement (subject, direct object, predicate nominal, or predicate adjective). For example, given the sentence *"The lie was horrific"*, the contextual pattern for *"lie"* would be ←**NSUBJ(be_horrific)**.

We also created compound relations for syntactic constructions that rely on pairs of constituents to be meaningful. For example, a preposition alone is not very informative, so we pair each preposition with the head of its object (e.g., *"in_jail"*). Specifically, we pair the dependency relation "prep" with its "pobj," "agent" with its "pobj", and "dative" with the "dobj" of its governing verb. We also create compound dependencies for "pcomp," and "advcl" relations and resolve the relative pronoun with its subject for "relcl" relations.

Basilisk has not previously been used to learn multi-word verb phrases, so we needed to define a VP representation. We represented each VP using the following syntax: VP([voice]<verb>)MOD(<modifier>)DOBJ(<noun>). The VP() identifies the head verb and voice (Active or Passive), and MOD() contains the first of any adverbs or particles included in the verb phrase. DOBJ() contains the head noun of a VP's direct object, if present. As we did with the contextual patterns, we treat *"have"*, *"do"*, and *"be,"* as special cases and join the verb with its complement. As an example, the verb phrase *"is clearly distorting"* would be represented as **"VP([active]be_distort)MOD(clearly)"**.

We observed that many of the most useful contextual patterns for identifying near-synonyms captured conjunction dependency relations. For example, the contextual pattern ←**CONJ(distortion)** occured with 6 seed terms (*exaggeration*, *fabrication*, *falsehood*, *lie*, *misrepresentation*, and *untruth*), as well as several other near-synonyms such as *disinformation*, *inaccuracy*, *crap*, and *dishonesty*. Near-synonyms also frequently appeared in conjoined verb phrases, such as *"misstate and inflate"* or *"misstate and embellish"*. As an example of a different type of dependency relation that proved to be useful, the compound pattern →**AgentPhrase(by_looter)** occurred with several near-synonym VPs for STEAL, such as *seize*, *ran-*

[3]We used all relations except "punct" and "det".

sack and *clean out*.

4.2 Near-Synonym Generation Results

We had no idea how many near-synonyms we could expect to find for each topic, so we configured Basilisk to learn 1,000 phrases[4] to err on the side of overgeneration. Basilisk learns in a fully automated process, but the resulting lists are not perfect so must be manually filtered to ensure a high-quality lexicon. The filtering process took us approximately 1.5 hours to review each list.[5] Most of the correct entries were among the first 400 terms generated by Basilisk.

Topic	FIRE	LIE	STEAL
# Phrases	142	177	146

Table 3: Near-Synonym Phrases per Topic

Table 3 shows the total number of near-synonyms acquired for each topic, after conflating active and passive voice variants, typos, and including the seed terms. These numbers show that the semantic lexicon induction algorithm enabled us to quickly produce many more near-synonym phrases per topic than we had found in the synonym lists of thesauri. Some of the discovered terms were quite interesting, such as *"infojunk"* and *"puffery"* for LIE, and sometimes unfamiliar to us but relevant, such as *"malversation"* and *"dacoity"* for STEAL.

5 Gold X-Phemism Data Set

To create a high-quality gold data set for x-phemism classification, we asked three people (not the authors) to label the near-synonym phrases for each topic on a scale from 1 to 5, where 1 is most dysphemistic, 3 is neutral, and 5 is most euphemistic. For each phrase, we computed the average score across the three annotators and assigned each phrase to a "gold" x-phemism category: phrases with score < 2.5 were labeled *dysphemistic*, phrases with score > 3.5 were labeled *euphemistic*, and the rest were labeled *neutral*.

To assess inter-annotator agreement, we assigned each annotator's score to one of the three x-phemism categories using the same ranges as above, and measured the category agreement for each pair of annotators using Cohen's kappa (κ). For LYING, the pairwise κ scores were $\{.64, .69, .77\}$ with average $\kappa = .70$. For FIRE, the κ scores were $\{.66, .68, .80\}$ with average $\kappa = .71$. For STEAL, the κ scores were $\{.66, .77, .79\}$ with average $\kappa = .74$. Since the mean κ scores were $\geq .70$ for all three topics, we concluded that the agreement was reasonably good.

Table 4 shows examples of near-synonym phrases[6] with their gold scores and category labels. For example, *crap* and *infojunk* were among the most dysphemistic phrases for LIE, while *invent* and *embellish* were among the most euphemistic phrases for LIE. Table 5 shows the distribution of labels in the gold data set.

