Inducing Generalizable and Interpretable Lexica

Yilin Geng*
University of Pennsylvania
link10@seas.upenn.edu

Zetian Wu*
Johns Hopkins University
zwu49@jhu.edu

Roshan Santhosh
University of Pennsylvania
roshansk@seas.upenn.edu

Tejas Srivastava
University of Pennsylvania
tjss@seas.upenn.edu

Lyle Ungar
University of Pennsylvania
ungar@cis.upenn.edu

João Sedoc
New York University
jsedoc@stern.nyu.edu

Abstract

Lexica – words and associated scores – are widely used as simple, interpretable, generalizable language features to predict sentiment, emotions, mental health, and personality. They also provide insight into the psychological features behind those moods and traits. Such lexica, historically created by human experts, are valuable to linguists, psychologists, and social scientists, but they take years of refinement and have limited coverage. In this paper, we investigate how the lexica that provide psycholinguistic insights could be computationally induced and how they should be assessed. We identify generalizability and interpretability as two essential properties of such lexica. We induce lexica using both context-oblivious and context-aware approaches, compare their predictive performance both within the training corpus and across various corpora, and evaluate their quality using crowd-worker assessment. We find that lexica induced from context-oblivious models are more generalizable and interpretable than those from more accurate context-aware transformer models. In addition, lexicon scores can identify explanatory words more reliably than a high performing transformer with feature-importance measures like SHAP.[1]

1 Introduction

Lexica – collections of words, often with associated weights – are widely used for interpretable models (Hayati et al., 2021; Pryzant et al., 2018), particularly in psychology (Boyd et al., 2022) and other social sciences. Lexica have been developed for areas as varied as sentiment and emotion (De Bruyne et al., 2022; Hamilton et al., 2016), moral foundations (Hopp et al., 2021), politeness (Li et al., 2020a), formality (Eder et al., 2021), concreteness and familiarity (Paetzold and Specia, 2016), and

	Generalizability	Interpretability
Lexica Vs. Lexica	**RQ1**	**RQ3**
Lexica Vs. Model	**RQ2**	**RQ4**

Figure 1: The relations between the proposed research questions

bilingual research (Shi et al., 2021; Patra et al., 2019). They are being created in hundreds of languages (Zhao and Schütze, 2019) and are increasingly used to augment modern deep learning models (Li et al., 2020b; Hu et al., 2019). Both supervised (Irvine and Callison-Burch, 2013) and unsupervised (Artetxe et al., 2019; Zhang et al., 2017; Kanayama and Nasukawa, 2012) methods have been proposed, some with an emphasis on supporting interpretation (Verhoeven and Daelemans, 2018; Clos and Wiratunga, 2017; Misra et al., 2015).

Some most widely used lexica were created by human experts (Pennebaker et al., 2001; Mohammad, 2018). However, these high-quality lexica often take years of refinement and have limited coverage. In comparison, computationally induced lexica are cheaper and lead to visible new insights provided by machine learning models for various corpora.

In computer science, closely related to lexicon development is "feature importance", which also computes a strength of association between words and an outcome of interest to support interpretation. Many methods have been used to extract feature importance from neural networks and other machine-learned models (Ribeiro et al., 2016; Kim

*Indicates equal contribution

[1]Code and induced lexica are available at https://github.com/wwbp/embedding-lexica-creation.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4430–4448
December 7-11, 2022

et al., 2020). One of the most popular of these measures is SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017), a mathematically principled way of computing feature importances based on Shapley values from game theory.

Different feature importances may serve different purposes, including “explaining the model" (i.e., showing why the model makes given predictions), versus “explaining the world" (i.e., providing insight into the data on which the model was trained and the world where that data was collected) (Chen et al., 2020; Liu and Ungar, 2021). For example, when extracting feature importance from the sentence, “The food was pretty and tasty!", an attention-based 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 “pretty" and “tasty".

These different feature importance measurements provide different interpretations for the same prediction of the same model on the same input. The interpretations that “explain the model" are, in general, more “faithful" to the model, reflecting how the model uses each feature, while the ones that “explain the world" are more consistent with human intuition and reflect some consensus in the world.

The two goals are not contradictory, but they have different priorities. In computer science research, feature importances are more often used to explain models. In contrast, social scientists such as psychologists use more expert-annotated lexica designed to explain the world. Our goal is to computationally build lexica that explain the world, with the help of feature importance measurements. Thus, our desirable lexica should solely be evaluated in terms of their faithfulness to the models. Our primary goal is to provide insights into scientific questions using the lexicon analysis (e.g., "how are political parties getting more polarized?" or "when is empathy good or bad for people?").

Instead of the faithfulness to the models, we identify generalizability and interpretability as key properties to assess the desirability of such lexica. Generalizability is crucial to high-quality lexica. For example, widely used lexica, such as LIWC (Pennebaker et al., 2001), works well in an extremely broad set of corpora (used in over 10,000 papers). If a lexicon has the ability to explain emotion/sentiment in the real world, it should generalize well from one corpus (e.g., food reviews on Yelp) to another (e.g., music lyrics). Interpretability is even more important. The words in the lexicon should reflect what humans view as being important for explaining the emotion, personality, political orientation or other labels being predicted.

To compare the degree of generalizability and interpretability of the lexica induced from context-aware or context-oblivious models and to gain insights into the lexica induction and assessment, we address the following four research questions (Figure 1):

- **RQ1:** How well do lexica made from context-aware or context-oblivious models generalize to different corpora?
- **RQ2:** How much predictive power do lexica lose relative to deep learning models?
- **RQ3:** How sensible do human raters view the words in lexica induced by context-aware and context-oblivious approaches?
- **RQ4:** How explainable are lexicon scores compared to feature importance measures from predictive models?

