Exploring sentiment constructions: connecting deep learning models with linguistic construction

Shu-Kai, Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Yu-Hsiang, Tseng
Graduate Institute of Linguistics
National Taiwan University
seantyh@ntu.edu.tw

Abstract

This paper presents a linguistically motivated novel framework that automatically identifies sentiment constructions in a corpus with only sentiment-annotated sentences. Construction, a crucial concept developed in Construction Grammar, is a form-meaning pair that relates a pattern with a specific communicative function. However, handcrafting constructions is laborious and often leads to sparse coverage in practice. We address the problem with a construction induction framework which includes three components: a deep-learning-based predictive model to capture the sentiment aspects of the text, a dynamic word parser that agglomerate tokens into (multi-)words units, and a score assignment mechanism to weigh those units based on their contributions to predictions. Units that score highly in the last step are the candid sentiment constructions. They are automatically post-processed with their linguistic contexts to create the final constructions. We experiment with the proposed framework on a sentiment-annotated corpus of online consumer reviews from Taiwan telecom. The proposed framework correctly assigned higher importance to handcrafted constructions. Furthermore, new constructions identified by the framework are validated by annotators' rating data.

1 Introduction

The computational treatment of language and sentiment in general has been widely and explored in affective NLP-related fields, among them *Sentiment Analysis* and its extensions have gained a lot attention and become an indispensable part in many social and commercial applications. The representation paradigms has shifted from the naïve *bag-of-word* to a more intricate *neural language models*. Promising advancements notwithstanding, the challenges in harnessing textual data, from social media in particular, are overwhelming and profound. As textual data requires substantial 'preprocessing' before it can be subjected to processing, and such step often presumes some perspectives that are not shared with *non-mainstream* ones on the nature of language/text.

Unfortunately, there is a conspicuous dearth of related studies that are explicitly aware of the underlying assumptions are not necessarily hold among language researchers. It has also been a paucity of literature on how emotion interfacing with *grammar in general* (Bednarek, 2008; Corver, 2013; Majid, 2012), which could be mainly due to the mainstream *poorly-cognitive* approach to framing linguistic constructs since the Saussurean structuralism, and the evidence of emotional impact in grammar is overlooked. For instance, Clark(2019) illustrates the contrast with two forms in English: 'let's talk about X' and 'let's talk X', where only the latter has the means to unambiguously manifest affect

content (i.e., 'let's talk X' serving as an emotive topic marker that is obligatorily emotive while 'let's talk about X' can be used emotively not not necessarily).

In this paper, we are in line with the usage-based constructionist approach to language (Goldberg, 2006; Goldberg, 2019; Goldberg, 2010), where language is considered as a form-meaning pair that relates a pattern with a specific communicative function, and following Clark(2019) by assuming that affective content would seep into the grammar as *marked constructions* from default representations. To explore the new direction, we propose a linguistically-motivated interpretable framework to automatically detect sentiment constructions (`SentiCon`). We first describe the components in the framework and validate the model predictions with human rating data.

2 Related Works

Construction is the central concept based on which linguistic theories explore how form-meaning pairs build up a complex language system we used in everyday communication (Fillmore, 1988; Goldberg, 2019). Instead of considering language as a clear-cut division between lexicon and syntax, constructions allow us to treat language as a lexicon-syntax continuum (Hoffmann and Trousdale, 2013). Specifically, the principle of form-meanings pairing operates at the lexical, phrase, or syntax level. For example, "dog" is a word and with which we denote a living animal. The lexical word pairs a form (sound/orthographic form) with a referential meaning in the external world; therefore, it is a construct. Similarly, a comparative sentence with open slots is also a construction, such as "Mary is elder than you." Idioms are also constructions, such as "out of the blue" or "take it for granted." These constructions are argued to operate through mapping directly between form and meaning, without transformation and derivation from an underlying deep linguistics structure. Furthermore, the construction is more than a useful linguistic concept. Empirical studies further show that these different levels of construction form a network, a *constructicon*, in the mental grammar of speakers (Bencini and Goldberg, 2000; Pulvermüller et al., 2013).