FIRE	**LIE**	**STEAL**	**GOLD**	
ax	crap	gut	1.00	D
flush out	infojunk	snatching	1.33	D
oust	fool	thuggery	1.66	D
expel	fakery	mug	2.00	D
disbar	deceive	burgle	2.33	D
severance	falsehood	rob	2.66	N
fire	lie	steal	3.00	N
dismiss	mistruth	despoliation	3.33	N
decommission	fabrication	malversation	3.66	E
downsize	misinform	overcharge	4.00	E
leave company	exaggerate	confiscate	4.33	E
furlough	invent	legerdemain	4.66	E
retire	embellish	–	5.00	E

Table 4: Examples of Gold Data Scores and Labels (D = dysphemistic, N = neutral, E = euphemistic)

	FIRE	**LIE**	**STEAL**
Euphemism	.30	.42	.24
Neutral	.29	.30	.35
Dysphemism	.41	.28	.41

Table 5: Class Distributions in Gold Data

6 X-phemism Classification with Sentiment Lexicons

Euphemisms and dysphemisms capture softer and harsher references to sensitive topics, so one could argue that this phenomenon falls within the realm of sentiment analysis. But x-phemisms are a distinctly different phenomenon. It may be tempting to equate euphemisms with positive sentiment and dysphemisms with negative sentiment, but x-phemisms refer to sensitive topics that typically

[4] Basilisk ran for 200 iterations learning 5 words per cycle.

[5] One of the authors did this filtering. Our goal was merely to obtain a list of near-synonyms to use for x-phemism classification, and not to evaluate the near-synonym generation per se since that is not the main contribution of our work.

[6] We display the phrases here as n-grams for readability, but they are actually represented syntactically. For example, *"leave company"* is represented as an active voice VP with head *"leave"* linked to a direct object with head *"company"*.

have strong affective polarity (usually negative). For example, vomiting is never a pleasant topic, no matter how it is referred to. Consequently, most euphemisms for vomiting still have negative polarity (e.g., *"be sick"* or *"lose your lunch"*). However some euphemisms can have neutral polarity, such as scientific or formal terms (e.g., *"regurgitation"*), and occasionally a euphemism will evoke positive polarity for a negative topic through metaphor (e.g., *"pushing up daisies"* for death). In this section, we investigate whether sentiment information can be beneficial for recognizing euphemisms and dysphemisms and establish baseline results for this task. We explore five properties associated with sentiment: affective polarity, connotation, intensity, arousal, and dominance.

As our first baseline, we assess the effectiveness of using positive/negative affective polarity (valence) information to label x-phemism phrases using two sentiment lexicons: the NRC EmoLex and VAD Lexicons (Mohammad and Turney, 2013; Mohammad, 2018a). For the specific emotions, we considered *anger*, *disgust*, *fear*, *sadness*, and *surprise* to be negative, and *anticipation*, *joy*, and *trust* to be positive. Another sentiment property related to x-phemisms is connotation. Euphemisms often include terms with positive connotation to soften a reference, and dysphemisms may include terms with negative connotation to make a reference more harsh. But importantly, connotation and x-phemisms are not the same phenomenon. For one, many terms with a strong connotation are not x-phemisms. Also, as with polarity, euphemisms can retain a negative connotation because the underlying topic has negative polarity. But since connotation and x-phemisms are related, we investigate whether connotation polarities from ConnotationWN (Feng et al., 2013; Kang et al., 2014) can be valuable for labeling x-phemisms.

We also explored the effectiveness of using affective intensity, arousal, and dominance information from the NRC Affective Intensity and VAD Lexicons (Mohammad, 2018b,a) for recognizing euphemistic and dysphemistic phrases. Dysphemisms are often harsh and can be downright rude, so we hypothesized that terms with high arousal may be dysphemistic. Conversely, euphemisms use softer and gentler language, so they may be associated with low arousal. Dominant terms correspond to power and control, so it would be logical to expect that high dominance may be associated with euphemisms and low dominance may be associated with dysphemisms (e.g., *"frail"* and *"weak"*) (Mohammad, 2018a).

For intensity, we used the NRC Affective Intensity Lexicon (Mohammad, 2018b), which associates words with specific emotions. We mapped the intensity scores so that high intensity values for negative emotions ranged from [0-0.5] (representing dysphemistic to neutral) and high intensity values for positive emotions ranged from [0.5-1] (representing neutral to euphemistic).