2 Related Work and Research Goals

Lexicon creation was traditionally done manually. In psychology, lexica such as LIWC were created based on judgments of expert annotators (Pennebaker et al., 2001). LIWC is unweighted, and can be viewed as having a weight of one for all words in the lexicon. Weighted lexica have also been created using crowdsourced annotations (Mohammad, 2018).

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-level 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 the coefficients from models or evaluate the impact of the features on the outputs by perturbing the inputs (Lundberg and Lee, 2017; Ribeiro et al., 2016).

For linear models, lexica can be constructed by directly using the coefficients or weights in the models. Similarly, for non-linear models, people attribute to features by examining gradients, which can also be used to induce lexica (Simonyan et al., 2013; Baehrens et al., 2010). Moreover, attention weights in more complex neural networks can serve the same function (Bahdanau 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 proper lexicon weights or faithful explanations (Jain and Wallace, 2019).

With the introduction of transformers (Vaswani et al., 2017), more complex context-aware models such as BERT (Devlin et al., 2019) (and variations such RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019)) often provide significantly better predictive performance. However, these developments present a larger challenge to interpret these more sophisticated models. On one hand, Sundararajan et al. (2017) proposed Integrated Gradients (IG) for these differentiable models that examines the path integral of the gradients based on input baselines. On the other hand, we can also interpret the model as a black box, without the access to the gradients. By observing the impact on the predictions of some carefully designed perturbations for each word (e.g., via removal or masking) in the input, we can compute the importance of each work to the prediction (Li et al., 2016; Kim et al., 2020).

SHAP is an important example of such input perturbations (Lundberg and Lee, 2017). Based on the Shapley Value from game theory, SHAP provides a class of approximations that evaluate the contributions of features in machine-learned models. Partition SHAP is one of these approximations, that computes the Shapley value for clustered features based on the partition trees, which provides a contextualized understanding of the input.

Many feature importance methods, such as marginal Shapley values, are designed to “explain the model". The induced lexica thus contain words that help explain what models are computing, which are not necessarily the words that are important for understanding the world – the sentences and the people who produce them. For example, attention weights may focus on the word “and", rather than adjacent words. We prefer feature importance such as conditional Shapley values that seek to “explain the world"; similarly, psychologists are also interested in lexica that “explain the world" to answer the 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 has been no broad assessment of the ability of induced lexica to “explain the world". Lai et al. (2019) compare feature importances across different models and feature importance metrics. However, the comparisons are based on the similarities of the most important features considered, and there is no metric to assess the quality of the lexica. Ding and Koehn (2021) provide an evaluation for the prediction interpretations in terms of plausibility and faithfulness. Although a good lexicon should support plausible interpretations, we are more interested here in how a plausible lexicon (independent of a particular given model prediction) can be induced and assessed, and thus address a different task.

Bearing in mind how social scientists actually use lexica, we focus on evaluating the generalizability and interpretability of the lexica induced by different automated approaches. Lai et al. (2019) shows that some models provide similar explanations of the predictions regardless of the feature-importance metrics used. We, therefore, choose a set of popular models with differing levels of complexity and different accessibility to context, along with the suitable interpretations for each, and test them on diverse sentiment and emotion corpora. We work with sentiment and emotion since they are well-studied domains, allowing us to focus on the lexica induction insights.

3 Datasets

Our experiments induce lexica using a mixture of common broad-coverage datasets such as Yelp[2], Amazon reviews (McAuley and Leskovec, 2013), and NRC Emotion (Mohammad and Turney, 2013). We use relatively tailored datasets such as Sentiment Treebank (Socher et al., 2013), EmoBank (Buechel and Hahn, 2017), Emotionlines (Hsu et al., 2018), Daily Dialog (Li et al., 2017), and

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

Song Lyrics (Mihalcea and Strapparava, 2012), for evaluations of the lexica (in order to evaluate their generalizability to remote corpora). We will refer to the datasets used for lexica induction and evaluation as the "lexicon-induction datasets", and the ones used only for evaluation as the "evaluation datasets". For large datasets, we use their balanced subsets.

The chosen datasets are from diverse sources, including Twitter, song lyrics, newswire, online reviews, and crowdsourced writing. They vary by size, sentence length, and vocabulary size (for detailed dataset statistics see Table 3 in Appendix A). This variety of datasets ensures robust comparisons between the lexica induction approaches.

Labels of all datasets are processed to 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 positive or negative. For these datasets, we are interested in both the "heads" and "tails" (the words with the highest and lowest scores) of the resulting lexica, as they indicate positivity and negativity, respectively. For the NRC dataset, models do binary emotion 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 fair comparisons, this work is done entirely in English; non-English words in the NRC datasets are filtered and removed.

4 Lexicon Induction Approaches

The core of lexicon induction is the assignment of scores to each word, reflecting its semantics; we do this using the relative importance of the words in contributing to the label prediction. This requires deciding which predictive model and which feature importance measure to use.

We explore different combinations of predictive models and means of computing feature importance as different approaches to create lexica. The models are trained to do text classification, and we select a set of sentiment and emotion tasks that are widely studied in order to yield the most insights. Although models like BERT use subwords as tokens, we compute only the word-level scores when inducing the lexica so that the lexica generated by different methods are comparable and the lexica are interpretable.

