On Positivity Bias in Negative Reviews

Madhusudhan Aithal
University of Colorado Boulder
madhuaithal@colorado.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Abstract

Prior work has revealed that positive words occur more frequently than negative words in human expressions, which is typically attributed to positivity bias, a tendency for people to report positive views of reality. But what about the language used in negative reviews? Consistent with prior work, we show that English negative reviews tend to contain more positive words than negative words, using a variety of datasets. We reconcile this observation with prior findings on the pragmatics of negation, and show that negations are commonly associated with positive words in negative reviews. Furthermore, in negative reviews, the majority of sentences with positive words express negative opinions based on sentiment classifiers, indicating some form of negation.

1 Introduction

A battery of studies have validated the Pollyanna hypothesis that positive words occur more frequently than negative words in human expressions, using corpora ranging from Google Books to Twitter (Dodds et al., 2015; Garcia et al., 2012; Boucher and Osgood, 1969; Kloumann et al., 2012). The typical interpretation is connected with the positivity bias, which broadly denotes 1) a tendency for people to report positive views of reality, 2) a tendency to hold positive expectations, views, and memories, and 3) a tendency to favor positive information in reasoning (Carr, 2011; Augustine et al., 2011; Hoorens, 2014). However, it remains an open question whether the Pollyanna hypothesis holds in negative reviews, where the communicative goal is to express negative opinions.

In this work, we use a wide variety of review datasets to examine the use of positive and negative words in negative reviews. Table 1 shows a negative review from Yelp. Although the overall opinion is clearly negative, the author expressed the excitement to try the place and deemed the food OK. Zooming into individual words, they used the same number of positive and negative words in this negative review. Interestingly, this short review has as many as three negations, one directly applied to "great" (hence "not great").

Food was ok...*not* the money they charge. I was unimpressed and will *not* return. I was excited to try this place and was so disappointed as my expectations were high. Service *not* great and The parking is awful.

Table 1: Example negative review on Yelp. Positive words are in blue and negative words are in red, based on Vader (Hutto and Gilbert, 2014). Negations are in italics. This short review contains three negations.

More generally, we find that negative reviews contain *more* positive words than negative words, which is consistent with the Pollyanna hypothesis. Two possible reasons may explain this observation: 1) negative reviews tend to still include positive opinions due to a naïve interpretation of the positivity bias, where positive words express positive sentiments without accounting for negation or other contextual meaning of these words; 2) negative reviews tend to use *indirect expressions* (i.e., applying negations to positive words) to indicate negative opinions (e.g., "not clean"). Note that a broad interpretation of positivity bias may encompass the second reason,[1] but indirect expressions could also be related to other factors, e.g., verbal politeness (Brown et al., 1987)).

We aim to delineate these two reasons by examining the use of negations. Our results provide support for the latter reason: negative reviews tend to use more negations than positive reviews. The

[1] Boucher and Osgood (1969) used a morphological analysis to show negative affixes are more commonly applied to positive words than negative words (unhappy vs. non-violent).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 294–304
August 1–6, 2021.

differences become even more salient when we compare negations applied to positive words vs. negative words. Finally, among sentences with positive words in negative reviews, the majority are classified as negative than as positive by sentiment classifiers, indicating some form of negation.

2 Related Work

In addition to positivity bias, our work is closely related to experimental studies on understanding the effect of direct (e.g., "bad") and indirect (e.g., "not good") wordings. Colston (1999) and Kamoen et al. (2015) observe no difference in people's interpretation of direct and indirect wordings in negative opinions; but direct wordings receive higher evaluations than indirect ones in positive opinions. In this work, we examine whether and how individuals use indirect wordings *in practice* (in negative reviews).

Our work is also related to Potts (2010), which finds that negation is used more frequently in negative reviews and is thus pragmatically negative. We extend Potts (2010) in two ways: 1) we demonstrate a high frequency of negation followed by positive words in negative reviews compared to other combinations, a new observation motivated through the lens of positivity bias; 2) we conduct a systematic study using a wide variety of datasets with multiple dictionaries.

Finally, our work builds on sentiment classification (Pang et al., 2002, 2008; Liu, 2012). The NLP community has made significant progress in recognizing the sentiment in texts of various languages, obtaining accuracies of over 95% (English) in binary classification (Devlin et al., 2019; Liu et al., 2019). Researchers have also developed novel approaches to identify fine-grained sentiments (e.g., aspect-level sentiment analysis (Schouten and Frasincar, 2015; Wang et al., 2016; Yang and Cardie, 2013)) as well as semi-supervised and unsupervised approaches (Hu et al., 2013; Zhou et al., 2010; Tan et al., 2011).

3 Datasets

We use a wide range of English review datasets to ensure that our results are robust across domains.

- Yelp.[2] We only consider restaurant reviews.
- IMDB movie reviews (Maas et al., 2011). This dataset provides train and test splits, so we follow their split when appropriate.

