Lexicon Creation for Interpretable NLP Models

Anonymous ACL submission

Abstract

Lexica–words and associated scores–are widely used as simple, interpretable, generalizable language features to predict sentiment, emotions, mental health, and personality traits. Applying different feature importance methods to different predictive models yields lexica of varying quality. In this paper, we train diverse sequence classification models, including context-oblivious (SVMs, Feed-forward neural networks) and context-sensitive (RoBERTa, DistilBERT) models, and generate lexica based on different feature importance measurements, including attention, masking, and SHAP (SHapley Additive exPlanations) values. We evaluate the generated lexica on their predictive performance on test sets within the same corpus domain and on their generalization to different but similar domains. We find that simple context-oblivious models produce lexica of similar accuracy within domain and of better accuracy across domains to those from complex context-sensitive models. Based on human evaluator ratings of these lexica, we also find that context-oblivious models generate similar lexica that are more aligned with human judgments.

1 Introduction

Lexica - sets of words, often with associated weights - are commonly used to characterize text with respect to features such as emotion (Mohammad et al., 2018; Kušen et al., 2017; Bandhakavi et al., 2017; Goel et al., 2017) style (Danescu-Niculescu-Mizil et al., 2013; Pavlick and Tetreault, 2016), political orientation, and attributes of the writer such as age, gender, happiness, and personality (Alm et al., 2005; Eichstaedt et al., 2018; Plank and Hovy, 2015; Preotiuc-Pietro et al., 2016; Schwartz et al., 2013). Such lexica are often manually created using Amazon Mechanical Turk or similar crowd workers (Dodds et al., 2015; Haralabopoulos and Simperl, 2017), and are widely used in fields such as psychology and political science. The most popular lexicon used in psychology, Linguistic Inquiry and Word Count (LIWC), has been cited over ten thousand times and translated into many different languages (Pennebaker et al., 2001). Other lexica, such as the NRC ones for predicting sentiment and emotion (SM Mohammad, 2013), are learned using linear regression on large data sets of labeled text or by computing similarities of vector embeddings of words (Sap et al., 2014; Sedoc et al., 2017).

Lexica are widely used in the social sciences both because they are easy to use–psychologists and political scientists are generally more comfortable with linear regression than with deep learning–and, more importantly, because they are easy to interpret. Social scientists are generally more interested in understanding phenomena than in maximizing predictive accuracy; to that end, they continue to develop and use a wide variety of lexica, ranging from broad-coverage psychology lexica such as LIWC (Pennebaker, 2011) (with over 70 categories including affiliation, achievement, and drive as well as syntactic categories such as pronouns and interrogatives) to specialized lexica such as those for word concreteness and familiarity (Brysbaert et al., 2014; Paetzold and Specia, 2016).

Closely related to lexica is "feature importance", which also computes a strength of association between words and an outcome of interest to support interpretation. A variety of methods are used to extract feature importance from neural networks and other machine-learned models (Kim et al., 2020a; Li et al., 2017a; Lundberg and Lee, 2017). One of the most popular feature importance measures is SHAP (Lundberg and Lee, 2017), a mathematically principled way of computing feature importances based on Shapley values.

Feature importances serve different functions, such as to "explain the model" (i.e. to understand why the model makes given predictions), versus to

"explain the world" (i.e. to provide insight into the data on which the model is trained) (Chen et al., 2020). For example, when extracting feature importance from the sentence, "The food was beautiful and delicious!", an attentional model might show that the highest attention was given to the word "and", the words with the highest Shapley values in a deep learned network might be "food" and "!", while a hand-compiled list of positive and negative words might select "beautiful" and "delicious". Lexica are generally designed to explain the world, and we focus on that use case.

We know that more sophisticated context-sensitive models (e.g., BERT) tend to make more accurate predictions than simpler ones (e.g., SVMs or simple neural networks using context-free word embeddings). However, sophisticated models are also often more difficult to explain and hence might yield less understandable lexica.

Good lexica should also generalize well from one domain (e.g., food reviews on Yelp) to another (e.g., student recommendations). It is not obvious whether using sequence information and context will yield lexica that generalize better, or merely that fit better within domain.

Since lexica are context-free language models, deriving them from sophisticated context-sensitive NLP models may or may not yield lexica that are better within or across domains. Thus, we want to know how much of a performance drop one should expect when replacing sophisticated models with easily interpretable lexica, both for within domain and across domains.

The paper contains three closely related sets of analyses:

- We compare a range of models (SVM, FNN, LSTM, BERT-related ones) on different prediction tasks and assess how well the models generalize to different corpus domains.
- We generate lexica from each of these models using a several feature importance measures, and assess how accurately they perform on the same domain, and how well they generalize to other domains.
- We show the extracted lexica to human evaluators to assess how well the extracted words correspond with human intuition.

The main findings are:

- Within the training corpus domain, although the context-sensitive models outperforms the context-oblivious ones, both types of models produce lexica with similar predictive performance.
- Lexica generated from simpler context-oblivious models have better across-domain generalization performance than those from more complex context-sensitive models.
- Lexica generated from different context-oblivious models are correlated and align better with human intuitions than those from context-sensitive models.