These approaches are categorized based on the models' access to the context of the input text: context-oblivious approaches in which the sequence information and context in the input are lost (SVM, FFN), versus context-aware approaches in which the sequence information is embedded in the representations and used for classification (LSTM, RoBERTa, and DistilBERT). The motivation for such categorization is that it remains unclear whether context would facilitate the creation of more generalizable and interpretable lexica (RQ1 and RQ3).

4.1 Context-oblivious Approaches

4.1.1 Frequency-based Baseline

The most intuitive way to score the words based on the classification datasets is to use the word frequency. Specifically, in what we called "univariate method", for each 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., binary scores, as the word's score for the lexicon. We have also tried another frequency-based baseline that combines tf-idf (term frequency-inverse document frequency) with logistic regression, and we picked the best of the two.

4.1.2 Bag-of-Vector Models with Single-token Importance (STI)

Bag-of-Vector Models (SVM and FFN) SVM and FFN are used as Bag-of-Vectors models, since they are popular and representative choices for linear and non-linear models with low model complexity. The inputs to both models are text embeddings, computed as the averaged FastText embedding for all the tokens in the text. As a result, they lose all the sequential information in the inputs, which makes them context-oblivious.

Single-token Importance (STI) Since the inputs to the models, text embeddings, are averaged token embeddings, they lie in the same embedding space as tokens. We can thus compute feature importance for individual 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.

4.2 Context-aware Approaches

4.2.1 LSTM with Attention

We choose LSTM as a representative example of the models explained by inspection. The inputs to the LSTM are sequences of fixed FastText embeddings, and model attention serves as the importance measurement.

Attention Weights as Explanations Attention has been used for model interpretation, with the belief that the attention weights indicate the relative importance of the tokens. However, it is still controversial whether attention is actually explanatory. 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 has a trivial function, since a random permutation of the attention coefficients does not significantly affect the predictions (Vashishth et al., 2019).

Diversity LSTM A recent paper investigated the contradictory claims about the quality of attention as a feature-importance measurement, and proposed techniques to improve the interpretability of the attention weights (Mohankumar et al., 2020). They reported 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 could 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 they claimed that it was the most interpretable LSTM model. Following this prior work, we use the difference between the attention weights of a token in positively-labeled and negatively-labeled 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 labeled positive and another for that token in all input data that are labeled 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 between the two attention scores is then used as the final score for the token.

4.2.2 BERT Variations with Masking and SHAP

BERT Variations (RoBERTa and DistilBERT) As stated, lexica creation is a task based on language understanding. Modern language models like BERT (Devlin et al., 2019) produce state-of-the-art results on many downstream NLP tasks, including the sequence classification tasks in this paper, and thus are believed to be able to capture the semantics. As a result, we included two variations of BERT with different network sizes in the comparison, namely RoBERTa and DistilBERT.

We use pretrained "distilbert-base-uncased" and "roberta-base" from HuggingFace library (Wolf et al., 2020) and fine-tuned them on binary emotion or sentiment text classification tasks. We used the last layers of models, following the standard approach for these models.

Feature importance in these complex models can hardly be interpreted by inspection. Here, we applied two model-agnostic methods.

Masking The importance of a token can be measured by the change in the model output when the token is replaced 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., 2016).

Partition SHAP SHAP 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 explained before, we believe that the SHAP that takes account of the correlation between words in each sentence is better at explaining the world. Partition SHAP is a variation of SHAP that uses a hierarchical clustering of the features (Lundberg and Lee, 2017). As a result, it is essentially computing the Owen values from game theory, where the partition of the players is considered (Owen, 1977). Partition SHAP assumes independence between sets of features instead of individual ones. The feature clustering can be done based on correlations, or any other distance metric, or even predefined rules (e.g., tokens in a cluster must be adjacent). Partition SHAP attributes to the clusters instead of individual features in the clusters. It is also much faster than other model-agnostic SHAP methods, such as kernel SHAP (Lundberg and Lee, 2017), since the complexity of partition SHAP is quadratic in the number of input features while the other methods are exponential in theory.

5 Evaluations and Results

The induced lexica are evaluated in terms of generalizability and interpretability, to address the four proposed research questions in Section 1. Examples of the induced lexica can be found in Appendix C.

5.1 Generalizability

We use predictive performance on within-corpus test sets and across-corpora evaluation sets as an indication of the generalizability of the induced lexica. The comparisons are made from two perspectives: Firstly, we compare lexica induced by different approaches against each other. This provides insight into the lexicon induction approach, such as how the sequence information helps to induce more generalizable lexica. Secondly, we assess how lexica, as simple linear classifiers, perform in predictive tasks compared to the sophisticated vector-embedding models.

To use lexica for predictive tasks, we rely on the lexicon scores to construct linear classifiers. Each lexicon-based classifier is a logistic regression model that classifies input sentences based on sentence scores, trained on a small subset that has the same distribution as the evaluation set. The sentence score is the average score for the lexical words in that sentence. In other words, the regression model learns the sentence score distribution of the evaluation set; thus, it serves as a calibration on a specific evaluation corpus. To make it a fair comparison between models and lexica, we do the exact calibrations using logistic regression models when evaluating model performances. In this case, we use the model outputs (logits) as the input of a logistic regression model and use the output of the regression model as the final prediction, rather than directly using the model logits for classification. The calibrations use small subsets separated from the evaluation sets, and the data in the calibration subsets is not seen in training or evaluation.

The predictive performance is presented in Appendix D as F1 scores averaged over all "evaluation datasets" and test sets of "lexicon-induction datasets". The model accuracy is in line with F1 scores and is included in Appendix D. One-tail paired t-tests are conducted to verify the significance of our observations (Appendix E). Similar comparisons are also conducted for emotion corpus and sentiment corpus separately, which are presented in the Appendix D. These comparisons confirm the stability of the observations when inducing lexica from different classification tasks and corpora.