[2] https://www.yelp.com/dataset.

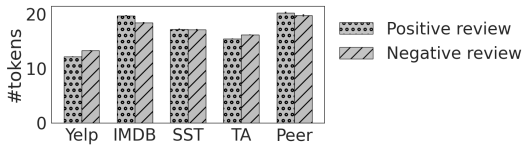


Figure 1: Sentence-length comparison. Although negative reviews can be much longer than positive reviews, sentences in positive reviews and negative reviews have similar lengths. Results on Amazon reviews are shown in the appendix. Tiny error bars show standard errors.

- Stanford sentiment treebank (SST) (Socher et al., 2013). SST contains processed snippets of reviews from the Rotten Tomatoes website (movie reviews). It has ground truth sentiment scores of reviews at the sentence level and the word level.
- Tripadvisor (Wang et al., 2010). This dataset consists of hotel reviews.
- PeerRead (Kang et al., 2018). We use reviews for papers in ACL, CoNLL, and ICLR.
- Amazon (Ni et al., 2019). This dataset contains Amazon reviews grouped by categories. We choose five categories that are substantially different from movies, hotels, and restaurants to ensure that our results are robust, namely, "Automotive", "Cellphones and accessories", "Luxury beauty", "Pet supplies", and "Sports and outdoors".

For datasets with ratings in 1-5 scale, we label reviews with ratings greater than 3 as positive and reviews with ratings less than 3 as negative following prior work (Pang et al., 2002), and ignore reviews with rating 3. Similarly, for datasets with ratings scale of 1-10 (IMDB, ICLR reviews in PeerRead), we label reviews with ratings greater than 6 as positive and review with ratings less than 5 as negative, and ignore reviews with ratings 5 and 6.

We use spaCy to tokenize the reviews in all datasets (Honnibal and Montani, 2017), except that Stanford Core NLP is used to tokenize SST reviews (Manning et al., 2014). We present results for Amazon reviews in the appendix, and our main results are robust on Amazon reviews. Our code is available at https://github.com/madhu-aithal/Positivity-Bias-in-Negative-Reviews.

Length of positive vs. negative opinions. In general, negative opinions tend to be longer than positive opinions ($p < 0.05$ after Bonferroni correction in 6 out 10 datasets; see the appendix for details). In comparison, the difference in length is smaller at the sentence level (Figure 1). Therefore, we use sentences as the basic unit in this work. To further

rule out sentence length as a confounding factor, we also present word-level results in the appendix.

4 Results

We first investigate the occurrences of positive words, negative words, and negations in reviews. We find that negative reviews contain more positive words than negative words in all datasets. We show that this observation relates to the prevalence of negation in negative reviews compared to positive reviews in all datasets. Furthermore, these negations are commonly associated with positive words in all datasets, and sentences with positive words tend to be negative based on sentence-level prediction, supporting the prevalence of indirect wordings in negative reviews.

4.1 Negative Reviews Have More Positive Words than Negative Words

We use lexicon-based methods to examine the frequency of positive and negative words in reviews. For most of the datasets, we randomly sample 5,000 positive reviews and 5,000 negative reviews to compute the lexicon distribution using LIWC (Pennebaker et al., 2007) and Vader (Hutto and Gilbert, 2014). In the case of SST, PeerRead, and negative reviews of Amazon Luxury Beauty, we use the entire dataset for our analysis as it has a relatively small number of reviews.

Figure 2 shows that as expected, negative reviews have more negative words and fewer positive words than positive reviews, based on Vader. Intriguingly, despite the negative nature of negative reviews, they tend to have more *positive* words than *negative* words ($p < 0.001$ on all datasets except SST after Bonferroni correction). Our results are robust at the word level and also hold based on LIWC and validate the Pollyanna hypothesis even in negative reviews.

4.2 Negative Reviews Have More Negations and Indirect Expressions

We hypothesize that in addition to the tendency to report positive views of reality, an important factor that can explain this observation in negative reviews is the use of indirect expressions (i.e., negation of positive words). To measure the amount of negation, we use two approaches: 1) a lexicon-driven approach based on Vader including *aint*, *cannot*, *not*, and *never* (Hutto and Gilbert, 2014)[3]; 2)

[3]See the appendix for the full list of negation lexicons.

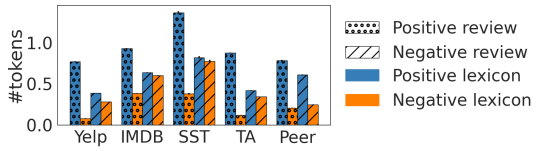
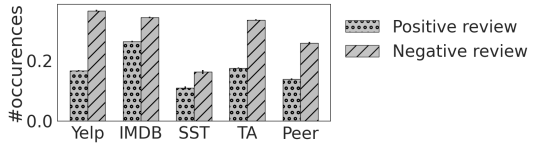
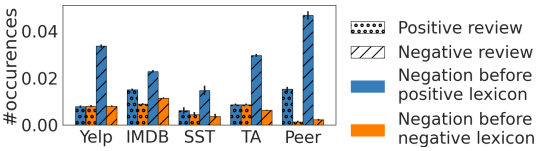


Figure 2: Number of positive and negative words based on Vader. Negative reviews have more positive words than negative words.