Construction occurs in different levels of analysis (e.g., word, phrase, sentence), and gradience and variation are also presented at a given level. Consider the example given by Bybee(2000), "I don't know". Although the sentence can be analyzed as a construction composed of pronoun, a negated form of auxiliary, and the main verb, the sentence expresses a distinct discourse-pragmatic meaning. This unique usage is further marked by its likely phonetic reduction involving replace the vowel of "don't" into schwa. If we replace the main verb with another low-frequency verb, such as "commute", these changes may not occur. The example showed that even within the same structural pattern, there are still possibilities of different form-meaning pairing, especially when complex communicative goals are in need.

Evaluative language is one of these complex communication scenarios. To convey judgment of evaluation, speakers not only use emotional words (e.g., happy, good, bad) but implicit evaluations, figurative speeches, or even sarcasm. Some of these expressions might be frequently used that they already consolidate into morpho-syntactic patterns stable enough to be considered as constructions. These constructions, if identified, serve as cornerstones to understand the evaluative messages in the text. However, finding these evaluative constructions are challenging. Manual labeling constructions from raw text always provide high-quality results, but it is time-consuming and challenging to cover a large corpus. Studies have proposed automatic construction extraction tools based on linguistic features, and associative measures (Dunn, 2017; Lee, 2018). Nevertheless, these implementations are more focused on identifying constructions in general rather than evaluative constructions.

In this paper, we aim to identify evaluative constructions (semi-)automatically with the help of deep learning models. Recently, pre-trained language models with transformers (Vaswani et al., 2017; Devlin et al., 2018) have

achieved great successes in numerous NLP tasks, including sentiment analysis, which is closely related to evaluative language. While the exact mechanism on how these models achieve sentiment analysis is not clear, studies are starting to show that model may capture some of the linguistic regularities, such as words' POSes, syntactic relations, or even constructions (Manning et al., 2020; Rogers et al., 2020; Tayyar Madabushi et al., 2020). We explore the possibilities that the deep learning model, when performing sentiment analysis, also learns the form-meaning pairing between the textual patterns and their sentiments. These textual patterns, or potential sentiment constructions, are then be extracted with Shapley scores. In the following sections, we first introduce the overall framework (Sec. 3) in which potential sentiment constructions are identified. Secondly, we compare the model results with human rating data and show model predictions are consistent with human evaluations (Sec. 4).

3 Construction Exploration Framework

The `SentiCon` framework consists of four components: (1) a classification model, (2) an explanation mechanism, i.e. Shapley value (Lundberg and Lee, 2017), (3) a soft word segmentation, and (4) linguistic pattern detector. Firstly, a classification model is trained to capture the relations between the textual form and its communicative meaning (e.g., the sentiment polarity of the text, etc.). Such a model involves the current transformer-based model, in which millions of trainable parameters are tuned to find the optimal set to predict the text label. However, the parameters are often difficult to interpret. Therefore, the explanation mechanism is employed to identify which part of the sentence is significant for the model. In addition, word segmentation is also involved in ensuring the explanation mechanism is correctly informed with the text's word boundaries. The framework explores the potential textual patterns that significantly contribute to the model predictions when working as a whole. Lastly, the framework is equipped with predefined linguistic patterns, with which the model weighs the potential textual patterns. The weighted scores are the model's predicted scores for each extracted potential construction.

3.1 Sentiment classification model

The first component of `SentiCon` is a prediction model for sentiment classification. The role of the classification model not only is to capture the relations between the textual input and the communicative meaning, but it also serves as the `explanandum` of the following explanation mechanism. The downstream explanation mechanism is model agnostic; different kinds of classifiers are all applicable. The only requirement is that the model must at least be capable of performing the sentiment classification. Transformer-based model is a proper choice because of its wide adoptions, performs well on SST-2 dataset, and it achieves the state-of-the-art in one of its variant (Kant et al., 2018; Jiang et al., 2020). Therefore, we choose to fine-tune the off-the-shelf pretrained BERT model for sentiment classification in the following experiment. The model training requires a dataset containing texts and their corresponding sentiment labels. It is noteworthy that model predictions are not directly used in the framework; instead, we are interested in the model (contribution) scores derived from the explanation mechanism.