Methods	Within-corpus		Across-corpora	
	Model	Lexi.	Model	Lexi.
Univariate		0.714		0.598
SVM_STI	0.791	**0.779**	0.687	**0.684**
FFN_STI	0.787	0.763	0.657	0.654
dLSTM[3]_Attn	**0.899**	0.756	0.654	0.609
DB[4]_Mask	0.825	0.761	0.755	0.650
DB[4]_SHAP		0.758		0.641
RB[5]_Mask	0.851	0.754	**0.768**	0.617
RB[5]_SHAP		0.774		0.649

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

5.1.1 Lexica Generalizability

Table 1 rows compare lexica induced by the various lexicon induction approaches introduced in Section 4.

As expected, the lexica induction approaches that are based on vector embeddings have observable advantages in the predictive performance compared to the frequency-based baseline (Table 1).

When it comes to the impact of the context-awareness or the sequence information, it is notable that context-oblivious bag-of-vector approaches with much simpler models produce comparable if not better lexica in terms of generalizability than the context-aware ones (Table 1). This indicates that the context and the model complexity do not contribute much to the lexica generalizability.

Meanwhile, the choice of interpretations does not have a consistent impact on the induced lexica generalizability. For example, the SHAP method yields more generalizable lexica than the masking method for RoBERTa, but performs similarly to the masking method for DistilBERT.

[3]diversityLSTM
[4]DistilBERT
[5]RoBERTa

5.1.2 Lexicon vs. Model: the use of lexica in predictive tasks

We compare the predictive performance of lexica and the predictive models by inspecting respectively the within-corpus and cross-corpora results in Table 1. For context-oblivious models, we find that the induced lexica, which are only linear classifiers, have negligible performance drop in predictive tasks compared to the model. As for the context-aware models, lexica always have worse performance than the models.

We can also synthetically compare the generalizability of lexica and models. Context-aware models perform better than context-oblivious ones within the training corpus as expected, and they also generalize better to other corpora. However, we do not see such a generalization advantage for the lexica induced using context-aware models, as we show in Section 5.1.1. This suggests that although the context contributes to model generalizability, it does not contribute to lexica generalizability.

There is a consistent reason for the performance drop and loss of generalization advantage observed for lexica induced using context-aware models: lexica themselves are context-oblivious. When generating lexica, we lose the sequence information learned by the context-aware models. As a result, although complex context-aware models generalize well to different domains, the lexica generated by them are not superior to those generated by simpler context-oblivious models.

5.2 Interpretability

The induced lexica are evaluated both as sets of words (without context) and as words within sentences (with context).

To measure the impact of the context-awareness on lexicon induction, the lexica induced using context-aware and context-oblivious approaches are presented to the annotators as sets of words without context. To evaluate how lexicon scores are explainable, we assess the ability of lexicon scores to highlight the explanatory words in the sentences, by comparing that to the capability of the best-performing model with different feature-importance measurements.

5.2.1 Lexica Interpretability

We split our lexica into two sets: one consists of words appearing only once in the training corpus, and the other includes the words appearing at least five times. We then 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 sentiment, we select the top and bottom 100 words (words with the most positive and the most negative scores), respectively, from each lexicon induced from sentiment classification tasks. For emotion classification tasks, only the top 100 words are drawn. We form multiple questionnaires for each one of the seven labels. An example of the questionnaires can be found in Appendix F.

Evaluators are required to choose from four categories for each word in the questionnaires (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 F.

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

In Table 2, we report the proportion of the reasonable words averaged across all sentiments and emotions for each lexicon induction approach. The detailed results for sentiment and emotion tasks are presented in Appendix F.

Methods	Sentiment Once	Sentiment Freq	Emotion Once	Emotion Freq
Univariate	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[3]_Attn	11.5	59.7	11.4	21
DB[4]_Mask	17.5	56.2	14.2	22.4
DB[4]_SHAP	11.2	35.4	7.8	18.4
RB[5]_Mask	12.2	35.4	9.4	19.6
RB[5]_SHAP	15.2	35.8	12	**23.2**

Table 2: Lexica interpretability human evaluation results: percentage of words annotated as “the word describes the **[sentiment/emotion]**" or “the word is related to the **[sentiment/emotion]**" averaged across all corpora for each method

Significantly more words, both rare ones and frequent ones, in lexica induced using context-oblivious approaches, are considered more reasonable by annotators than those in lexica induced using context-aware approaches (Table 2). This observation is especially evident for lexica induced from sentiment tasks, for which, lexica from context-

RoBERTa + Masking:	Great shop with lots of ideas . Prices are very reasonable.	1(a)
RoBERTa + PartitionSHAP:	Great shop with lots of ideas. Prices are very reasonable.	1(b)
Lexica Scores:	Great shop with lots of ideas. Prices are very reasonable.	1(c)
RoBERTa + Masking:	Worst experience I had in a restaurant. The burger came little burnt and the waiter was very rude.	2(a)
RoBERTa + PartitionSHAP:	Worst experience I had in a restaurant. The burger came little burnt and the waiter was very rude.	2(b)
Lexica Scores:	Worst experience I had in a restaurant. The burger came little burnt and the waiter was very rude.	2(c)
RoBERTa + Masking:	The movie takes such a speedy swan dive from excellent to interesting to familiar before landing squarely on stupid.	3(a)
RoBERTa + PartitionSHAP:	The movie takes such a speedy swan dive from excellent to interesting to familiar before landing squarely on stupid.	3(b)
Lexica Scores:	The movie takes such a speedy swan dive from excellent to interesting to familiar before landing squarely on stupid.	3(c)
RoBERTa + Masking:	They are anything but fabulous . Very disappointing experience.	4(a)
RoBERTa + PartitionSHAP:	They are anything but fabulous . Very disappointing experience.	4(b)
Lexica Scores:	They are anything but fabulous . Very disappointing experience.	4(c)