The code and data for all experiments available on GitHub.[1]

2 Related Work and Research Goals

We define a lexicon as a list of words along with associated scores, where higher scores correspond to some notion of word importance.

Lexicon creation was traditionally done manually. In psychology, lexica such as LIWC were created based on judgements of expert annotators (Pennebaker, 2011). LIWC is unweighted, and can be viewed as having a weight of 1 for all words in the lexicon. More rigorously weighted lexica have been created using crowd-sourced annotations, such as labMT (Garcia et al., 2015).

Recent work in computer science induces lexica using computational approaches (Pryzant et al., 2018). Lexica can be generated by methods ranging from using linear regression coefficients to computing word scores by "inverting" feed forward network (Sedoc et al., 2020). The word-wise score can also be obtained using attention distributions or word frequency vectors. The extracted lexica have been applied to many tasks, including feature extraction (Mohammad et al., 2018), emotion prediction (Sedoc et al., 2020), linguistic analysis, or causal domain theories (Pryzant et al., 2018).

Although the term "lexicon" is often not explicitly mentioned, methods that compute the feature importance of words in machine-learned models produce lexica. These approaches generally use coefficients from linear models or explain by analyzing the inputs and outputs of the models using methods such as Shapley values (Ribeiro et al., 2016; Lundberg and Lee, 2017)

For linear models, lexica can be constructed by directly using the coefficients or weights in the model. Similarly, attention weights in more complex neural networks can serve as lexica (Bahdanau

[1] https://github.com/xxx/lexica_creation Code and data will be released upon acceptance.

et al., 2016; Luong et al., 2015). Attention provides some insights into certain types of models and tasks (Vashishth et al., 2019), but it is less clear whether it produces useful lexicon weights or faithful explanations (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019).

With the introduction of transformers (Vaswani et al., 2017), more complex context-sensitive models such as BERT(Devlin et al., 2019) (and variations such RoBERTa (Liu et al., 2019) and DistilBERT(Sanh et al., 2020)) often provide significantly better predictive performance.

Understanding these more complex models by inspection is infeasible. Instead, by observing the effect on the output of carefully designed perturbations of each word (e.g., via removal or masking), we can compute its importance (Kim et al., 2020b).

Shapley values and their many variations and approximations are a key example of such input perturbations (Lundberg and Lee, 2017). Among them, deepSHAP is designed based on DeepLIFT for deep networks and partitionSHAP computes Shapley value for clustered features, which is equivalent to computing the Owen values and provides a contextualized understanding of the input.

Many feature importance methods, such as marginal Shapley values are designed to "explain the model." The generated lexica thus contain words that are important to the models, which are not necessarily the words that are important for understanding the world. For example, attention weights may focus on the word 'and', rather than adjacent words. We seek feature importance such as conditional Shapley values that seek to 'explain the world'; similarly, psychologists want to answer questions like "What words typify empathetic people?"(Buechel et al., 2018) and "What does Twitter language of people with ADHD reveal about how they perceive the world?" (Guntuku et al., 2019)

To date, there is no broad assessment of the ability of generated lexica to 'explain the world'. Our goal is assess this. Previous research compares the lexica created by different models and different metrics in respect of similarities of the most important features obtained from each approaches. (Lai et al., 2019) compares feature importances across different models and different feature importance metrics. However, the comparisons do not judge the relative quality of the lexica either subjectively or by their generalization accuracy.

Bearing in mind how social scientists actually use lexica, we focus on evaluating the interpretability and generalizability of the lexica generated by different approaches. (Lai et al., 2019) show that some models provide similar explanations regardless of the feature importance metric used. We therefore choose a set of popular models with differing levels of complexity, along with the suitable interpretation for each, and test them on diverse sentiment and emotion datasets. The generalizability of the lexica are evaluated both within and across these datasets. Finally, human evaluations are used to assess the quality of lexica-produced interpretation.

3 Datasets

Our experiments use a mixture of common broad-coverage datasets such as Yelp, Amazon reviews, and NRC Emotion; and relatively tailored datasets such as EmoBank, Daily Dialog, and Song Lyrics. For large datasets, we use their balanced subsets. Table 1 shows Dataset references and basic statistics. These datasets are from a variety of sources including Twitter, song lyrics, newswire, online reviews, and crowd-sourced writing. They vary by size, average length, and vocabulary size. This variety of datasets ensures more robust and fair comparisons between the lexica generalization methods.

Labels of all datasets are processed so that they could be used for binary classifications. The datasets can be divided into two categories. The Yelp and Amazon datasets are for sentiment classification: the models classify reviews as being positive or negative. For these datasets, we are interested in both the 'head' and 'tail' of the resulting lexica, as they indicate positivity and negativity, respectively. For datasets including NRC, Dialog, and Song Lyrics, models do binary classification for five different emotions (joy, fear, anger, sadness, and surprise). In these cases, we are only interested in the 'head' of the lexica, because those are the words most closely associated with the corresponding emotion.

