Why are reckless socks not (more of) a thing? Towards an empirical classification of evaluative concepts.

Lucien Baumgartner (lucien.baumgartner@philos.uzh.ch)
Department of Philosophy, Zürichbergstrasse 43
8044 Zurich, Switzerland

Abstract

This paper proposes new empirical classifiers for evaluative concepts, including thin concepts like *good* or *bad* and thick concepts such as *honest* or *disgusting*, based on quantitative corpus linguistics. Prior work in experimental philosophy has shown that sentiment analysis can be used to track differences between concept classes. Building on this, Task 1 investigates whether the relationship between sentiment and evaluativeness is parabolic rather than linear. Task 2 extends this question to the differences between evaluative and non-evaluative concept classes. The results of both Tasks show that the linear and the parabolic logistic regression classifiers perform equally well. Interestingly, this study also finds that adjectives attributed to animate entities (e.g. "*generous* customer") generally have a higher probability to be evaluative concepts than those attributed to inanimate entities (e.g. "*dry* soil").

Keywords: experimental philosophy; evaluative language; thick concepts; thin concepts; corpus linguistics

Introduction

Over the last decades, there has been an ongoing debate on whether there are principled differences between thin and thick concepts (Eklund, 2011; Kirchin, 2019; Roberts, 2013; Tappolet, 2004; Väyrynen, 2013). Thin concepts such as *good* or *bad* are commonly understood to evaluate, i.e. to express approval or disapproval. Thick concepts like *honest* or *cruel*, on the other hand, evaluate as well but have descriptive content beyond that. If we say "Amy is honest", for example, we positively evaluate Amy and, at the same time, characterize her as somebody who is truthful and genuine. In comparison, "Amy is good" evaluates Amy as a person, but there is no information about why the speaker approves of her.

In addition to these two classes of evaluative concepts, a third one needs to be added. Value-associated concepts like *homeless* or *sunny* can be used to describe but also, in a way, carry an evaluation. However, the relationship between the descriptive and the evaluative is importantly different compared to thick and thin concepts. It has been hypothesized by Reuter, Baumgartner, and Willemsen (ms) that adjectives like *homeless* or *sunny* do not evaluate in the sense of expressing approval or disapproval for having certain descriptive features. For example, being homeless is considered undesirable, but we do not necessarily evaluate a person negatively for living on the streets. Hence, unlike thin and thick concepts, value-associated concepts are merely *associated* with certain values (Reuter et al., ms).

The few experimental-philosophical studies that exist on evaluative concept classes, especially on thick and thin concepts, are mostly based on online surveys (Willemsen & Reuter, 2020, 2021). Recently, however, Willemsen, Baumgartner, Frohofer, and Reuter (2021) have conducted a corpus study comparing the use of evaluative adjectives by legal professionals and laypeople. The authors show that sentiment values, which are commonly used in opinion mining, track differences among descriptive concepts and thick concepts from different domains (thick epistemic, thick ethical, and thick legal concepts).

In this paper, I provide evidence on how sentiment values can be used to *classify* evaluative and non-evaluative concepts, rather than just *measure* differences between them. Specifically, I look at the concepts featured in Reuter et al. (ms), i.e. descriptive, thin, thick moral, thick non-moral, and value-associated concepts. For the classification tasks, I adapt the experimental setup of Willemsen et al. (2021). In Task 1, I build linear and polynomial classifiers for evaluative (thin, thick moral, thick non-moral) vs. non-evaluative (descriptive, value-associated) concepts. Based on these models, I test the hypothesis that the relationship between sentiment and evaluativeness is parabolic rather than linear. Task 2 extends this hypothesis to the *pairwise* differences between evaluative and non-evaluative concept classes (e.g. descriptive vs. thin concepts). However, the results of Task 1 & 2 show that the linear and the parabolic classifier perform equally well.