The sentiment resources provide scores between 0 and 1. For polarities and connotation, 0 represents the strongest negative score and 1 represents the strongest positive score. For arousal, and dominance, the range is low (0) to high (1). We expect high arousal to be associated with dysphemism, so to be consistent with the other properties we reverse its range and replace each score S with 1-S. We score multi-word phrases by taking the average score of their words. Once a phrase receives a score S, we map S to one of the three x-phemism categories as follows: $S \leq 0.25 \Rightarrow$ *dysphemism*, $0.25 < \text{S} < 0.75 \Rightarrow$ *neutral*, and $\text{S} \geq .75 \Rightarrow$ *euphemism*. We chose these ranges to conservatively divide the space into quadrants, so that scores in the lowest quadrant represent dysphemism, scores in the highest quadrant represent euphemism, and scores in the middle are considered neutral.

6.1 Lexicon Results

Table 6 shows the results for the sentiment lexicon experiments. We report F-scores for the euphemism (**Euph**), neutral (**Neu**), and dysphemism (**Dysph**) categories as well as a macro-average F-score (**Avg**). The best-performing lexicon across all three topics was ConnotationWN (ConnoWN).

We also experimented with combining multiple dictionaries to see if they were complementary. For these experiments, each dictionary labeled a phrase as euphemistic, dysphemistic, or neutral (as described earlier) or none (i.e., no label if the word was not present in the lexicon). The most frequent label was then assigned to the phrase, except that 'none' labels were ignored. ConnotationWN's label was used to break ties. We evaluated all pairs of lexicons and the best pair turned out to be ConnotationWN plus Valence, which we refer to as BestPair in Table 6. We also tried using all of the dictionaries, shown as AllDicts in Table 6. Combining dictionaries did improve performance, with

BestPair performing best for FIRE and STEAL, and AllDicts performing best for LIE.

Overall, connotation and valence (affective polarity) were the most useful sentiment properties for recognizing x-phemisms. But, thus far we have considered only the words in a phrase. In the next section, we explore an approach that exploits the sentence contexts around the phrases.

FIRE	**Euph**	**Neu**	**Dysph**	**Avg**
EmoLex	.00	.00	.28	.09
Dominance	.00	**.40**	.07	.15
Intensity	.05	.29	.18	.17
Arousal	.05	.35	.15	.18
Valence	.05	.26	.25	.19
ConnoWN	.28	**.40**	**.50**	.39
AllDicts	.39	.30	.48	.39
BestPair	**.40**	.33	.47	**.40**
LIE	**Euph**	**Neu**	**Dysph**	**Avg**
EmoLex	.11	.04	.37	.17
Intensity	.03	.38	.15	.19
Dominance	.09	**.42**	.15	.22
Arousal	.03	**.42**	.25	.23
Valence	.12	.31	.37	.26
ConnoWN	.31	.32	.38	.34
BestPair	**.33**	.33	.38	.35
AllDicts	.32	.40	**.43**	**.39**
STEAL	**Euph**	**Neu**	**Dysph**	**Avg**
EmoLex	.20	.00	.21	.13
Intensity	.00	.19	.18	.13
Arousal	.00	.30	.21	.17
Valence	.20	.20	.23	.21
Dominance	.21	**.39**	.07	.22
ConnoWN	.20	.21	**.43**	.28
AllDicts	.28	.31	.41	.33
BestPair	**.40**	.33	.41	**.38**

Table 6: Results for Sentiment Lexicons (F-scores)

7 X-phemism Classification with Contextual Sentiment Analysis

We hypothesized that the contexts around euphemisms and dysphemisms would be different in terms of sentiment. People often use euphemisms when they want to be comforting, supportive, or put a positive spin on a subject. In obituaries, for example, euphemisms for death are often accompanied by references to peace, heaven, flowers, and courage. In contrast, grisly murder mystery novels often use dysphemisms, speaking about death using harsh or graphic language. X-phemisms are also prevalent in political discourse. People frequently use euphemisms to argue for the merits of a particular subject (e.g., *"enhanced interrogation"* is a euphemism invoked to justify the use of TORTURE). Conversely, people use dysphemisms when arguing against something (e.g., *"baby killing"* to refer to ABORTION).

To investigate this hypothesis, we developed models to classify a phrase with respect to x-phemism categories using sentiment analysis of its sentence contexts. We use the Gigaword corpus and experiment with both sentiment lexicons and a sentiment classifier to evaluate sentence polarity.