To allow easy comparison, this work is done entirely in English; non-English words in the NRC datasets are removed.

4 Lexicon Generation Methods

A lexicon generation method consists of a predictive model, based on context-oblivious or context-sensitive embeddings and an associated means of computing feature importance.

Datasets		Training/Validation Size	Test Size	Mean Seq Length
Yelp_Subset [www.yelp.com/dataset]		27592/3398	3426	132.8
Amazon_FineFood_Subset (McAuley and Leskovec, 2013)		25794/3258	3188	96.3
Amazon_Toys_Subset (He and McAuley, 2016)		17666/2094	2158	125.9
NRC (SM Mohammad, 2013)	Joy	12646/1576	1548	18.3
	Fear	4046/510	578	19.1
	Anger	2390/270	322	19.2
	Sadness	5780/780	662	18.3
	Surprise	4886/600	606	18.2
Song (Mihalcea and Strapparava, 2012)	Joy	Only Used for Evaluation	202	55.8
	Fear		262	56.0
	Anger		284	56.4
	Sadness		298	55.8
	Surprise		302	55.6
Dialog (Li et al., 2017b)	Joy		8134	14.5
	Fear		314	15.8
	Anger		1872	15.9
	Sadness		2150	15.0
	Surprise		3134	13.6
Emotionlines (Hsu et al., 2018)	Joy		3420	10.2
	Fear		492	11.3
	Anger		1518	10.8
	Sadness		996	11.8
	Surprise		3314	9.8
Emobank_Valence (Buechel and Hahn, 2017)			7410	18.0

Table 1: Datasets Information

4.1 Feature Importance Measurements

We explore several feature importance measurements, including uni-variant, "Single-Token Importance" (computing model outputs for embeddings of single words), attention weights, masking, and partitionSHAP.

Uni-variant The most easy-to-apply method for lexicon generation is to calculate the correlation between word frequencies and sentence scores. Specifically, for one word, we count its frequencies of occurrence in every sentence in the dataset, and calculate the Pearson correlation between the word's counts and sentence labels, i.e. scores, as the word's score for the lexicon.

Single-token Importance (STI) Bag-of-Word models like SVMs and FFNs allow us to feed in text data as vector embeddings. In our experiments, we average FastText embeddings of all the tokens in each text. The averaging results in context-oblivious text embeddings that lie in the same embedding space as tokens. We can thus compute feature importances for tokens by feeding their embeddings directly into models trained on text embeddings. Then the outputs of the models serve as their relative importance. We call this 'Single-Token Importance' (STI) measurement.

Attention weights Attention has been used for model interpretation, with focus on the intermediate outputs of the encoder used to measure the relative importance of the token. However, there is still ongoing debate about what such measurement is actually explaining. Some authors claim that attention weights do not explain the reasoning behind model predictions (Jain and Wallace, 2019; Serrano and Smith, 2019), while others claim that attention weights do capture linguistic insights and can explain the models' decisions (Vashishth et al., 2019; Wiegreffe and Pinter, 2019). Others argue that attention often have little function, since a random permutation of the attention coefficients does not significantly affect the predictions. (Vashishth et al., 2019)

Masking The importance of a token can be measured by the change in the model output that results when the token is preplaced with a special mask token. This allows us to explain sophisticated models by simple input perturbation without having to

make sense of millions of model parameters (Li et al., 2017a).

PartitionSHAP SHAP (SHapley Additive exPlanations) values allow more sophisticated ways of evaluating the contributions of features to the model prediction, enabling the replacement of a token and associated tokens with words drawn from a background distribution. As described above, some Shapley values (marginal) explain the model; we prefer conditional ones that attempt to explain the word by taking account of the correlation between words in each sentence. PartitionSHAP is a variation of SHAP that uses a hierarchical clustering of the features. As a result, it is essentially computing the Owen values from the game theory. PartitionSHAP assumes independence between groups of features instead of individual ones. The feature clustering can be done based on correlations, or any distance metric, or predefined rules (e.g., tokens in a cluster must be adjacent). PartitionSHAP attributes to the clusters instead of individual features in the clusters.

PartitionSHAP is a model-agnostic evaluation. However, it is much faster than other model-agnostic SHAP methods such as kernelSHAP (Lundberg and Lee, 2017), as the complexity of partition SHAP is quadratic in the number of input features while the other methods are exponential in theory .

4.2 Lexicon Generation Models

Most feature importance measurements can be used with most models to create lexica. We analyze variety of representative models, ranging from traditional context-oblivious models like support vector machines (SVM) and feed-forward neural networks (FFN) to modern context-sensitive models like variations of BERT.

SVM and FFN: Context-Oblivious Bag-of-Word Models SVM and FFN are used for Bag-of-Word models, as they are popular and representative choices for linear and non-linear models. SVM is a simple model which we use to explore how model complexity affects lexicon generation. FFN, as one of the simplest deep networks, also makes a good comparison with more complicated context-sensitive neural networks. In both models, we use context-oblivious FastText embeddings and generate lexica using STI.