Moreover, this study also takes into account contextual differences in regards to *what* or *whom* the adjectives are actually attributed to (or predicated of). I show that propositions like "I wear reckless socks.", namely where a thick term is attributed to an inanimate entity, are less frequent than corresponding thick attributions to animates, e.g. "My cat is reckless." Accordingly, animacy should be considered as a predictor (among others) for whether a term is evaluative or not. This research provides a push in the direction of more context-sensitive modeling of evaluativeness, although still being concerned with the standard content of concepts.

This project represents a first step away from purely intuition-based conceptual work and towards machine-based applications of philosophical frameworks of evaluative language. In the long run, machine-based classification will allow experimental philosophy to empirically assess the accuracy of theoretical predictions based on natural language data.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Rationale

Willemsen et al. (2021) focus on the sentiment dispersion in coordinating conjunctions containing evaluative adjectives, as in "The Eiffel Tower is *beautiful and imposing*." Adjectives in *and*-conjunctions typically have a similar sentiment polarity and intensity (Elhadad & McKeown, 1990; Hatzivassiloglou & McKeown, 1997). Put simply, positively evaluating adjectives are commonly used in conjunction with other positive adjectives, descriptive ones are paired with neutral ones, and negative with negative ones.

Based on these considerations, Willemsen et al. (2021) preselect a list of adjectives and hand-code their respective concept class. Willemsen et al. (2021) investigate different kinds of thick concepts (epistemic, ethical, legal) as well as descriptive concepts. For this study, however, I am interested in the broader class of evaluative concepts and not just thick concepts. Hence, I use the pre-coded list of concepts featured in Reuter et al. (ms) instead, which contains descriptive, thick moral, thick non-moral, thin, and value-associated concepts. Each of the examined conjunctions consists of one of these pre-selected adjectives, the so-called 'target adjective', and a freely variable 'conjoined adjective'. The conjoined adjective gets annotated with sentiment values ranging from -1 (extremely negative), over the neutral midpoint (i.e. 0), to 1 (extremely positive).

The rationale behind this standardization is that for each target adjective, e.g. *honest*, we know both *honest*'s concept class (as theoretically defined) as well as its sentiment dispersion across all the conjunctions it is part of (by virtue of the conjoined adjective). In a classification task, it should thus be possible to use the sentiment distribution of *honest*'s conjoined adjectives as a predictor of *honest*'s evaluativeness.

The Linear and the Parabolic Relationship Assumptions

Willemsen et al. (2021) use simple *linear* methods, such as ANOVA and estimated marginal means (EMMs), to measure sentiment differences between the aforementioned concept classes. Why? Think about it this way: if X is conjoined with *horrendous* or *marvelous*—both very negative adjectives—, the term is more likely to be evaluative than, say, if it is conjoined with *administrative*—which is much more neutral. In other words: the probability for a term X to be evaluative (vs. non-evaluative) decreases, the closer it is to the neutral midpoint of the sentiment scale; inversely, it increases, the bigger its distance to the neutral midpoint. In Willemsen et al. (2021), this relationship is assumed to be constant, or, linear, as illustrated by the blue lines in Figure 1. At first glance, this sounds rather intuitive.

The linear hypothesis, however, has a potential problem: why would one assume that the strength of this relationship is constant along the sentiment spectrum? Proponents of a non-linear relationship could object that a sentiment change around the neutral midpoint of the scale has less of an effect on the probability for X to be evaluative, compared to a corresponding sentiment change at the extremes of the scale. For instance, the probability for *dry* to be evaluative should be similarly low across "dry and dusty soil" and "dry and rocky soil". On the other hand, there seems to be a more important difference for *generous* in "generous and friendly customer" and "impeccable and generous customer". In other words, one could argue that the linear assumption underestimates the effects of sentiment changes at the extremes of the scale and overestimates the ones around the neutral midpoint. According to this, the relationship follows a parabola, i.e. a polynomial function, as depicted by the red line in Figure 1. This study will test whether the parabolic relationship assumption leads to a better classification of evaluative concepts compared to the linear hypothesis.