(a) Overall negation.



(b) Negation before positive and negative lexicons.

Figure 3: Negative reviews generally have more negations at the sentence level (Figure 3a). Among those negations, Figure 3b shows that there are substantially more negations before positive lexicons in negative reviews than any other combinations.

the negation relation in dependency parsing.[4] We present the results based on Vader negation in the main paper as it may have higher precision, and all results hold using dependency parsing.

Negative reviews have more negations than positive reviews in all datasets. Figure 3a presents the number of negations at the sentence level. In all datasets, negative reviews have more negations than positive reviews ($p < 0.001$ in all datasets). In fact, the number of negations in negative reviews almost doubles that in positive reviews in Yelp, TripAdvisor, and PeerRead (see samples in the appendix). This observation is robust at the word level, which accounts for the fact that negative reviews tend to be longer.

Negations are commonly associated with positive words in negative reviews. To further examine the relation between negations and sentimental lexicons, we investigate the occurrences of negations immediately followed by positive words and negative words. Figure 3b shows that there are more negations before positive words in negative reviews than any other combination ($p < 0.001$ in all datasets). The difference is especially salient in

[4]We used spaCy for dependency parsing (Honnibal and Montani, 2017).

Dataset	Positive words associated with negations
Yelp	recommend, sure, like, good, care, great, special, impressed, fresh, help, ready, enjoy, friendly, honor, helpful, clean, happy, accept, greeted, amazing
IMDB	like, care, funny, help, sure, recommend, good, save, fit, great, special, interesting, enjoy, well, play, better, giving, original, convincing, true
PeerRead	clear, sure, convincing, convinced, ready, well, true, clearly, surprising, novel, convincingly, recommend, guarantee, improve, interesting, support, satisfactory, help, acceptable, convince

Table 2: Most frequent positive words that immediately follow negations in negative reviews, based on Vader.

Yelp, TripAdvisor, and PeerRead. In particular, in negative reviews in PeerRead, negation before positive lexicon are approximately 20 times as frequent as negation before negative lexicon. These results demonstrate the prevalence of indirect wordings when people express negative opinions. Moreover, using indirect expressions to express negative opinions (negation before positive words) is also common in positive reviews for IMDB and PeerRead.

Table 2 shows the 20 most common words that immediately follow negations in Yelp, IMDB, and PeerRead, highlighting the prevalence of "not clear", "not convincing", and "not surprising" in negative reviews of NLP/ML submissions.

A natural question is how much of the usage of positive words in Figure 2 can be explained by negations before positive words. We find that it is sufficient to explain 11.3% on average. For instance, negative reviews in Yelp have 0.389 positive words per sentence, out of which 0.033 words follow a negation. This accounts for 8.7% of the usage of positive words. This suggests that negations before positive words only account for a small fraction of positive words, despite that they dominate other combinations of negations and sentiment lexicon. We hypothesize for positive words in negative reviews, they may be negated in ways beyond immediate preceding negations (e.g., "nor is the food great" and "fail to support").

Similarly, the number of negations followed by positive/negative words is a fraction of all the negations (14.2% in negative reviews and 9.7% in positive reviews). For example, "I will not return" counts as negation but there is no sentimental lexicon. We hypothesize that these negations also tend to express negative sentiments.

4.3 Sentence-level Sentiment Classification

To capture the sentiment of sentences with positive words or negations beyond negations immediately followed by positive words, we rely on sentiment classifiers. Specifically, we use sentence-level classification to quantify the extent of negative sentences in those contexts compared to the overall average in negative reviews.

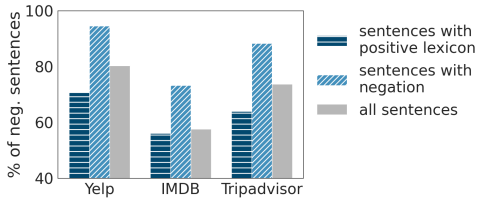
We fine-tune BERT (Devlin et al., 2019) to perform review-level classification for each dataset except SST and PeerRead. This is because all reviews in SST are very short and sentences in negative reviews are mostly negative whether negation occurs or not. In the case of PeerRead, the number of samples is too small to fine-tune the BERT model. For all other datasets except IMDB and Amazon Luxury Beauty, we randomly sample 25K positive reviews and 25K negative reviews as the training set, and 5K positive reviews and 5K negative reviews as the test set. For IMDB, we use 12.5K positive and 12.5K negative training samples provided for fine-tuning, and for Luxury Beauty, we use a balanced set of 2.3K positive and 2.3K negative samples for fine-tuning. We use 90% of the training samples to fine-tune the BERT model and 10% as the development set to select hyperparameters. We achieved accuracies varying from 94% to 98% for the test set reviews in all datasets. See the appendix for details of the data split and accuracies.