LSTM: Context-Oblivious Model with Local Sequence Information We choose LSTM as a representative example of the models explained by inspection. Although LSTM is based on context-oblivious embeddings (FastText embeddings are context-free), it differs from the SVM and FFN as it takes advantage of the local sequence information. The inputs to the LSTM are sequence of FastText embeddings and retain their structural information.

A recent paper investigated the contradictory claims about the quality of attention as a feature importance method, and proposed techniques to improve the interpretability of the attention weights (Mohankumar et al., 2020). They report that high similarities among LSTM encoders across time impair the interpretability of the attention weights and that by reducing such similarities using the diversity LSTM they proposed, attention weights can be more interpretable. The diversity LSTM minimizes the conicity (similarity) of the hidden states while maximizing the log-likelihood of the training data.

We include the diversity LSTM from Mohankumar et al. (2020) in our comparison, as the authors showed it to be the most interpretable LSTM model. Following this prior work, We use the difference between the attention weights of a token in positively-labelled and negatively-labelled data as the metric to build the lexicon. To elaborate, in order to compute a score for a token, we compute an average attention weight for that token in all input data that are labelled positive and another for that token in all input data that are labelled negative. The reason for computing the two average scores is that attention weights do not have signs and do not distinguish between 'important to form a positive text' and 'important to form a negative text'. The difference of the two attention scores is then used as the final importance score.

DistilBERT and RoBERTa: Context-Sensitive Models BERT (Devlin et al., 2019) and variants like RoBERTa (Liu et al., 2019) produce state-of-the-art results on many natural language processing tasks, including sequence classification tasks in this paper. However, BERT models often have several hundred million parameters; it is controversial whether larger models necessarily lead to better performances on downstream tasks. Sanh et al. (2020) reach similar performances on many downstream tasks using much smaller language

models pretrained with knowledge distillation.

DistilBERT (Sanh et al., 2020), with a triple loss, shows that a 40% smaller network pretrained through distillation via the supervision of a bigger transformer-based language model can achieve similar performance on a variety of downstream tasks.

On the other hand, RoBERTa is a BERT model pretrained with careful choice of hyperparameters. With well-made design choices, RoBERTa improves the performance of original BERT (Devlin et al., 2019) and achieves similar performance as many state-of-the-art models published after BERT.

We use distilbert-base-uncased and roberta-base as pretrained models and fine-tune them on binary emotion or sentiment sequence classification tasks. We used the last layer of BERT, following the standard approach in the RoBERTa (Devlin et al., 2019). Then we adopt both masking and partionSHAP to generate lexica from the fine-tuned models.

5 Results and Discussion

Our aims were the following: 1) to understand how well different models and lexica perform in predictive tasks within the same domain as the training data and how well they generalize to datasets across different corpus domains and 2) to understand how humans rate the lexica.

5.1 Predictive Performance

We establish baselines for the models and lexica by assessing the models and lexica on a test set of the same dataset. Logistic regression is used as a calibration for both models and lexica, with prediction accuracy and F1 score as outputs. We conduct the same assessments for datasets across different corpus domains.

To elaborate, the logistic regression is used to find the threshold between binary predictions. Specifically, to use lexica for classification, the score for each sentence is obtained by averaging the lexical scores of words in that sentence. A logistic regression model takes these sentence scores as input and outputs binary classification results. Therefore, the regression model learns the sentence score distribution of the dataset we aim to evaluate on; thus, it serves as a calibration on the entire evaluation dataset. To make it a fair comparison between models and lexica, we did the same calibration using logistic regression models when evaluating model performance. In this case, we used the model outputs (logits) as the input of a logistic regression model and obtained the final predictions rather than directly using the model logits for classification.

The within-domain assessments use the test sets of those *lexicon-generating datasets*, datasets used to train the models and to create the lexica, while the across-dataset assessments use all data in the *evaluation datasets*, datasets used for evaluation only.

We compare the lexica generation methods by model and lexica predictive performance, measured by F1 scores, averaged over all lexicon-generating and evaluation datasets, as presented in Table 2 for both within-domain and across-domain performance. Furthermore, we conduct one-tail paired t-test to verify the significance of our observations (Appendix B).

	within-domain		**across-domain**	
Methods	**Model**	**Lexi.**	**Model**	**Lexi.**
Univariant		0.714		0.597
SVM_STI	0.791	**0.779**	0.68	**0.677**
FFN_STI	0.787	0.763	0.656	0.652
dLSTM[1]_Attn	**0.899**	0.756	0.652	0.604
DB[2]_Mask	0.825	0.761	0.749	0.641
DB[2]_SHAP	0.825	0.747	0.749	0.635
RB[3]_Mask	0.851	0.754	**0.767**	0.614
RB[3]_SHAP	0.851	0.774	**0.767**	0.646

Table 2: Mean predictive F-1 scores of models and lexica within and across corpus domain(s)

The model accuracy is in line with F1 scores and is included in Appendix A. Similar comparisons were also conducted for datasets with different types of labels separately, which are presented in the Appendix A. The comparison of the methods for datasets with different label types provided some insights on the stability of the methods when generating the lexica for different tasks.