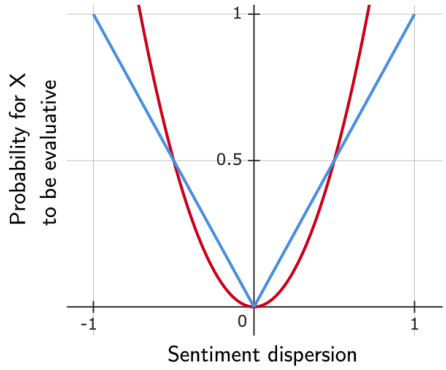

Figure 1: Theoretical linear (blue) and parabolic (red) trends for the classification based on sentiment.

Animacy as a Predictor

Not every adjective can be plausibly attributed to all kinds of entities. The semantic content of adjectives often restricts their range of application. Consider, for example:

(1) Sarah collects beach pebbles. She thinks they are smooth and colorful.

(2) Megan loves Justin Bieber. She thinks he is honest and sensitive.

The adjectives in (1) describe an inanimate entity (i.e. beach pebbles), whereas in (2) they are used to characterize a person, the pop star Justin Bieber. It is hard to see how thick concepts such as *honest* and *sensitive* could be attributed to beach pebbles, even more so on a regular basis. We can, of course, come up with creative uses, as in "John, a stone is more sensitive than you!", but they are arguably less frequent than literal instances. Another example is *dead*: we might occasionally hear that "the party this weekend was dead", but more often than not it is used to say that a formerly alive being—a person, animal, plant, etc.—died (cf. Dahl, 2008). Thus *to what* or *to whom* the adjectives are attributed matters. Animate and inanimate noun phrases ("shrimp" vs. "wooden shrimp") and verbs ("to think" vs. "to crystallize") often come with grammatical and ontological restrictions (e.g., Dahl, 2008; De

Swart & De Hoop, 2018; García, Primus, & Himmelmann, 2018; Schumacher, 2018). We expect this also to be reflected in the frequency of use of evaluative adjectives. Especially thick *moral* concepts seem to be predestined to be applied to animates, rather than concrete objects ("selfish windowsill") or abstract entities ("compassionate prime number"). This raises the question of whether there is a more general relation between evaluativeness and animacy. To allow for a more detailed analysis, I thus suggest including the animacy of the attributed entity as a predictor.

Data

The data for this study consists of 22,500 Reddit comments, which were initially collected using the Pushift API (Baumgartner et al., 2020). Each comment contains a coordinating conjunction of two adjectives. For simplicity's sake, only *and*-conjunctions are considered, as conjunctions with *but*, *or*, or *yet* work differently (Elhadad & McKeown, 1990; Hatzivassiloglou & McKeown, 1997). Comments which include a negation of the adjectives (e.g. *not*, *hardly*, *barely*) or any other adverbial modifier (e.g. *very*, *rather*, *mostly*) were discarded.

The target adjectives consist of a pre-selection of 45 terms featured in Reuter et al. (ms). The selected adjectives have already been annotated with their respective concept class. Analogous to Reuter et al. (ms), I distinguish between descriptive (e.g. *yellow*, *dry*), thick moral (e.g. *compassionate*, *cruel*), thick non-moral (e.g. *delicious*, *disgusting*), thin (e.g. *good*, *bad*), and value-associated concepts (e.g. *quiet*, *homeless*). Each conjunction in the data contains one of the preselected target adjectives and freely variable conjoined adjective.[1] The conjoined adjectives were annotated with sentiment values from the SentiWords dictionary (Esuli & Sebastiani, 2006; Gatti, Guerini, & Turchi, 2016; Baccianella, Esuli, & Sebastiani, 2010).[2] The dictionary codes a term's sentiment intensity on a scale from $-1 \leqslant x \leqslant 1$.