Figure 2: Comparisons between highlighting explanatory words in sentences using "lexicon scores induced by FFN from Yelp dataset" and "RoBERTa finetuned on Yelp dataset with different feature importance scores" (red for positive sentiment, blue for negative sentiment)

oblivious model contain double the number of "reasonable words" as lexica from context-aware models. Such good performance, however, cannot be simply due to the naive model structures, since lexica generated by the frequency-based baseline are not considered similar in quality.

Although lexica induced using different feature importance measures for the same BERT models yield similar generalization accuracy, they are, in fact, very different. For example, the lexica induced using masking from DistilBERT models have significantly better performance in interpretability compared to the lexica induced using SHAP from the same models.

Finally, human evaluation interpretability results remain consistent when investigating the correlations between the lexica (Table 19 in Appendix E). We notice that context-oblivious approaches induce similar lexica (with an average correlation of 0.88), while lexica induced using context-aware approaches differ substantially from each other (with average correlations ranging from 0.11 to 0.63).

5.2.2 Lexicon vs. Model: the use of lexica to support interpretation

To interpret sentiment and emotion in text, people often use predictive models with some feature importance measurements. Alternatively, one can inspect the lexicon scores associated with those words in the text, and use that as an interpretation. We take the lexica induced using FFN (the best-rated lexica from the crowdsourced evaluation) and compare it to RoBERTa (the best performing model in predictive tasks) with various feature importance measurements as interpretations for text instances.

To evaluate and compare these interpretations, we highlight the words with the highest and lowest scores in a set of texts considered by each method (e.g., instances in Figure 2). And to assure comparability, thresholds are selected so that all methods highlight a similar number of words across the corpus. On any given sentence, the number of highlighted words is thus allowed to vary across methods.

From Figure 2, we can observe that masking is not a reliable interpretation method for the RoBERTa model. It often attributes importance to neutral background words and punctuation. SHAP performs better than masking, but it tends to attribute importance to neutral words adjacent to the positive/negative words. (as in Figure 2 [1(b), 2(b), 4(b)]) and sometimes to punctuation (as in Figure 2 [4(b)]).

Lexicon scores of neutral words and punctuation are reliable and stable. Plus, they are more sensitive to the change of the positivity of the adjectives than SHAP (as shown by comparing Figure 2 [3(c), 3(b)]).

Lexica are oblivious to context, and thus cannot identify negativity in expressions such as "anything but fabulous" (as in Figure 2 [4(c)]). However, it is controversial whether generally positive words still have positive meanings when they are used in a negative context. For example, does "fabulous" still carry positive meaning when it is used in "anything but fabulous"?

6 Conclusion

Comparing lexicon induction approaches based on various models – interpreted by different feature importance measures, and tested on various corpora – yields insights into what works best for inducing lexica and supporting interpretation.

We observe that context improves model generalizability, but not lexicon generalizability. The simpler context-oblivious models produce lexica with better generalizability: better predictive performances both within the training corpus and across different corpora. Lexica induced using context-aware models lose the superiority in across-domain generalizability of context-aware models.

When we induce lexica from the context-aware models, we lose the sequence information learned, as the lexica themselves are context-oblivious. That also leads to a surprising finding that, for context-oblivious models, linear classifiers using lexica scores do not show much performance drop compared to more complex models.

Lexica generated from context-oblivious models not only generalize better, but also align closer with human intuition. Human evaluation shows that more words in lexica induced using context-oblivious models are considered reasonable than in lexica induced using context-aware models, regardless of whether the words are rare or frequent.

We also find that the lexica generated from different context-oblivious models are correlated, while lexica generated from different context-aware models vary more.

Furthermore, lexicon scores can more reliably identify explanatory words in texts than feature-importance measures applied to transformer models.

7 Future Work

Lexica used in computational social science range from ad hoc sets of words selected by a single investigator to carefully crafted and validated word collections. Future work should compare the quality of computer-generated lexica such as the ones included in this paper against this range of human-constructed lexica.

We also found that feature importances of words in context are highly unstable, and that such instability can be observed across various models and feature importance measures. Future work should investigate the scale of the instability and the reasons for it.

8 Limitations

This work identified two desirable properties of lexica, generalizability and interpretability, and evaluated lexica induced using various approaches in terms of these two properties. However, to make the most obvious comparisons, this work induced lexica only using well-established sentiment classification tasks as representative examples. Lexicon induction from other tasks should be explored to ensure that the results are globally consistent.

In this paper, we found that better-performing context-aware models generate worse lexica. This work only tested existing feature importance measurements; future work can search for improved interpretation methods for context-aware models.

Finally, we only looked at English corpora. It remains to be verified that these results generalize across languages.

Acknowledgment

We would like to thank our anonymous reviewers for their helpful and constructive suggestions.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

R Boyd, Ashwini Ashokkumar, Sarah Seraj, and J Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*.

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? *arXiv preprint arXiv:2006.16234*.

Jérémie Clos and Nirmalie Wiratunga. 2017. *Lexicon Induction for Interpretable Text Classification*, pages 498–510. Springer International Publishing.