We use the BERT model fine-tuned on reviews to predict sentiment of sentences. Note that this prediction entails a distribution shift as sentences are shorter than full reviews used to fine-tune BERT models. However, this is a common strategy for evaluating rationales in the interpretable machine learning literature and there exists evidence that transformer-based models provide strong performance despite the distribution shift in the form of reduced inputs (DeYoung et al., 2020; Hsu et al., 2020; Carton et al., 2020).[5]

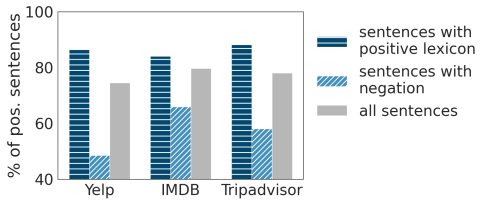
Figure 4a shows that sentences with positive words in negative reviews are more likely to be negative than to be positive (65.1% on average across all datasets; notably, IMDB is lower but still at 56.13%, above 50%).[6] It suggests that the majority of positive words are negated in some way. While the remaining minority of sentences with positive

[5] Bastan et al. (2020) investigates the reverse direction, i.e., from paragraph-level predictions to document-level predictions.

[6] Similar trends hold if we adjust the estimates using TPR, TNR, FPR, and FNR. See the appendix.



(a) Fractions of negative sentences in negative reviews.



(b) Fractions of positive sentences in positive reviews.

Figure 4: Sentence-level prediction results based on fine-tuned BERT classifiers. In negative reviews, sentences with positive words tend to be negative, and sentences with negations are overwhelmingly negative. In comparison, sentences with negations are more balanced (44.7% negative) in positive reviews.

words are indeed positive and align with the tendency to report positive views, our results highlight the important role of indirect expressions in explaining the positive words in negative reviews.

Furthermore, sentences with negation tend to be negative (88.6%) based on our classifiers, confirming our hypothesis that most negations are used to express negative sentiments in negative reviews. This is even higher than the average fraction of negative sentences (73.1%) among all sentences in negative reviews. In comparison, Figure 4b shows that positive words in positive reviews tend to reflect positive sentiments, indicating no common use of negation associated with positive words. Meanwhile, negations are not usually associated with negative sentiments in positive reviews (44.7%), substantially lower than negations associated with negative sentiments in negative reviews (88.6%).

5 Conclusion

In this paper, we investigate positivity bias in negative reviews and highlight the role of indirect expressions in understanding this phenomenon. We show that negations followed by positive words are more prevalent than any other combination in negative reviews. Given that these indirect wordings account for only 11.3% of the occurrences of positive words in negative reviews, we further show that such sentences with positive words tend to be negative, based on sentiment classifiers.

While our findings support the prevalence of indirect expressions, we do not take sentiment intensity into account. In practice, "not good" provides a different meaning from "not amazing". We believe exploring the relationship between negation and semantic intensity is a promising direction. Our lexical-driven approaches are limited by the lexicons included in the dictionaries, which are typically evaluated independent of the context, so their sentiment may be different in the specific context.[7] Similarly, our sentence-level prediction results are limited by the distribution shift when applying BERT trained on documents to sentences. It is reassuring that our high-level results hold across multiple datasets based on both lexical-driven approaches and sentence-level prediction.

As our study focuses on negative reviews in English, it is important to examine the generalizability of our results. For instance, it is important to understand to what extent the observed positivity bias in general expressions is driven by such indirect expressions. Another natural extension is to investigate other languages. Although our findings are limited to English reviews, we believe that they may be applicable to negative opinions in other languages, as Pollyanna hypothesis (Boucher and Osgood, 1969) has been validated across languages and cultures. Finally, our work has implications for sentiment-related applications in NLP. The prevalence of indirect expressions in negative reviews underscores the importance of modeling and understanding negation in sentiment analysis and sentiment transfer (Ettinger, 2020).

In general, we believe that online reviews not only provide valuable data for teaching machines to recognize sentiments but also allow us to understand how humans express sentiments. We hope that our work encourages future work to further investigate the framing choices when we express emotions and opinions, and their implications on NLP applications.

Acknowledgments

We thank anonymous reviewers and the members of the Chicago Human+AI lab for their helpful comments. This work was supported in part by an Amazon research award, a Salesforce research award, and NSF IIS-1941973.

[7] One reviewer pointed out an interesting hypothesis: judges assume the nicest interpretation of a word out of context in the annotation process, as a result, the Pollyanna hypothesis may be an instrumentation bias.

References

Adam A Augustine, Matthias R. Mehl, and Randy J. Larsen. 2011. A Positivity Bias in Written and Spoken English and Its Moderation by Personality and Gender. *Social Psychological and Personality Science*, 2(5):508–515. Publisher: SAGE Publications Inc.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In *Proceedings of COLING*.