The comments typically span over multiple sentences and make heavy use of coreferences (anaphora and cataphora). Since I am only interested in the sentences containing the aforementioned adjective conjunctions (rather than the complete comment respectively), coreferences can ultimately lead to a loss of semantic information, if left unresolved. Hence, I applied a coreference resolution algorithm by Zeldes and Zhang (2016).[3] The same algorithm also detects the animacy state (*animate* or *inanimate*) and the entity type (e.g., *abstract*, *person*, *object*, etc.) of named and non-named entities mentioned in the texts.[4] Together with the corresponding dependency trees, this information is used to determine whether the adjective conjunction is attributed to (or predicated of) an animate or inanimate entity.[5] Due to malformed sentences, not all comments could be annotated. Since I do not intend to perform the classification task on an adjectival level, the variations in sample size are negligible. After the annotation step, the corpus retains 18,301 sentences that contain the desired target structures.[6]

Methods

Task 1

Task 1 is to perform a classification of evaluative versus non-evaluative concepts based on logistic regression, comparing the linear and the parabolic assumption. Sequential likelihood-ratio tests and backward stepwise regression are used to determine the best polynomial predictor for the polynomial model (see below). The data is split randomly into a test and training set, based on a 20-80% ratio. The training includes 10-fold cross-validation to select the best linear and polynomial models respectively.[7]

Task 2

In Task 2, the results are disentangled a bit further. This is motivated by the fact that the non-evaluative class is composed of two very different sub-classes, namely descriptive and value-associated concepts. As shown in Figure 2, descriptive and value-associated concepts have very different sentiment distributions. This indicates that pooling descriptive and value-associated together as non-evaluative class might be problematic. Hence, classification accuracy might improve for binary classifications of concept classes, say, descriptive vs. thin concepts.

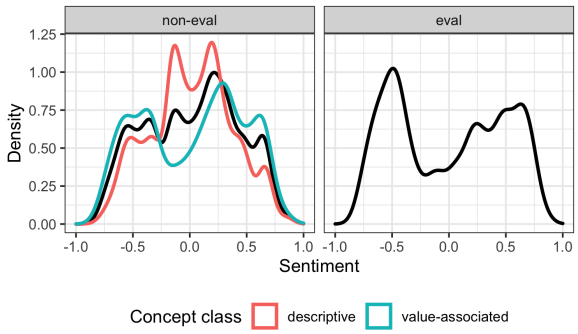

Figure 2: Sentiment distributions. The black lines show the pooled dispersion.

Hence, Task 2 is to examine how well the linear and polynomial classifiers work for 2×3 distinct concept class pairs,

[1] To guarantee a balanced sample, we initially collected 500 comments for each adjective ($45 \times 500 = 22,500$).

[2] For the sentiment annotation I used the `quanteda`-package (v3.0.0) in R (v4.1.0).

[3] The classifier's performance is reviewed in Sukthanker, Poria, Cambria, and Thirunavukarasu (2020).

[4] Both the coreference resolution and the animacy detection are conducted with `xrenner` (v2.2.0.0) by Zeldes and Zhang (2016), based on the the pretrained Electra model for GUM7, using Python (v3.7.11).

[5] The dependency parsing was conducted using the `stanza` toolkit (v1.3.0) provided by the Stanford NLP Group (Qi, Zhang, Zhang, Bolton, & Manning, 2020) in Python (v3.7.11).

[6] The data, analyses, and the full selection of target adjectives, are available on the OSF repository at `https://osf.io/s5n6g/`.

[7] The models were built with `caret` (v6.0-90) in R (v4.1.0).

each featuring one evaluative (thin, thick moral, thick non-moral) and one non-evaluative (descriptive, value-associated) concept class.

Model Specification

The linear model in Task 1 is pretty straightforward: the dependent variable is a dummy for non-evaluative (descriptive and value-associated) terms and evaluative (thin, thick moral, and thick non-moral) terms. As predictors, I use the interaction of absolute sentiment values (continuous) and animacy (two-level factor),[8] and an interaction between the absolute sentiment values and a dummy coding for the polarity of the sentiment values ($x < 0$: negative; $x \geqslant 0$: positive). The reason for using absolute sentiment values and a polarity dummy is that we want to be able to discriminate between different polarities without having to average the effects over the whole sentiment continuum.