3.2 Shapley value

The classification model which is performant in sentiment classification usually involves highly complicated non-linear mappings. It is difficult, if not impossible, to directly readout which features or parts of the text are responsible for the model prediction. Therefore, we often need another explanation model to help us credit the text chunks that contribute to the model predictions. Many explanation mechanisms are available, for example, LIME (Ribeiro et al., 2016), concept activation vectors (Kim et al., 2018), etc., but the Shapley (Lundberg and Lee, 2017) value provides a unique insight in the text-based model input.

We use Shapley value to serve as an explana-

tion mechanism in SentiCon. The role of the Shapeley value is to identify the contributing text chunks to the classification model's predictions, and from which we derived the contribution scores of text chunks. Shapley values compute the feature attributions given a model. The attribution is formalized as the difference between the feature coalitions with or without the target feature. However, the exact value requires a complete enumeration of all possible coalitions, which is prohibitively expensive in the text data, the approximated value, $\hat{\phi}$, is computed with sampling.

$$\hat{\phi}_t = \frac{1}{M} \sum_{m=1}^{M} \left(f(x_{+t}^m) - f(x_{-t}^m) \right) \quad (1)$$

The structure embedded in the language provided further constraints on computing Shapley values. We followed the partition approach implemented in SHAP (Lundberg, 2021). Taking advantage of the linguistic properties of Chinese, each character is agglomerated into larger units (i.e., *words*, chunks, prefabs). The tendency on which characters can be grouped is model with their respective word boundary probabilities, determined with a soft parser (see Sec 3.3 for details). These probabilities provide a distance metric on which we cluster these tokens into a hierarchical binary tree structure. The partition approach of computing Shapley values then proceeds in a divide-and-conquer fashion, in which it only needs to compute the left, $\hat{\phi}_{\text{left}}$ or right branch, $\hat{\phi}_{\text{right}}$, at a time. The values of each node in the tree, $\hat{\phi}_{\text{node}}$, is the *interaction* between the branches:

$$\hat{\phi}_{\text{node}} = f(x_{11}) - f(x_{10}) - f(x_{01}) + f(x_{00}) \quad (2)$$

where $f(x_{11})$ denotes the model predictions when both branches are presented, $f(x_{10})$ and $f(x + 01)$ denotes either left or right branch is presented, and $f(x_{00})$ denotes all of which are absent.

The contribution scores of the character group under a specific node, $\hat{\phi}_{\text{group}}$, are computed from their antecedents and descendants:

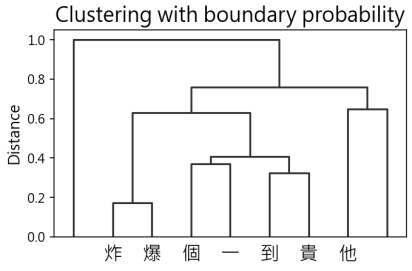

Figure 1: Hierarchical clustering, using single linkage, with soft word boundary probabilities. The first and the last token are the special CLS and SEP tokens.

$$\hat{\phi}_{\text{group}} = \hat{\phi}_{\text{upper}} + \hat{\phi}_{\text{lower}}$$
$$\hat{\phi}_{\text{lower}} = \hat{\phi}_{\text{left}} + \hat{\phi}_{\text{right}} + \hat{\phi}_{\text{node}}$$
$$\hat{\phi}_{\text{upper}} = 0.5 \cdot \hat{\phi}_{\text{upper}}^{parent}$$

The difference between the scores of the parent and its children signifies the non-compositionality of the corresponding character groups. If the parent node has significant interactions (with respect to its children), the character groups are considered a potential construction pattern. The corresponding score, $\hat{\phi}_{\text{group}}$, is its contribution to the model prediction.

3.3 Word segmentation & POS tagging

The partition approach in SHAP implementation facilitates the efficiency of computation, but how the text is partitioned into a binary tree is highly pertinent to a good explanation, especially in the context of SentiCon. Ideally, if researchers know *a priori* the constructions, prefabs, or multi-word expressions used in the text, the framework can partition the text following the linguistic knowledge and assign contributing scores to each of them. However, the existing constructions (especially in Chinese) are usually sparse and have low coverage on the actual corpus. It is also the reason why we want to develop the SentiCon in the first place, i.e., to (semi-)automatically explore sentiment-laden constructions. Therefore, we attempt another approach to construct the partition with word boundary possibilities provided by a soft word parser.