Jerry Boucher and Charles E Osgood. 1969. The pollyanna hypothesis. *Journal of verbal learning and verbal behavior*, 8(1):1–8.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Alan Carr. 2011. *Positive psychology: The science of happiness and human strengths*. Routledge.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of EMNLP*.

Herbert L Colston. 1999. "not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes*, 28(3):237–256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*.

Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. 2015. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

David Garcia, Antonios Garas, and Frank Schweitzer. 2012. Positive words carry less information than negative words. *EPJ Data Science*, 1(1):3.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Vera Hoorens. 2014. Positivity Bias. In Alex C. Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 4938–4941. Springer Netherlands, Dordrecht.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. *Finding of EMNLP*.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of WWW*.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM*.

Naomi Kamoen, Maria BJ Mos, and Willem FS Dekker. 2015. A hotel that is not bad isn't good. the effects of valence framing and expectation in online reviews on text, reviewer and product appreciation. *Journal of Pragmatics*, 75:28–43.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of NAACL*.

Isabel M Kloumann, Christopher M Danforth, Kameron Decker Harris, Catherine A Bliss, and Peter Sheridan Dodds. 2012. Positivity of the english language. *PloS one*, 7(1):e29484.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of ACL (system demonstrations)*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of ACL*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Christopher Potts. 2010. On the negativity of negation. In *Semantics and Linguistic Theory*, volume 20, pages 636–659.

Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of KDD*.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of KDD*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of EMNLP*.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of ACL*.

Shusen Zhou, Qingcai Chen, and Xiaolong Wang. 2010. Active deep networks for semi-supervised sentiment classification. In *Proceedings of COLING*.

A Vader Lexicons

Table 3 shows the list of negation lexicons in Vader.

aint, arent, cannot, cant, couldnt, darent, didnt, doesnt, ain't, aren't, can't, couldn't, daren't, didn't, doesn't, dont, hadnt, hasnt, havent, isnt, mightnt, mustnt, neither, don't, hadn't, hasn't, haven't, isn't, mightn't, mustn't, neednt, needn't, never, none, nope, nor, not, nothing, nowhere, oughtnt, shant, shouldnt, uhuh, wasnt, werent, oughtn't, shan't, shouldn't, uh-uh, wasn't, weren't, without, wont, wouldnt, won't, wouldn't, rarely, seldom, despite

Table 3: Negation lexicons in Vader used for our negation analysis.

B Samples from PeerRead

Table 4 shows a list of 6 sentences with negation selected from random negative PeerRead reviews. Negations are mostly associated with positive words, both directly and indirectly.

Please do *not* make incredibly unscientific statements like this one :"
I'm *not* convinced about the value of having this artificial dataset.
For example, at the end of sec 4.4, " This result is *not* surprising, given that FOV-R contains additional information ....
It is *not* clear whether the improvements (if there is) of the ensemble disappear after data-augmentation.
Empirical analysis is *not* satisfactory.
But I'm *not* sure from reading the paper.

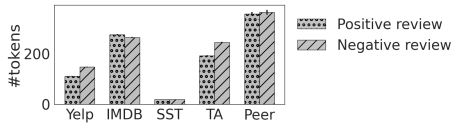
Table 4: Sentences with negation sampled from negative reviews of PeerRead. Positive words are in blue and negative words are in red. Negations are in italics.

C Additional Plots

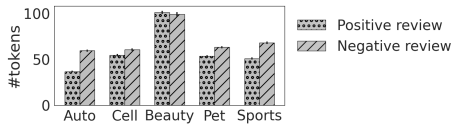
Length distribution. See Figure 5 for review-level length and Figure 6 for sentence-level length distribution for Amazon reviews.

Lexicon distribution. Figure 7 shows the sentiment lexicon distribution of all reviews using LIWC. Figure 8 shows the lexicon distribution of Amazon reviews using Vader.

Negation distribution. See Figure 9 and Figure 11 for the negation distribution of Amazon reviews using Vader and dependency parsing respectively. Figure 10 shows the negation distribution found using dependency parsing for non-Amazon reviews.



(a) SST, Yelp, IMDB, and Tripadvisor (non-Amazon datasets).



(b) Amazon datasets.

Figure 5: Review-level length distribution. This shows the length comparisons of positive and negative reviews of different datasets. The values represent the average number of tokens present in each review. Negative reviews are longer than positive reviews in all datasets except IMDB and Amazon Luxury Beauty.