The parabolic or polynomial model in Task 1 has the same dependent variable as the linear model. As predictors, I use the sentiment values (continuous) to the power *i* and animacy (two-level factor), as well as their interaction. Note that with this model we do not need to limit ourselves to absolute sentiment values since we use a polynomial function. To determine the best exponent, I ran sequential likelihood-ratio tests for different *i*s. The root model simply classifies whether a target adjective is evaluative or not, based on the conjunct sentiment values and the animacy of the object of predication.[9] The subsequent models just iteratively include more polynomials, as in

$$y \sim (I(sentiment)^2 + ... + I(sentiment)^n) * animacy$$

Table 1 shows likelihood-ratio tests for the root model (Model 1) and the polynomial models from degree 2 to 6 (Model 2-6):

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Model 1	18297	24690.28			
Model 2	18295	23854.98	2	835.30	0.0000
Model 3	18293	23841.11	2	13.87	0.0010
Model 4	18291	23762.45	2	78.66	0.0000
Model 5	18289	23749.17	2	13.28	0.0013
Model 6	18287	23746.97	2	2.19	0.3340

Table 1: Sequential likelihood-ratio tests on nested logistic regression models.

The likelihood-ratio tests compare the goodness-of-fit of Models 1-6 sequentially, i.e. Model 1 is compared with Model 2, Model 2 with Model 3, etc. The comparison shows that up to Model 5, the null hypothesis, i.e. that the nested model is better than the complex model, can be rejected on a 0.05-alpha level. This means that simply adding higher degree polynomial terms steadily improves the model up to $i = 5$. Consequently, Model 5 is selected.

A backward stepwise regression further validates that all predictors are significant on 0.05-alpha level, and thus informative, except for the untransformed sentiment values (p-value = 0.1931). However, the latter has to be included still, since its polynomial transformations are also used. Thus, Model 5 can be used without dropping a predictor. This also validates the inclusion of animacy as a predictor.

Task 2 uses the same predictors as the linear model and the best polynomial model in Task 1. However, instead of the pooled dummy (i.e. non-evaluative vs. evaluative) in Task 1, the independent variables in Task 2 are 2×3 concept class dummies (e.g. descriptive vs. thin, descriptive vs. thick moral, etc.).

Results

Task 1

The accuracy of the best polynomial model (fraction of predictions our model got right) is 62.47% (95% CI: 60.88%, 64.04%; p-value < 0.001), which is significantly higher than the no-information rate (56.79%). This means the model is significantly more accurate than just picking the most prevalent observed class. The model's F1 value is 0.5226, based on a Recall of 0.4753 and a Precision of 0.5802.

The linear model, on the other hand, has an accuracy of 61.24 % (CI: 59.64%, 62.82%, p-value < 0.001), with F1 = 0.4785, based on a Recall of 0.4115 and a Precision of 0.5716. Evidently, the difference between the models' accuracy is not significant, as the confidence intervals overlap. Therefore, the linear and the polynomial model are roughly equivalent.

		True Class	
		non-eval	eval
Pred.	non-eval	0.4753	0.2617
	eval	0.5247	0.7383

Table 2: Confusion matrix for the polynomial model.

Table 2 shows the confusion matrix for the polynomial model. 47.53% non-evaluative and 73.83% evaluative concepts were correctly predicted, which indicates that non-evaluative concepts are less homogeneous than evaluative concepts. Table 3 shows the same for the linear model. Apparently, the linear model is slightly worse at correctly classifying non-evaluative concepts correctly (41.15%) than the polynomial model, and slightly better in the case of evaluative concepts (76.53%).

		True Class	
		non-eval	eval
Pred.	non-eval	0.4115	0.2347
	eval	0.5885	0.7653

Table 3: Confusion matrix for the linear model.

[8] For the animacy variable, I coded persons, plants, and animals as animate entities, the rest as inanimate ones.

[9] Note that the root model for the likelihood-ratio tests is *not* the linear model specified above. Hence, the increasing goodness-of-fit discussed below does not license the inference that the polynomial models are better than the linear model.