Lexicon Generation Methods Comparisons Modern sequence models like LSTM, RoBERTa, and DistilBERT are larger and more complicated networks than a simple feed-forward neural network and certainly much more complex than SVMs. Thus, these models are expected to better fit the training data. With proper regularization, modern sequence models performed significantly better on data in the same corpus domain as the training data.

In terms of the generalizability, BERT-related models achieve better predictive performance on

data in different corpus domains than the datasets they are fine-tuned on (e.g., fine-tuning using Yelp comments and evaluating on song lyrics). On the other hand, LSTM does not generalize well outside of its training data domain. Unlike BERT-related models, LSTM models are limited to local sequence information.

The lexicon performance does not necessarily agree with the model performance. We can see from Table 2 that compared to the complex context-sensitive models, simple context-oblivious models can generate lexica with comparable if not better predictive performance within the same corpus domain as the training data. Meanwhile, the lexica generated from context-oblivious models also generalize better across other corpus domains than those from complex context-sensitive models.

This is reasonable since lexica are context-oblivious language models. When generating lexica, we lose the sequence information in the context-sensitive models. Although complex context-sensitive models generalize well to different domains, the lexica generated by them are not superior to those generated by simpler context-oblivious models.

It is worth noting that simpler models do not guarantee more generalizable lexica. Although the model complexity does not contribute to lexica generalizability, we still need a model sufficient to capture the correlations. In Table 2, it can be observed that the the lexica generated using SVM with STI measurement generalize significantly better that those generated using uni-variant correlation.

Meanwhile, different interpretation methods do not impact lexica generalizability as much as expected. SHAP yields better lexica than masking method for RoBERTa, but performs similarly as masking for DistilBERT.

5.2 Human Evaluation

Besides comparing generalization metrics from the model side, we also conduct the human evaluation for the created lexica. First, we split our lexica into two sets: one consists of words appearing only once in the training corpus, and the other includes the rest (words appearing at least five times). We group the words in both sets by seven different predictive labels: two sentiments (positive, negative) and five emotions (joy, fear, anger, sadness, and surprise).

To obtain words describing positive and negative sentiments, we select the top and bottom 100 words (words with the most positive and the most negative scores), respectively, from each lexica generated for sentiment classification tasks. The words describing emotions are drawn from each lexica generated for emotion classification tasks (top 100 words). Then we form multiple questionnaires for each one of seven labels.

Evaluators are required to choose from four categories for each word in the questionnaire (e.g., to evaluate the words in 'joy' lexica, four categories are *Describes Joy*, *Related to Joy*, *Not Related to Joy* and *Do Not Know*). Further details can be found in Appendix C.

We combine the responses of the questionnaires to determine whether a word is considered reasonable for the lexica. If 80% responses classify a word to either one of the first two categories, we say that it is considered a reasonable candidate for the lexica by human evaluators.

For each lexicon generation methods, we then report the proportion of the reasonable words averaged across sentiments and emotions, respectively. Table 3. The detailed results for each sentiment and emotion are presented in Appendix C.

Methods	Sentiment		Emotion	
	Once	**Freq**	**Once**	**Freq**
Univariant	7	32.9	2.2	13
SVM_STI	31.2	59.5	16.4	**22.6**
FFN_STI	**37.2**	**63.7**	**16.6**	22
dLSTM[1]_Attn	11.5	59.7	11.4	21
DB[2]_Mask	17.5	56.2	14.2	22.4
DB[2]_SHAP	10.2	35.5	4.8	15.2
RB[3]_Mask	12.2	35.4	9.4	19.6
RB[3]_SHAP	8.9	34.7	11	19.8

Table 3: Human evaluation results: percentage of words annotated as describing or related

It can be observed in Table 3 that the significantly more words, both rare ones and frequent ones, in lexica from context-oblivious models are considered reasonable by annotators than those in lexica from context-sensitive models. This is more obvious for sentiment tasks, where the amount of 'reasonable words' in lexica from context-oblivious models is almost twice as the amount in lexica from context-sensitive models.

Such good performance, however, cannot simply resort to the uncomplicated model structures since we also find that lexica generated by uni-variant method, the simplest one in all our methods, are usually not consistent with the human understanding.

By investigating the correlations between the lexica (Table 17 in Appendix B), we notice that context-oblivious methods generate similar lexica (with average correlation 0.88), while lexica generated by other methods differ from each other a lot (with average correlation ranging from 0.11 to 0.63), even for lexica generated by the same model using different interpretations or the ones generated using the same interpretation for the models of the same type. Such an observation is consistent with human evaluation results, where the lexica generated by context-oblivious models always have similar proportions.

6 Conclusion

Comparing lexicon generation methods, which are based on various models, interpreted by different feature importance measures, and tested on a large range of datasets, yields insights into what works better or worse for lexicon development and for model interpretation.