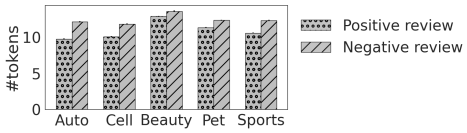
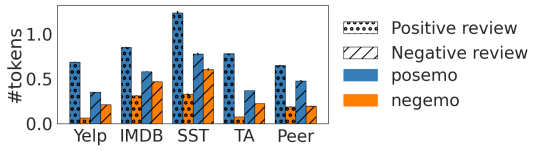
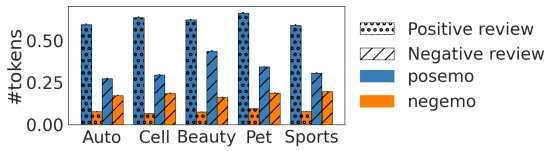


Figure 6: Sentence-level length distribution of Amazon datasets.



(a) Non-Amazon datasets.



(b) Amazon datasets.

Figure 7: Lexicon distributions based on LIWC. Figure 7a and Figure 7b shows the lexicon distribution of reviews using *posemo* and *negemo* LIWC categories. In all datasets, negative reviews have fewer positive emotions than positive reviews. They also have more positive words than negative words. This trend is similar to that obtained using Vader lexicons in case of non-Amazon reviews.

In case of negation distributions found using dependency parsing, we used Vader to identify positive and negative words.

Sentiment predictions. See Figure 4a and Figure 12 for the fractions of negative sentences in negative non-Amazon reviews measured by the BERT model. See Figure 13 for fractions of negative sentences in negative reviews of Amazon. Figure 14

Dataset	Training set	Validation set	Test set	Test accuracy (%)
Yelp	45000	5000	10000	97.51
IMDB	22500	2500	10000	94.38
Tripadvisor	45000	5000	10000	96.66
Automotive	45000	5000	10000	95.65
Cellphones and accessories	45000	5000	10000	95.39
Luxury beauty	4195	467	3040	96.10
Pet supplies	45000	5000	10000	95.60
Sports and outdoors	45000	5000	10000	95.12

Table 5: Dataset split and test accuracies of BERT fine-tuning. For all datasets except IMDB, Luxury Beauty, we use 45K samples as training set, 5K as validation set, and 10K as test set, randomly sampled from the entire dataset. In the case of IMDB, we use 22.5K samples for training and 2.5K samples for validation, randomly sampled from the provided training set of size 25K. We then use 10K samples randomly sampled from the provided test set of size 25K for testing purposes. In the case of Amazon Luxury Beauty, we use a balanced set of 4195 samples for training and 467 samples for validation. We then use 3K samples (imbalanced) randomly sampled from the dataset for testing. All these random samplings were done without replacement.

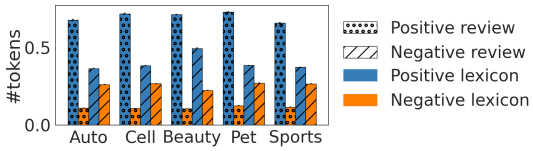
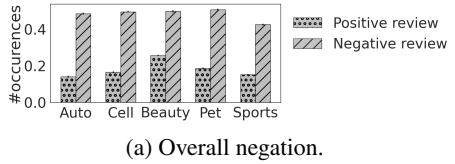
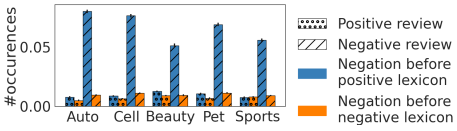


Figure 8: Lexicon distribution of Amazon datasets using Vader. Negative reviews have more positive words than negative words, similar to the trend in SST, Yelp, IMDB, Tripadvisor, and PeerRead.

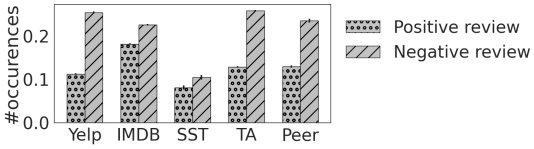


(a) Overall negation.

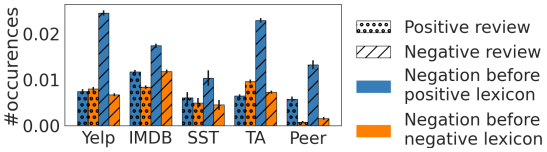


(b) Negation before positive and negative lexicons.

Figure 9: Negation distribution of Amazon datasets using Vader lexicons. Negative reviews use more negation words compared to positive reviews. Negative reviews have substantially more negation words associated with positive words than negative words.



(a) Overall negation.



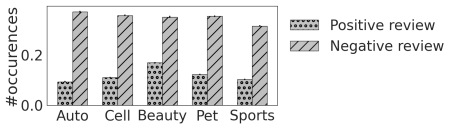
(b) Negation before positive and negative lexicons.

Figure 10: Negation distribution using dependency parsing - non-Amazon datasets. In all non-Amazon datasets, negative reviews use more negation words than positive reviews. This observation is inline with the negation results obtained using Vader lexicons. Dependency parsing is used to extract negations from reviews, and to identify words associated with a negation word.