However polysemy and metaphor pose a major challenge: many phrases have multiple meanings. To address this problem, we create a subcorpus for each topic by extracting Gigaword articles that contain a seed term for that topic (see Table 2). The seed terms can also be ambiguous, but we expect that the resulting subcorpus will have a higher density of articles about the intended topic than the Gigaword corpus as a whole. Given a candidate x-phemism phrase for a topic, we then extract sentences containing that phrase from the topic's subcorpus. Our expectation is that most documents that contain both the x-phemism phrase and a seed term for the topic will be relevant to the topic.

Once we have a set of sentence contexts for an x-phemism phrase, our first contextual model uses sentiment lexicons to determine each sentence's polarity. For each topic, we use the best-performing lexicons reported in Section 6.1 (i.e., BestPair for FIRE and STEAL, and AllDicts for LIE). First, each word found in the lexicons is labeled positive for scores > 0.5 or negative for scores < 0.5.[7] We then assign a polarity to each sentence based on majority vote among its labeled words. Sentences with an equal number of positive and negative words, or no labeled words, are ignored.

X-phemisms are relative to a topic that itself often has strong affective polarity, so given a phrase P, our goal is to determine whether P's contexts are positive or negative *relative to the topic*. To assess this, we generate a polarity distribution across *all* sentences in the topic's subcorpus. We will refer to all sentences in the subcorpus for topic T as Sents(T) and the sentences in the subcorpus that mention phrase P as Sents(T, P). We define POS(S) as the percent of sentences S labeled positive, and

[7] If a word occurred in multiple lexicons, ConnotationWN was given precedence.

NEG(S) as the percent of sentences S labeled negative, and classify each phrase P as follows:

If POS(Sents(T, P)) > POS(Sents(T)) + γ
Then label P as *euphemistic*

If NEG(Sents(T, P)) > NEG(Sents(T)) + γ
Then label P as *dysphemistic*

Else label P as *neutral*

We set γ = 0.10 for our experiments.[8] Intuitively, the γ parameter dictates that a phrase is labeled as euphemistic (or dysphemistic) only if its sentence contexts have a positive (or negative) percentage at least 10% higher than the sentence contexts for the topic as a whole.

Our second contextual model uses a sentiment classifier instead of lexicons to assign polarity to each sentence. We used a reimplementation of the NRC-Canada sentiment classifier (Mohammad et al., 2013), which performed well in the SemEval 2013 Task 2. Given a sentence, the classifier returns probabilities that the sentence is positive, negative, or neutral. We label each sentence with the polarity that has the highest probability.

Since the classifier provides labels for all three polarities (whereas we only got positive and negative polarities from the lexicons), we use a slightly different procedure to label a phrase. First, we compute the percent of subcorpus sentences containing phrase P that are assigned each polarity (POS, NEG, NEU), and compute the percent of all subcorpus sentences assigned each polarity. Then we compute the difference for each polarity. For example, Δ(POS) = POS(Sents(T, P))-POS(Sents(T)). This represents the difference between the percent of Positive sentences containing P and the percent of Positive sentences in the subcorpus as a whole. Finally, we label phrase P based on the polarity that had the largest difference: POS $\Rightarrow$ *euphemistic*, NEG $\Rightarrow$ *dysphemistic*, NEU $\Rightarrow$ *neutral*.

7.1 Contextual Sentiment Results

Table 7 shows F-score results for the contextual models on our gold data. We evaluated three contextual models that use different mechanisms to label the affective polarity of a sentence: ContextNRC uses the NRC sentiment classifier, ContextAllDicts uses the AllDicts lexicon method, and ContextBestPair uses the BestPair lexicon method. For the sake of comparison, we also re-display the results for the best lexicon model (*BestDictModel*) presented in Section 6.1 for each topic.

[8] We chose $\gamma = .10$ based on intuition without experimentation, so a different value could perform better.

	Euph	Neu	Dysph	Avg
FIRE				
BestDictModel	.40	**.33**	**.47**	**.40**
ContextNRC	.28	.18	.37	.28
ContextAllDict	**.52**	.19	.18	.30
ContextBestPair	.31	.26	.45	.34
LIE				
BestDictModel	.32	.40	.43	.39
ContextBestPair	.42	.41	.35	.39
ContextNRC	.56	.19	**.46**	.40
ContextAllDicts	**.67**	**.42**	.31	**.47**
STEAL				
BestDictModel	.40	**.33**	.41	.38
ContextNRC	.24	.25	**.52**	.34
ContextBestPair	.42	.24	.47	.38
ContextAllDicts	**.61**	.29	.40	**.43**

Table 7: Results for Contextual Analysis (F-scores)

For LIE and STEAL, the best contextual model outperformed the best lexicon method, improving the F-score from .39 $\rightarrow$ .47 for LIE and from .38 $\rightarrow$ .43 for STEAL. For FIRE, the contextual models showed lower performance. We observed that phrases for the FIRE topic exhibited more lexical ambiguity than the other topics, so the subcorpus extracted for FIRE was more noisy than for the other topics. This likely contributed to the inferior performance of the contextual models on this topic.