Luna De Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2022. Joint emotion label space modeling for affect lexica. *Computer Speech & Language*, 71:101257.

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.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5034–5052, Online. Association for Computational Linguistics.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2021. Acquiring a formality-informed lexical resource for style analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2028–2041, Online. Association for Computational Linguistics.

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.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53(1):232–246.

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).

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–523, Atlanta, Georgia. Association for Computational Linguistics.

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.

Hiroshi Kanayama and Tetsuya Nasukawa. 2012. Unsupervised lexicon induction for clause-level detection of evaluations. *Natural Language Engineering*, 18(1):83–107.

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

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. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020a. Studying politeness across cultures using english twitter and mandarin weibo. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–15.

Wei Li, Luyao Zhu, Yong Shi, Kun Guo, and Erik Cambria. 2020b. User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models. *Applied Soft Computing*, 94:106435.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. 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.

Tony Liu and Lyle Ungar. 2021. Towards cotenable and causal shapley feature explanations. *AAAI 2021 Workshop: Trustworthy AI for Healthcare*.

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. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Julian J. McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. *CoRR*, abs/1303.4402.

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.

Dipendra Kumar Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. 2015. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 992–1002, Beijing, China. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. 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.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

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, Online. Association for Computational Linguistics.

Guilliermo Owen. 1977. Values of games with a priori unions. In *Mathematical economics and game theory*, pages 76–88. Springer.

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.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

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

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, New Orleans, Louisiana. 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. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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.

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.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 813–826, Online. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *CoRR*, abs/1909.11218.

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.

Ben Verhoeven and Walter Daelemans. 2018. Discourse lexicon induction for multiple languages and its use for gender profiling. *Digital Scholarship in the Humanities*, 34(1):208–220.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2019. A multilingual BPE embedding space for universal sentiment lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3506–3517, Florence, Italy. Association for Computational Linguistics.

A Dataset Information

Information on the datasets used in experiments of this paper can be found in Table 3.

B Model Information

Information on the model architectures and specific settings used in this paper can be found in Table 4.

We use AdamW as the optimizer. The learning rate is 1e-4 for FFN and 1e-5 for all other neural network models. We conduct an early stop strategy which monitors the change of accuracy to determine whether to stop training or not, and the patience is 7.

C Lexica Examples

The examples of our induced lexica are presented in Table 5.

D Generalization Results for Sentiment and Emotion Classifications

Detailed generalization results can be found in Table 6 - Table 11.

E Supportive Statistical Analysis

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 12 - Table 14.

t-Test for Inter-Model and Inter-Lexicon Comparisons We conduct paired t-tests on f-1 scores for model pairs and for lexicon pairs. As above, we test on within-domain and across-domain datasets separately. The results for models are in Table 15 and Table 16. The results for lexica are in Table 17 and Table 18. The null hypothesis is the models or methods have the same generalization performance.

Pearson Correlation between Lexica We calculate the averaged Pearson correlation coefficient for lexica induced by every pair of methods, and present the numbers in Table 19.

F Human Evaluation Details

We run our human evaluations on Amazon Mechanical Turk. Our HITs are in batches of 50 words, with 10 attention checks per HIT. Each HIT is evaluated by 5 workers. The compensation for each HIT was $1.00 or $0.02 per word rated. The median time for each HIT depends on the task, but is slightly less than 5 minutes. Figure 3 shows the first page of the HIT for positive sentiment.

For all HITs, we remove invalid responses based on their performance on attention check words, i.e., if one response makes more than 2 mistakes on check words, it is considered invalid and will thus be filtered. We also mark the WorkerID of invalid results to avoid them partaking other HITs.

We calculate the average Cohen's kappa coefficient of HITs to evaluate inter-rater reliability. The values for different tasks, e.g., positive/negative, joy/sadness, etc., are between 0.431 and 0.576, which show a sensible consistency among different workers.

Lastly, we calculate the proportion of words considered reasonable in all the induced lexica. The results for sentiment and emotions are in Table 20 and Table 21 respectively.

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 (McAuley and Leskovec, 2013)		17666/2094	2158	125.9
NRC (Mohammad and Turney, 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., 2017)	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
SST2 (Socher et al., 2013)			872	20.2

Table 3: Details on the datasets used for training and evaluation.

Model	Architecture	Input	Output
SVM	Linear SVM Regularization C = 25	300	1
FFN	Linear: 300*1024 1024*512 512*128 128*2 Activation: Relu	300	2
diversity LSTM	The same as in Mohankumar et al., 2020, with attention weights outputted	128*300	2
DistilBERT	The same as in Sanh et al., 2019	128*768	2
RoBERTa	The same as in Liu et al., 2019	128*768	2

Table 4: Details on the model architectures used to induce lexica.