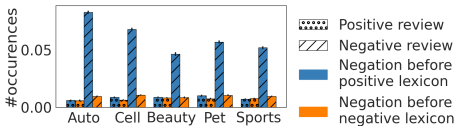
shows the fractions of positive sentences in positive reviews. Some of the fractions in our results are computed based on the TPR, TNR, FPR, and FNR of the BERT model. We used test set reviews of the datasets to compute these metrics as they give more accurate estimate of percentage of positive and negative sentences in reviews. All BERT classifiers that we used for predicting the sentiment of sentences are fine-tuned using the reviews of corresponding datasets. Table 5 shows the dataset split and test accuracies of BERT fine-tuning.

Hyperparameter tuning. We did hyperparameter tuning by varying number of epochs, batch size, and learning rate. We fine-tuned BERT for 4 epochs with batch sizes of 2, 4 and 8, with a learning rates of 1e-5 and 2e-5. Based on validation accuracies, the model trained for 2 epochs, with a batch size of 8 and learning rate of 2e-5 turned out to be the best performing model for most of the datasets.

Word-level results. Figure 15 shows the lexicon distribution using LIWC and Vader. See Figure 16 and Figure 17 for word-level results of negation distribution using Vader and dependency parsing respectively.



(a) Overall negation.



(b) Negation before positive and negative lexicons.

Figure 11: Negation distribution using dependency parsing - Amazon datasets. Figure 11a shows that negative reviews have substantially more negation words than positive words. Figure 11b shows the negation distribution associated with positive and negative words. This corresponds to about 16.68% of all negation words used in the positive and negative reviews based on our dictionary. Negative reviews also have substantially more negations before positive words, compared to other combinations.

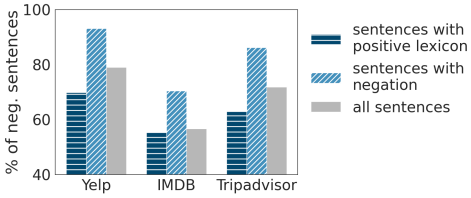
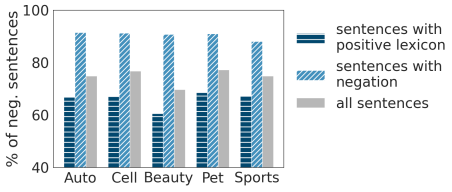
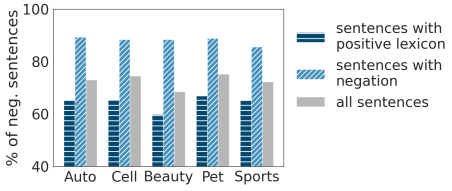


Figure 12: Fractions of negative sentences in negative reviews of Yelp, IMDB, and Tripadvisor. These fractions are corrected using TPR, TNR, FPR, and FNR. It can be seen that higher proportion of negative reviews with negation are classified as negative by our BERT model. This shows that negations in negative reviews are mostly used to express negative opinions. This observation holds for other datasets also.

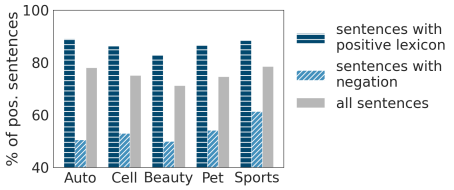


(a) Fractions based on accuracy.

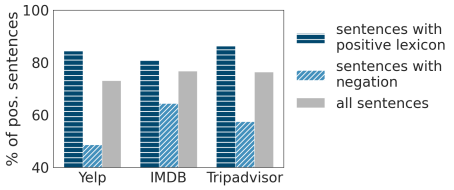


(b) Fractions based on TPR, TNR, FPR, and FNR.

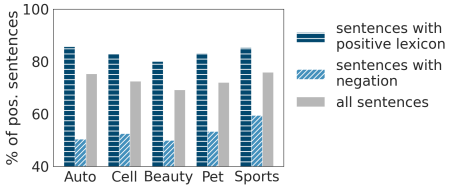
Figure 13: Fractions of negative sentences in negative Amazon reviews based on fine-tuned BERT classifiers. The distribution confirms our hypothesis that most negations are used to express negative sentiments.



(a) Fractions based on accuracy.

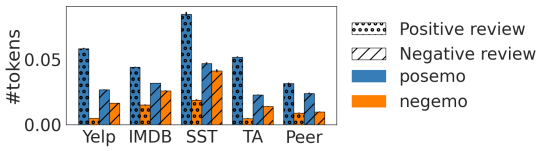


(b) Fractions based on TPR, TNR, FPR, and FNR.

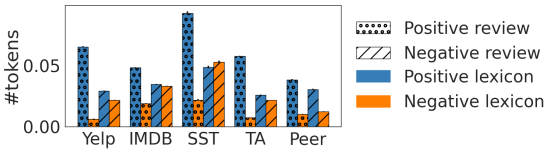


(c) Fractions based on TPR, TNR, FPR, and FNR.