Table 8 shows the recall (R) and precision (P) breakdown for the best performing model for each topic. Euphemisms had the best recall and precision for LIE and STEAL, but lower recall for FIRE. Precision was lowest for the neutral category overall, indicating that too many euphemistic and dysphemistic phrases are being labeled as neutral.

	Euph		Neu		Dysph	
	R	P	R	P	R	P
FIRE	.31	.58	.47	.25	.44	.51
LIE	.64	.69	.52	.35	.24	.46
STEAL	.68	.56	.32	.26	.33	.50

Table 8: Recall and Precision of Best Models

Our observation is that the models perform best on strongly euphemistic or dysphemistic phrases, and they have the most trouble categorizing metaphorical expressions, such as *"ax"* for FIRE. It makes sense that the lexicon-based models would have difficulty with these cases, but we had hoped that the contextual models would fare better. We suspect that polysemy is especially problematic for metaphorical phrases, resulting in a subcorpus

for the topic that contains many irrelevant contexts. Incorporating understanding of metaphor seems to be an important direction for future research.

8 Conclusions

This paper presented the first effort to recognize euphemisms and dysphemisms using natural language processing. Our research examined the relationship between x-phemisms and sentiment analysis, exploring whether information about affective polarity, connotation, arousal, intensity, and dominance could be beneficial for this task. We used semantic lexicon induction to generate near-synonyms for three topics, and developed lexicon-based and context-based sentiment analysis methods to classify phrases as *euphemistic*, *dysphemistic*, or *neutral*. We found that affective polarity and connotation information were useful for this task, and that identifying sentiment in sentence contexts around a phrase was generally more effective than labeling the phrases themselves. Promising avenues for future work include incorporating methods for recognizing politeness, formality, and metaphor. Euphemisms and dysphemisms are an exceedingly rich linguistic phenomenon, and we hope that our research will encourage more work on this interesting yet challenging problem.

Acknowledgments

We gratefully thank Shelley Felt, Shauna Felt, and Claire Moore for their help annotating the gold data for this research.

References

Keith Allan. 2009. The connotations of English colour terms: Colour-based X-phemisms. *Journal of Pragmatics*, 41.

Keith Allan and Kate Burridge. 1991. *Euphemism and Dysphemism: Language Used as Shield and Weapon*. Oxford University Press, New York.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Anne Bertram. 1998. *NTC's Dictionary of Euphemisms*. NTC, Chicago.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*.

Juri Ganitkevitch, , Benjamin Van Durme, and Chris Callison-burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

D. Gupta, J. Carbonell, A. Gershman, S. Klein, and D. Miller. 2015. Unsupervised phrasal near-synonym generation from text corpora. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

R. W. Holder. 2002. *How Not To Say What You Mean: A Dictionary of Euphemisms*. Oxford University Press, Oxford.

Jun Seok Kang, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning connotation over the word+sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu (Margento). 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.

Rijul Magu and Jiebo Luo. 2014. Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*.

Saif M. Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Kerry L. Pfaff, Raymond W. Gibbs Jr., and Michael D. Johnson. 1997. Metaphor in using and understanding euphemism and dysphemism. *Applied Psycholinguistics*, 18.

A. Qadir and E. Riloff. 2012. Ensemble-based semantic lexicon induction for semantic tagging. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*.

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.

Hugh Rawson. 2003. *Rawson's Dictionary of Euphemisms and Other Doubletalk*. Castle, Chicago, IL.

E. Riloff, J. Wiebe, and T. Wilson. 2003. Learning Subjective Nouns using Extraction Pattern Bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pages 25–32.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

S. Somasundaran and J. Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

M. Thelen and E. Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pa ttern Contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 214–221.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

M. Walker, P. Anand, R. Abbott, and R. Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Alan Wallington, Rodrigo Agerri, John Barnden, Mark Lee, and Tim Rumbell. 2011. Affect transfer by metaphor for an intelligent conversational agent. In *Affective Computing and Sentiment Analysis*, pages 53–66. Springer.

Michael Wiegand, Josef Ruppenhoffer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words - a feature-based approach. In *NAACL-HLT 2018*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics*.