Now, how does animacy affect the predicted probability for a term to be evaluative (vs. non-evaluative)? Figure 3 shows the predicted probabilities for a term to be evaluative rather than non-evaluative, based on the interaction of the sentiment values (x-axis) and the animacy of the object of predication (color), for the polynomial model.

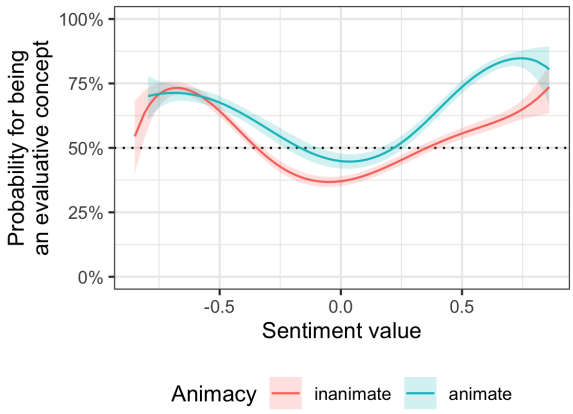

Figure 3: Predicted probabilities for membership in evaluative concept class compared to the non-evaluative class (polynomial model).

There are a few interesting takeaways: Firstly, adjective attributions involving an *animate* entity have a higher probability to be evaluative overall. Secondly, the predicted probability for a term to be evaluative drops significantly below 50% around the midpoint. This means that sentiment-neutral terms have a higher probability to be non-evaluative regardless of animacy, even though there is still a significant difference between animate and inanimate states. Thirdly, the more extreme the sentiment values, the higher the probability for a term to be evaluative. These three findings follow the expectations stated earlier. Yet, one can also see that extremely negative terms ($x < -0.5$) are *not* significantly associated with any animacy state, whereas extremely positive ones ($x > 0.5$) are. This indicates a strong relationship between animacy and sentiment for positive terms, whereas it is more negligible for negative terms. In short, the animacy predictor shows interesting effects, especially the unexpected difference between negative and positive terms.

In summary, there is no significant difference in accuracy between the linear and the polynomial classifier, which makes them equally viable for classification tasks. The main source of false classifications is the sentiment heterogeneity in the class of non-evaluative concepts, i.e. between descriptive and value-associated concepts. The effect for animacy is not constant along the sentiment spectrum, indicating differences between polarities.

Task 2

Table 4 and Table 5 show the accuracy of the polynomial classifiers and the linear models respectively, as well as the

	Accuracy	Lower	Upper	NIR
D : Thin	0.6785	0.6517	0.7044	0.6076
D : Thick M	0.6755	0.6515	0.6988	0.5061
D : Thick NM	0.6546	0.6303	0.6783	0.5310
VAC : Thin	0.6355	0.6089	0.6616	0.6181
VAC : Thick M	0.5943	0.5700	0.6183	0.5099
VAC : Thick NM	0.5817	0.5573	0.6059	0.5009

Note: D: Descriptive; Thick M: Thick Moral; Thick NM: Thick Non-Moral; VAC: Value-Associated Concepts.

Table 4: Model evaluation metrics for pairwise logistic regression models using polynomials.

no-information rate (NIR). All the classifiers (rows) are significantly more accurate than the prediction based on the NIR (on 0.05-alpha level). The models *within* each Table are based on different subsets of data, and hence not directly comparable among themselves. Nonetheless, it seems that the classifiers are generally more accurate in comparing evaluative concept classes to descriptive concepts than they are to value-associated concepts. It is possible to compare respective linear and the polynomial models, though: The polynomial (Table 4) and the linear classifier (Table 5) do *not* perform significantly differently. The only exception is the classification of value-associated vs. thick non-moral concepts, where the linear model (55.15% [52.69%, 57.59%] accuracy) performs significantly worse than the polynomial model (58.17% [55.73%, 60.59%] accuracy).