Context-sensitive models perform better than context-oblivious models within corpus domains and generalize better to other domains, but such an advantage is not observed for the predictive performance and generalizability of the produced lexica. The simpler context-oblivious models produce lexica that have similar or better predictive performance than those generated from more complex context-sensitive models, both within the corpus domain of the training data and across different domains.

Lexica are context-oblivious language models, so it is plausible that the sequence information learned by context-sensitive models is largely lost when generating the lexica, removing the advantage on across-domain generalizability that we observe for models.

Context-oblivious models do not only generate lexica that generalize better but also align better with human intuition. Human evaluation shows that much more words in lexica from context-oblivious models are considered reasonable than those in lexica from context-sensitive models and such observation is consistent for both rare and frequent words.

What is more, the lexica generated from different context-oblivious models are correlated, while lexica generated from different context-sensitive are quite different.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate.

Anil Bandhakavi, Nirmalie Wiratunga, Stewart Massie, and Deepak Padmanabhan. 2017. Lexicon generation for emotion detection from text. *IEEE intelligent systems*, 32(1):102–108.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the Model or True to the Data?

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 (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

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

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.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

David Garcia, Antonios Garas, and Frank Schweitzer. 2015. The language-dependent relationship between word happiness and frequency. *Proceedings of the National Academy of Sciences*, 112(23):E2983–E2983.

Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. 2017. Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65.

Sharath Chandra Guntuku, J. Russell Ramsay, Raina M. Merchant, and Lyle H. Ungar. 2019. Language of adhd in adults on social media. *Journal of Attention Disorders*, 23(12):1475–1485. PMID: 29115168.

Giannis Haralabopoulos and Elena Simperl. 2017. Crowdsourcing for beyond polarity sentiment analysis a pure emotion lexicon. *arXiv preprint arXiv:1710.04203*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020a. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167. Association for Computational Linguistics.

Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. 2020b. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167. Association for Computational Linguistics.

Ema Kušen, Giuseppe Cascavilla, Kathrin Figl, Mauro Conti, and Mark Strembeck. 2017. Identifying emotions in social media: comparison of word-emotion lexicons. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 132–137. IEEE.

Vivian Lai, Zheng Cai, and Chenhao Tan. 2019. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 486–495, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. Understanding Neural Networks through Representation Erasure.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 897–908, New York, NY, USA. Association for Computing Machinery.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards Transparent and Explainable Attention Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 435–440, San Diego, California. Association for Computational Linguistics.

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

James Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the dark triad of personality through twitter behavior. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 761–770.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded Lexicon Induction for Interpretable Social Science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1664–1673, Marseille, France. European Language Resources Association.

João Sedoc, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2017. Predicting emotional word ratings using distributional representations and signed clustering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 564–571, Valencia, Spain. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

PD Turney SM Mohammad. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

A Generalization Results for Sentiment and Emotion Classifications

Table 4 - 9

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.783	0.776
SVM_STI	0.855	0.853	**0.852**	**0.851**
FFN_STI	0.856	0.852	0.834	0.832
dLSTM[1]_Attn	0.881	0.879	0.837	0.825
DB[2]_Mask	0.9	0.9	0.841	0.838
DB[2]_SHAP	0.9	0.9	0.841	0.832
RB[3]_Mask	**0.918**	**0.919**	0.825	0.826
RB[3]_SHAP	**0.918**	**0.919**	0.847	0.841

Table 4: Within-domain performance of models and lexica for sentiment classification task

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.636	0.622
SVM_STI	0.716	0.714	**0.713**	**0.712**
FFN_STI	0.696	0.678	0.692	0.688
dLSTM[1]_Attn	0.688	0.671	0.673	0.638
DB[2]_Mask	0.788	0.784	0.682	0.671
DB[2]_SHAP	0.788	0.784	0.68	0.661
RB[3]_Mask	**0.809**	**0.808**	0.646	0.644
RB[3]_SHAP	**0.809**	**0.808**	0.686	0.673

Table 5: Across-domain performance of models and lexica for sentiment classification task

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.674	0.66
SVM_STI	0.734	0.733	0.716	**0.714**
FFN_STI	0.73	0.728	0.698	0.698
dLSTM[1]_Attn	**0.887**	**0.887**	0.702	0.695
DB[2]_Mask	0.759	0.76	0.71	0.694
DB[2]_SHAP	0.759	0.76	0.703	0.677
RB[3]_Mask	0.787	0.788	0.699	0.689
RB[3]_SHAP	0.787	0.788	**0.722**	**0.715**

Table 6: Within-domain performance of models and lexica for emotion classification task

B Statistical Comparison of Lexica

t-Test for Comparison between Models and Corresponding Lexica We conduct paired t-test on f-1 scores of models and lexica generated from them. We test on emotion tasks, sentiment tasks and all the tasks together. The null hypothesis is that the model has the same generalization performance with the lexicon. Results can be found in Table 10 - 12.