Yelp_FFN_Negative		Yelp_FFN_Positive		Yelp_DistilBERT_Mask_Negative		Yelp_DistilBERT_Mask_Positive	
Word	**Score**	**Word**	**Score**	**Word**	**Score**	**Word**	**Score**
discusting	-25.336	bookmarked	23.126	diminished	-1.45103	enticed	0.49419
uninviting	-25.203	expertly	20.613	saddest	-1.23786	famished	0.48921
unprofessionalism	-24.825	terrific	20.422	butchered	-1.10995	magically	0.45964
undrinkable	-24.606	invaluable	19.827	weirdest	-0.92964	harried	0.41756
unedible	-24.191	bookmark	19.760	marginal	-0.83127	brilliant	0.41124
unprofessional	-24.010	thorough	19.671	slowest	-0.78917	vines	0.39751
unwelcoming	-23.636	adore	19.669	absence	-0.74505	overcharging	0.39042
unappetizing	-23.060	mouthwatering	19.662	patchy	-0.70485	traditionally	0.38383
tastless	-22.990	fabulous	19.336	embarrassment	-0.68027	triangles	0.38312
unsanitary	-22.861	cutest	18.936	lacks	-0.67126	blessed	0.36406
inedible	-22.086	fantastic	18.791	poorest	-0.66787	souvent	0.36043
scammed	-21.987	superb	18.707	won't	-0.65272	hooray	0.35292
undercooked	-21.934	unbeatable	18.588	hated	-0.64843	gimmick	0.35001
underseasoned	-21.803	marvelous	18.347	disgraceful	-0.61634	tornado	0.35000
tasteless	-21.778	wonderful	18.303	lacking	-0.57571	takeaway	0.34061
disgusting	-21.667	sweetest	18.224	smattering	-0.57051	excellently	0.33728
degraded	-21.387	amazing	18.050	unwilling	-0.56928	compelling	0.33546
disrespected	-21.134	tremendous	17.969	disregard	-0.56529	godsend	0.33355
unacceptable	-21.036	gorgeous	17.949	thoughtless	-0.56109	sympathetic	0.33345
flavorless	-20.997	versatile	17.852	regrettable	-0.55726	phenomenal	0.33086
insulted	-20.826	assisted	17.822	insulting	-0.55290	hardy	0.33078
inexcusable	-20.806	incredible	17.720	atrocious	-0.54360	np	0.32545
disrespectful	-20.680	stunning	17.601	devoid	-0.53357	troubles	0.32544
apologizes	-20.557	superbly	17.582	comical	-0.53046	congratulations	0.32380
substandard	-20.364	jackpot	17.579	shameful	-0.52006	depended	0.31622
insulting	-20.305	skillfully	17.576	speechless	-0.51451	scalloped	0.31149
vomited	-20.176	adorable	17.571	dumbest	-0.51364	catsup	0.31093
disgusted	-20.057	seamless	17.532	declining	-0.51047	proudly	0.30586
uneatable	-19.964	scrumptious	17.413	overbooked	-0.49153	souper	0.30103
humiliated	-19.904	delightful	17.384	subpar	-0.49027	rewarded	0.29642
lifeless	-19.889	seamlessly	17.304	worst	-0.48168	legit	0.28702
disjointed	-19.837	knowledgeable	17.225	destroy	-0.47974	steered	0.28629
miserably	-19.836	enjoyed	17.140	yucky	-0.46464	hustling	0.28310
appalling	-19.670	personable	16.930	disappointing	-0.45567	psyched	0.28096
overcooked	-19.602	impeccably	16.926	ruining	-0.45190	appreciative	0.28092
apologized	-19.583	amazing-	16.912	tastiest	-0.44454	joking	0.27745
reeked	-19.574	mazing	16.847	horrid	-0.44354	powerful	0.27428
disrepair	-19.520	recommande	16.819	disturbing	-0.44239	perfected	0.27220
degrading	-19.249	thoughtful	16.681	obscene	-0.44145	avg	0.26942
pathetic	-19.181	unforgettable	16.500	fiend	-0.43779	cages	0.26939
apologize	-18.936	insightful	16.465	questionable	-0.42931	utmost	0.26898
uninspired	-18.829	guided	16.447	flavorless	-0.42255	deconstructed	0.26239
grossly	-18.755	phenomenal	16.412	disgrace	-0.40968	flippant	0.26056
disgraceful	-18.710	savored	16.359	stingy	-0.40387	reassured	0.25918
deplorable	-18.703	fab	16.335	displeasure	-0.40305	polo	0.25645
wasting	-18.646	unsurpassed	16.239	offended	-0.40262	seamless	0.25565
lied	-18.550	adored	16.047	slim	-0.39541	cokes	0.25446
rudest	-18.539	knowledgable	15.949	disgustingly	-0.39347	shy	0.25239
shoddy	-18.493	beautifully	15.904	wretched	-0.38803	painless	0.25238
stunk	-18.481	excellently	15.655	inaccurate	-0.38731	shined	0.25084

Table 5: Sentiment lexica examples induced using the FFN model and the DistilBERT with masking.

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 3: An example for the Amazon Mechanical Turk HIT (positive sentiment).

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^2_Attn	0.881	0.879	0.837	0.825
DB^3_Mask	0.900	0.900	0.841	0.838
DB^3_SHAP	0.900	0.900	0.841	0.832
RB^4_Mask	**0.918**	**0.919**	0.825	0.826
RB^4_SHAP	**0.918**	**0.919**	0.847	0.841

Table 6: Within-corpus performance of models and lexica for sentiment classification task.

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.635	0.621
SVM_STI	0.721	0.719	**0.718**	**0.717**
FFN_STI	0.693	0.677	0.690	0.686
dLSTM^2_Attn	0.687	0.670	0.673	0.641
DB^3_Mask	0.790	0.787	0.688	0.679
DB^3_SHAP	0.790	0.787	0.683	0.666
RB^4_Mask	**0.805**	**0.804**	0.647	0.645
RB^4_SHAP	**0.805**	**0.804**	0.686	0.675

Table 7: Across-corpora 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^2_Attn	**0.887**	**0.887**	0.702	0.695
DB^3_Mask	0.759	0.76	0.710	0.694
DB^3_SHAP	0.759	0.76	0.703	0.677
RB^4_Mask	0.787	0.788	0.699	0.689
RB^4_SHAP	0.787	0.788	**0.722**	**0.715**

Table 8: Within-corpus performance of models and lexica for emotion classification task.