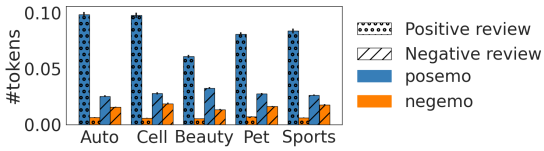
Figure 14: Fractions of positive sentences in positive reviews. We can see that negations in positive reviews are more balanced with positive and negative sentences when compared to negative reviews. Also, sentences with positive lexicons are mostly positive (86.5%). There are very few negative sentences with positive lexicons. This holds for all datasets.



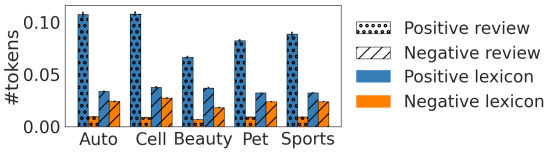
(a) LIWC distribution of non-Amazon datasets.



(b) Vader distribution of non-Amazon datasets.

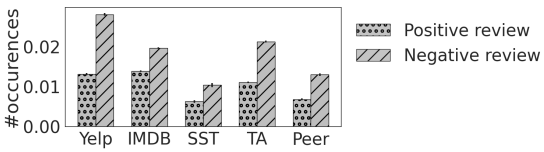


(c) LIWC distribution of Amazon datasets.

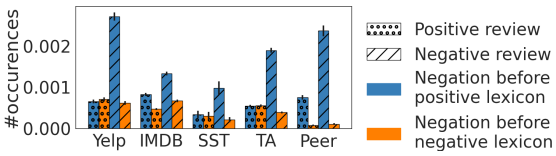


(d) Vader distribution of Amazon datasets.

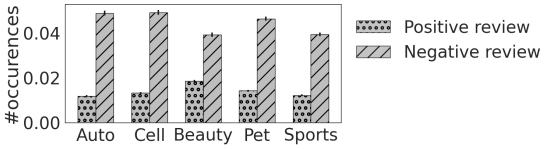
Figure 15: Word-level lexicon distribution. At the word-level, positive reviews have more positive words than negative reviews. However, negative reviews contain more positive words than negative words (except SST with Vader). The trend that we observe in the sentence-level results can be seen here as well.



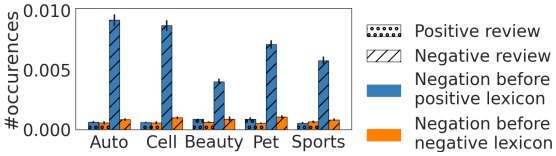
(a) Non-Amazon datasets.



(b) Non-Amazon datasets.

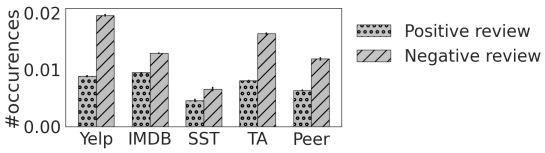


(c) Amazon datasets.

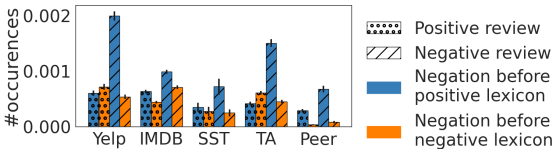


(d) Amazon datasets.

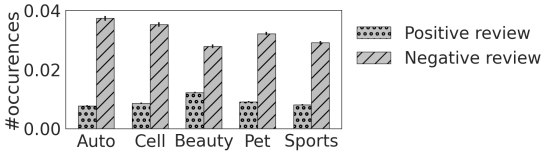
Figure 16: Word-level negation distribution using Vader. Figure 16a and Figure 16c indicate the more frequent use of negation in negative reviews than in positive reviews at the word-level. Negative reviews have more negations before positive words in all datasets. This difference is substantially large in case of Yelp, Tripadvisor, PeerRead and Amazon reviews. This shows that although negative reviews have more positive words than negative words, these positive words are associated with negations.



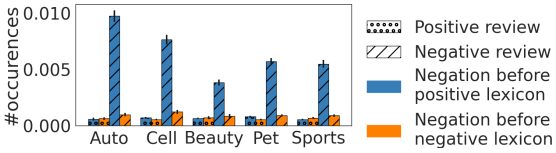
(a) Overall negation in non-Amazon datasets.



(b) Negation before positive and negative lexicons in non-Amazon datasets.



(c) Overall negation in Amazon datasets.



(d) Negation before positive and negative lexicons in Amazon datasets.

Figure 17: Word-level negation distribution of all reviews using dependency parsing. With dependency parsing, we observe the same pattern as in Figure 16. Negative reviews in Yelp, Tripadvisor, PeerRead and Amazon datasets have substantially more negations in general and also before positive words. This high number of negation associated with positive words partially explains the higher proportion of positive words in negative reviews.