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.582	0.546
SVM_STI	0.622	0.616	0.615	**0.612**
FFN_STI	0.598	0.591	0.584	0.577
dLSTM[1]_Attn	0.611	0.605	0.576	0.537
DB[2]_Mask	0.68	0.672	0.614	0.578
DB[2]_SHAP	0.68	0.672	0.608	0.578
RB[3]_Mask	**0.684**	**0.683**	0.591	0.557
RB[3]_SHAP	**0.684**	**0.683**	**0.621**	0.601

Table 7: Across-domain performance of models and lexica for emotion classification task

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.726	0.714
SVM_STI	0.792	0.791	**0.781**	**0.779**
FFN_STI	0.79	0.787	0.764	0.763
dLSTM[1]_Attn	**0.899**	**0.899**	0.764	0.756
DB[2]_Mask	0.825	0.825	0.772	0.761
DB[2]_SHAP	0.825	0.825	0.766	0.747
RB[3]_Mask	0.85	0.851	0.759	0.754
RB[3]_SHAP	0.85	0.851	**0.78**	0.774

Table 8: Within-domain averaged performance of models and lexica over both sentiment and emotion classification tasks

t-Test for Model and Lexicon Comparison Separately We conduct paired t-tests on models and lexica's f-1 scores for different datasets separately. As former, we test on within-domain and across-domain datasets separately. Results are in the Table 13 - 16. The null hypothesis is the models or methods have the same generialization performance..

Pearson Correlation between lexica We calculate pearson correlation coefficient between every two lexica generated from different methods and put the results in Table 17.

C Human Evaluation

We ran out human evaluations of Amazon Mechanical Turk. Our HITs were in batches of 50 words with 10 attention checks per HIT. Five crowdworkers evaluated each HIT. The compensation for each HIT was $1.00 or $0.02 per word rated. The median time for each HIT depended on the task but was slightly less than 5 minutes. Figure 1 shows the first page of the HIT for positive sentiment.

Method	Model		Lexicon	
	Acc	**F1**	**Acc**	**F1**
Univariant			0.619	0.597
SVM_STI	0.683	0.68	**0.679**	**0.677**
FFN_STI	0.668	0.656	0.657	0.652
dLSTM[1]_Attn	0.664	0.652	0.641	0.604
DB[2]_Mask	0.753	0.749	0.66	0.641
DB[2]_SHAP	0.753	0.749	0.657	0.635
RB[3]_Mask	**0.767**	**0.767**	0.628	0.614
RB[3]_SHAP	**0.767**	**0.767**	0.663	0.646

Table 9: Across-domain averaged performance of models and lexica over both sentiment and emotion classification tasks

Methods	within-domain		across-domain	
	Acc	**F1**	**Acc**	**F1**
SVM_STI	**0.483**	**0.444**	**0.185**	**0.327**
FFN_STI	**0.065**	**0.089**	**0.305**	**0.173**
dLSTM[1]_Attn	0.026	0.019	0.007	0.001
DB[2]_Mask	0.016	0.014	5e-14	3e-11
DB[2]_SHAP	0.006	0.004	6e-13	2e-10
RB[3]_Mask	0.012	0.011	4e-17	1e-14
RB[3]_SHAP	0.005	0.003	5e-13	8e-11

Table 10: p-Values of paired t-tests for f-1 scores between models and lexica over sentiment classification tasks

Methods	within-domain		across-domain	
	Acc	**F1**	**Acc**	**F1**
SVM_STI	0.028	0.025	**0.084**	**0.307**
FFN_STI	**0.101**	**0.114**	**0.293**	**0.383**
dLSTM[1]_Attn	0.017	0.013	0.005	0.017
DB[2]_Mask	5e-4	0.003	7e-4	3e-4
DB[2]_SHAP	0.004	0.015	0.006	0.003
RB[3]_Mask	3e-4	7e-4	2e-5	8e-5
RB[3]_SHAP	7e-4	0.002	0.005	0.002

Table 11: p-Values of paired t-tests for f-1 scores between models and lexica over emotion classification tasks

Methods	within-domain		across-domain	
	Acc	**F1**	**Acc**	**F1**
SVM_STI	**0.051**	0.044	0.033	**0.142**
FFN_STI	0.031	0.040	**0.057**	**0.548**
dLSTM[1]_Attn	0.008	0.005	2e-4	8e-5
DB[2]_Mask	9e-5	1e-4	6e-14	4e-13
DB[2]_SHAP	6e-5	5e-4	2e-11	5e-11
RB[3]_Mask	2e-5	2e-5	2e-17	2e-16
RB[3]_SHAP	7e-6	1e-5	6e-12	7e-12

Table 12: p-Values of paired t-tests for f-1 scores between models and lexica over both sentiment and emotion classification tasks

	FFN	dLSTM[1]	DB[2]	RB[3]
SVM	0.549	**0.014**	**0.002**	**4e-5**
FFN		**0.017**	**0.003**	**7e-5**
dLSTM[1]			0.095	0.220
DB[2]				**7e-5**

Table 13: p-Values of paired t-tests for within-domain model f-1 scores

	FFN	dLSTM[1]	DB[2]	RB[3]
SVM	**0.005**	**0.012**	**2e-11**	**5e-12**
FFN		0.730	**9e-14**	**1e-11**
dLSTM[1]			**1e-10**	**1e-11**
DB[2]				**0.007**