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.581	0.545
SVM_STI	0.627	0.622	0.620	**0.618**
FFN_STI	0.599	0.590	0.587	0.579
dLSTM^2_Attn	0.613	0.607	0.578	0.541
DB^3_Mask	0.686	0.679	0.620	0.587
DB^3_SHAP	0.686	0.679	0.613	0.586
RB^4_Mask	**0.688**	**0.686**	0.597	0.564
RB^4_SHAP	**0.688**	**0.686**	**0.625**	0.605

Table 9: Across-corpora 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^2_Attn	**0.899**	**0.899**	0.764	0.756
DB^3_Mask	0.825	0.825	0.772	0.761
DB^3_SHAP	0.825	0.825	0.766	0.747
RB^4_Mask	0.850	0.851	0.759	0.754
RB^4_SHAP	0.850	0.851	**0.780**	0.774

Table 10: Within-corpus averaged performance of models and lexica over both sentiment and emotion classification tasks.

Method	Model		Lexicon	
	Acc	F1	Acc	F1
Univariant			0.620	0.598
SVM_STI	0.690	0.687	**0.685**	**0.684**
FFN_STI	0.668	0.657	0.659	0.654
dLSTM^2_Attn	0.665	0.654	0.644	0.609
DB^3_Mask	0.758	0.755	0.667	0.650
DB^3_SHAP	0.758	0.755	0.661	0.641
RB^4_Mask	**0.768**	**0.768**	0.630	0.617
RB^4_SHAP	**0.768**	**0.768**	0.665	0.649

Table 11: Across-corpora 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^2_Attn	0.026	0.019	0.007	0.001
DB^3_Mask	0.016	0.014	5e-14	3e-11
DB^3_SHAP	0.006	0.004	6e-13	2e-10
RB^4_Mask	0.012	0.011	4e-17	1e-14
RB^4_SHAP	0.005	0.003	5e-13	8e-11

Table 12: 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^2_Attn	0.017	0.013	0.005	0.017
DB^3_Mask	5e-4	0.003	7e-4	3e-4
DB^3_SHAP	0.004	0.015	0.006	0.003
RB^4_Mask	3e-4	7e-4	2e-5	8e-5
RB^4_SHAP	7e-4	0.002	0.005	0.002

Table 13: 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[2]_Attn	0.008	0.005	2e-4	8e-5
DB[3]_Mask	9e-5	1e-4	6e-14	4e-13
DB[3]_SHAP	6e-5	5e-4	2e-11	5e-11
RB[4]_Mask	2e-5	2e-5	2e-17	2e-16
RB[4]_SHAP	7e-6	1e-5	6e-12	7e-12

Table 14: p-Values of paired t-tests for f-1 scores between models and lexica over both sentiment and emotion classification tasks.

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

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

	FFN	dLSTM[2]	DB[3]	RB[4]
SVM	**0.005**	**0.012**	**2e-11**	**5e-12**
FFN		0.730	**9e-14**	**1e-11**
dLSTM[2]			**1e-10**	**1e-11**
DB[3]				**0.007**

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

	SVM	FFN	dLSTM^2_Attn	DB^3_Mask	DB^3_SHAP	RB^4_Mask	RB^4_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^2_Attn				0.533	0.504	0.853	**0.003**
DB^3_Mask					0.116	0.349	0.163
DB^3_SHAP						0.579	0.052
RB^4_Mask							**0.029**

Table 17: p-Values of paired t-tests for within-domain lexicon f-1 scores.

	SVM	FFN	dLSTM^2_Attn	DB^3_Mask	DB^3_SHAP	RB^4_Mask	RB^4_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^2_Attn				**4e-4**	**0.006**	0.270	**2e-4**
DB^3_Mask					0.311	**2e-5**	0.470
DB^3_SHAP						**0.019**	0.067
RB^4_Mask							**5e-5**

Table 18: p-Values of paired t-tests for across-domain lexicon f-1 scores.

	SVM	FFN	dLSTM^2_Attn	DB^3_Mask	DB^3_SHAP	RB^4_Mask	RB^4_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^2_Attn				0.18	0.28	0.15	0.29
DB^3_Mask					0.22	0.32	0.24
DB^3_SHAP						0.11	0.63
RB^4_Mask							0.33

Table 19: Averaged Pearson correlation between lexica induced by different approaches.

Methods	**Positive**		**Negative**	
	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^2_Attn	8.3	60.7	14.7	58.7
DB^3_Mask	10.7	50.3	24.3	62
DB^3_SHAP	11.7	30	10.7	40.7
RB^4_Mask	8	22	16.3	48.7
RB^4_SHAP	9.7	29.3	20.7	42.3

Table 20: Percentage of words annotated as “the word describes the **[sentiment]**" or “the word is related to the **[sentiment]**".

Methods	**Joy**		**Anger**		**Fear**		**Sadness**		**Surprise**	
	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^2_Attn	11	25	12	18	18	30	7	17	**9**	**15**
DB^3_Mask	16	31	**19**	19	25	**33**	**8**	18	3	11
DB^3_SHAP	18	22	10	**21**	6	20	2	18	3	11
RB^4_Mask	18	25	3	14	14	28	**8**	22	4	9
RB^4_SHAP	**29**	29	11	17	14	32	2	**23**	4	**15**

Table 21: Percentage of words annotated as “the word describes the **[emotion]**" or “the word is related to the **[emotion]**".