	Accuracy	Lower	Upper	NIR
D : Thin	0.6769	0.6501	0.7029	0.6076
D : Thick M	0.6742	0.6502	0.6975	0.5061
D : Thick NM	0.6423	0.6178	0.6662	0.5310
VAC : Thin	0.6181	0.5912	0.6444	0.6181
VAC : Thick M	0.6017	0.5774	0.6257	0.5099
VAC : Thick NM	0.5515	0.5269	0.5759	0.5009

Note: D: Descriptive; Thick M: Thick Moral; Thick NM: Thick Non-Moral; VAC: Value-Associated Concepts.

Table 5: Model evaluation metrics for pairwise logistic regression models without polynomials.

In sum, the results for Task 2 indicate that descriptive concepts are more distinct from evaluative concepts (thin, thick moral, and thick non-moral) than value-associated are from the latter. The linear and polynomial models do not perform significantly differently.

Discussion

This study demonstrates that the sentiment dispersion in *and*-conjunctions can be used to distinguish evaluative from non-evaluative terms. In Task 1, I compared the predictive accuracy of classifiers based on the linear and the parabolic assumption. There was no significant difference in accuracy between the two classification models. In Task 2, the classification was further specified to discriminate between pairs of

evaluative concept classes (i.e. thin, thick moral, thick non-moral) and non-evaluative concept classes (i.e. descriptive, value-associated). The accuracy for these models ranged between 55.15%–67.85% (all significantly above the respective non-information rate). The joint evidence from Tasks 1 & 2 suggests that the parabolic relationship assumption does not yield significantly more accurate results than the linear assumption. The results also show that sentiment values are a relevant metric for classifying evaluative concept classes, but also indicate they do not tell the whole story.

It was furthermore found that adjective attributions to animate entities result in a higher probability for said adjective to be an evaluative concept. However, the effect does not seem to be consistent, but rather displays a certain asymmetry. On the negative end of the spectrum, we find that the more negative the conjoined sentiment values for a term are, the smaller the predicted effect of animacy. The same does *not* hold on the positive end of the sentiment spectrum, though. Rather, positive evaluations of animates such as "My gum tree is beautiful and strong." are much more likely than those of inanimates, e.g. "The Golden Gate Bridge is beautiful and strong."

One explanation for this asymmetry might be that the use of negative evaluatives comes with considerable social costs and, accordingly, commits the speaker to the evaluation, while positive evaluations can be used more inflationary (Willemsen & Reuter, 2020, 2021; Willemsen, Baumgartner, Cepollaro, & Reuter, ms). For example, we often use terms such as "friendly" or "honest" to characterize a person who is *merely decent*, whereas "cruel" or "ugly" typically mean more than *simply subpar*. Willemsen and Reuter (2020, 2021), Willemsen et al. (ms), and Baumgartner, Reuter, and Willemsen (ms) found that the evaluation of positive thin and thick terms is significantly easier to cancel than that of negative ones. Participants in their studies judge statements like "Amy is generous, but by that, I don't mean to say anything positive about Amy." to be significantly less contradictory than their negative counterparts, e.g. "Amy is selfish, but by that, I don't mean to say anything negative about Amy." This difference in the treatment of positive and negative evaluations has been dubbed the *Polarity Effect* (e.g., Willemsen & Reuter, 2020, 2021). The authors suggest that positive terms have two standard usage modes, a positively evaluating one and a more neutral one. Negative terms, on the other hand, generally carry a negative evaluation.

The tentative evidence for the effect of animacy points in the same direction. However, Willemsen and Reuter (2021) and Baumgartner et al. (ms) tested whether the Polarity Effect is affected by the fact that the adjective is used to describe the *character of a person* ("Amy is ...") versus their *behavior* ("What Amy did last week is ..."). The authors did not find significant differences, which would potentially speak against the potential influence of animacy on the Polarity Effect. Yet, the behavior of a person might be understood to be indicative of more general character traits, thus blurring the lines between character and behavior condition. The same might be happening when we felicitously use the noun phrase "reckless socks" as a characterization of the person wearing them (*for* wearing them), rather than of the socks themselves. Including more distinct animacy conditions thus might help to further disentangle the Polarity Effect.