Table 14: p-Values of paired t-tests for across-domain model f-1 scores

	SVM	FFN	dLSTM[1]_Attn	DB[2]_Mask	DB[2]_SHAP	RB[3]_Mask	RB[3]_SHAP
Univariant	**0.001**	**0.006**	**8e-4**	**0.004**	**0.033**	**0.006**	**6e-6**
SVM		**0.006**	**0.008**	0.065	0.064	**0.044**	0.550
FFN			0.364	0.857	0.344	0.363	0.199
dLSTM[1]_Attn				0.533	0.504	0.853	**0.003**
DB[2]_Mask					0.116	0.349	0.163
DB[2]_SHAP						0.579	0.052
RB[3]_Mask							**0.029**

Table 15: p-Values of paired t-tests for within-domain lexicon f-1 scores

	SVM	FFN	dLSTM[1]_Attn	DB[2]_Mask	DB[2]_SHAP	RB[3]_Mask	RB[3]_SHAP
Univariant	**3e-8**	**2e-4**	0.610	**4e-5**	**1e-8**	0.095	**2e-7**
SVM		**5e-4**	**2e-8**	**0.002**	**4e-4**	**2e-7**	**0.002**
FFN			**7e-5**	0.375	0.173	**0.005**	0.602
dLSTM[1]_Attn				**4e-4**	**0.006**	0.270	**2e-4**
DB[2]_Mask					0.311	**2e-5**	0.470
DB[2]_SHAP						**0.019**	0.067
RB[3]_Mask							**5e-5**

Table 16: p-Values of paired t-tests for across-domain lexicon f-1 scores

	SVM	FFN	dLSTM[1]_Attn	DB[2]_Mask	DB[2]_SHAP	RB[3]_Mask	RB[3]_SHAP
Univariant	0.27	0.30	0.45	0.13	0.42	0.12	0.37
SVM		**0.88**	0.26	0.22	0.21	0.18	0.24
FFN			0.27	0.21	0.21	0.17	0.23
dLSTM[1]_Attn				0.18	0.28	0.15	0.29
DB[2]_Mask					0.22	0.32	0.24
DB[2]_SHAP						0.11	0.63
RB[3]_Mask							0.33

Table 17: Pearson correlation between lexica from different methods

	Positive		**Negative**	
Methods	**One-time**	**Frequent**	**One-time**	**Frequent**
Univariant	5.7	46	8.3	19.7
SVM_STI	20	57.3	42.3	61.7
FFN_STI	**28.3**	**63.7**	**46**	**63.7**
dLSTM[1]_Attn	8.3	60.7	14.7	58.7
DB[2]_Mask	10.7	50.3	24.3	62
DB[2]_SHAP	9.7	30.3	10.7	40.7
RB[3]_Mask	8	22	16.3	48.7
RB[3]_SHAP	6.7	28	11	41.3

Table 18: Percentage of words classified as pos/neg description or related word from top 100 of lexica

	Joy		**Anger**		**Fear**		**Sadness**		**Surprise**	
Methods	**Once**	**Freq**	**Once**	**Freq**	**Once**	**Freq**	**Once**	**Freq**	**Once**	**Freq**
Univariant	6	19	0	13	3	14	1	13	1	6
SVM_STI	16	38	15	16	**35**	31	**8**	17	8	11
FFN_STI	**21**	**39**	**19**	15	28	28	6	17	**9**	11
dLSTM[1]_Attn	11	25	12	18	18	30	7	17	**9**	**15**
DB[2]_Mask	16	31	**19**	19	25	**33**	**8**	18	3	11
DB[2]_SHAP	12	15	6	**20**	2	18	2	14	2	9
RB[3]_Mask	18	25	3	14	14	28	**8**	**22**	4	9
RB[3]_SHAP	24	21	9	18	14	29	3	18	5	13

Table 19: Percentage of words classified as emotion description or related word from top 100 of lexica

Please Note

- You have to be an **English Native Speaker**
- You have to complete judgments for all sentences. **All fields are required**.

Instructions

Some words describe sentiment, which means a positive or negative emotion while other words relate to sentiment or emotion (eg, might cause it).

This task focuses on **positive** sentiment. For example, the word *fantastic* describes positive sentiment and the word *cake* relates to positive sentiment. In this task, you will be given a set of words. For each word, you will decide between the following choices:

a) the word describes positive sentiment
b) the word is related to positive sentiment (e.g. might cause it)
c) the word does not have any positive sentiment
d) don't know (e.g. you don't know the word)

	Positive sentiment	Related to Positive sentiment	Unrelated Word	Don't know
great	X			
skiing		X		
deadline			X	
further			X	
the			X	
alsike				X

Please confirm the following worker criteria:

☐ I have read the instructions
☐ I have read the examples
☐ I am a native English speaker
☐ I agree to be part of future research studies.

Positive Sentiment Rating

Figure 1: Image of the Amazon Mechanical Turk HIT