Overall, the models presented in this paper appear to be solid baseline models and follow arguably sound theoretical expectations. Future attempts at improving the classifiers would best be directed towards achieving more context-sensitive predictions. The current design is limited to sentiment dispersion in *and*-conjunctions. Future research would profit from additionally including disjunctive or contrastive conjunctions such as *but* and *or*, which function differently from coordinating conjunctions (e.g. Elhadad & Mckeown 1990; Hatzivassiloglou & McKeown, 1997). Admitting adverbial modifiers and intensifiers such as *mostly* or *really*, as well as negations would also contribute to a more complete picture. Other potential improvements include topic-based sentiment analysis or more complex classifiers like support vector machines or ensemble algorithms.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th international conference on language resources and evaluation* (pp. 2200–2204).

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J., & Io, P. (2020). The Pushshift Reddit Dataset. *ArXiv Preprint*, 2001.08435v1.

Baumgartner, L., Reuter, K., & Willemsen, P. (ms). The Polarity Effect of Evaluative Language. *Preprint*, http://philsci-archive.pitt.edu/20146/.

Dahl, Ö. (2008). Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua*, *118*(2), 141–150.

De Swart, P., & De Hoop, H. (2018). Shifting animacy. *Theoretical Linguistics*, *44*(1-2), 1–23.

Eklund, M. (2011). What are Thick Concepts? *Canadian Journal of Philosophy*, *41*(1), 25–49.

Elhadad, M., & McKeown, K. R. (1990). Generating Connectives. In *Proceedings of the 13th conference on computational linguistics* (pp. 97–101).

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th international conference on language resources and evaluation* (pp. 417–422).

García, M. G., Primus, B., & Himmelmann, N. P. (2018). Shifting from animacy to agentivity. *Theoretical Linguistics*, *44*(1-2), 25–39.

Gatti, L., Guerini, M., & Turchi, M. (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, *7*(4), 409–421.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics* (pp. 174–181).

Kirchin, S. (2019). Thick and Thin Concepts. *International Encyclopedia of Ethics*, 1–10.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for computational linguistics system demonstrations* (pp. 101–108).

Reuter, K., Baumgartner, L., & Willemsen, P. (ms). Tracing Thick and Thin Concepts Through Corpora. *Preprint*, http://philsci-archive.pitt.edu/id/eprint/20584.

Roberts, D. (2013). Thick Concepts. *Philosophy Compass*, *8*(8), 677–688.

Schumacher, P. B. (2018). On type composition and agentivity. *Theoretical Linguistics*, *44*(1-2), 81–91.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion*, *59*, 139–162.

Tappolet, C. (2004). Through thick and thin: good and its determinates. *Dialectica*, *58*(2), 207–221.

Väyrynen, P. (2013). *The Lewd, the Rude and the Nasty: A Study of Thick Concepts in Ethics*. Oxford University Press.

Willemsen, P., Baumgartner, L., Cepollaro, B., & Reuter, K. (ms). *Acting in the Zone of Moral Indifference: Social Expectations Explain the Polarity Effect of Evaluative Language.*

Willemsen, P., Baumgartner, L., Frohofer, S., & Reuter, K. (2021). Examining evaluativity in legal discourse : A comparative corpus-linguistic study of thick concepts. *PsyArXiv Preprints*, yxsp9.

Willemsen, P., & Reuter, K. (2020). Separability and the effect of valence: an empirical study of Thick Concepts. In *Proceedings of the 42th annual conference of the cognitive science society* (pp. 794–800).

Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*, *10*(2), 135–146.

Zeldes, A., & Zhang, S. (2016). When Annotation Schemes Change Rules Help: A Configurable Approach to Coreference Resolution beyond OntoNotes. In *Proceedings of the workshop on coreference resolution beyond ontonotes* (pp. 92–101).