他_0.58 Nh_0.50/ VH_0.04/SPC_0.03
貴_0.64 VH_0.50/ VC_0.04/ VJ_0.04
到_0.41 VH_0.15/VCL_0.12/ VJ_0.10
一_0.45 Neu_0.53/ VH_0.05/ Nb_0.02
個_0.43 Nf_0.37/ VH_0.05/ Nc_0.04
爆_0.57 VH_0.32/ Nv_0.27/ Na_0.14
炸_0.31 VH_0.30/ Nv_0.28/ Na_0.12

Figure 2: Soft-tagging results of the example sentence, *"it is awfully overpriced"*. The first column indicates the probabilities of being word boundaries. Following are the top 3 POS classes. The tag set follows CKIP POS tagset from Academia Sinica, Taiwan.

A Chinese word segementer and a POS tagger are used to ensure the partition approach of Shapley value is properly informed with linguistic structure. The pipeline (Hsieh and Tseng, 2020) includes two modules: word segmentation and part-of-speech tagging. This pipeline is unique in that it highlights the profound issues of *wordhood* in Chinese (Huang et al., 2017). Instead of assigning hard word boundaries and POS for each *word*, the pipeline operates on character levels. It assigns a word boundary probability on each character and the probabilities of which it functions as a given POS (Figure 2). The word probabilities are used by the partition mechanism (Sec 3.2) as a metric to construct token clusterings (Figure 1). Figure 1 also illustrate the strength of soft segmentation: it captures the potential structure between words otherwise defined. Specifically, 貴到一個, guì dào yī ge, "overpriced to an (extent)", would be traditionally segmented as four independent words but are shown to be grouped together under soft segmentation. Such cluster information is essential in discovering potential patterns.

3.4 Pattern weighting

After the Shapley values identified the potential construction patterns in the sequence, a post-processing step used the POS information to weight these patterns. This step provides further constraints on the potential patterns. For example, a pattern containing an adverb-verb structure is more likely to be a potential construction; therefore, it should be assigned more positive weights.

To detect the underlying structure in the sequence, we cross-product the POS matrix, P_{seq}, with a structure kernel, Q_s, for each structure, s, in a predefined set, $\mathcal{S}$. The POS matrix is provided by the soft tagger (see Figure 2 for an example). Each m token in a sequence has a POS probability distribution over k POS tags, resulting in a $m \times k$ matrix. The kernel is itself a $p \times k$ binary matrix, where each element in the matrix denotes the corresponding POS tag in that position. The weight of each sequence is computed as

$$w_{\text{seq}} = \text{MaxPool}_{s \in \mathcal{S}} \max(P_{\text{seq}} * Q_s) \quad (3)$$

where $*$ denotes the cross-product operator. The scores are max-pooling across different kernels and result in the final pattern weight.

4 Experiment

We experiment with the proposed framework on discovering potential construction on a dataset of evaluative text about telecom service in Taiwan. The dataset contains 2,622 text sequences, each of which is under length of 100 characters. Each sequence is manually labeled with a sentiment polarity, positive or negative. This experiment aims to test if the proposed framework can find potential constructions in the dataset with only sequence-level annotation.

The experiment proceeds as two analyses. First, a confirmatory analysis test the model's ability to identify construction is sentiment-laden. Specifically, the model computes the construction score for each manually written construction, and we test the relations between the model scores and constructions polarities. Secondly, the exploratory analysis is set to discover the potential patterns from raw sequences.

Both of the analyses used the same classification model trained from the evaluative text dataset. The dataset is split into the training and testing set by a ratio of 9:1. The classification accuracy on the testing set is 77%. To further compare the constructions identified by

the model, we manually compiled another list of sentiment constructions. This list consists of 39 sentiment constructions, written in the form of regular expression patterns. These constructions are handcrafted by a native speaker who is a graduate student in the Graduate Institute of Linguistics, National Taiwan University, and each of the constructions is annotated with its sentiment polarity (positive or negative). The manual compiled list is used in the following confirmatory analysis.

4.1 Confirmatory analysis

The purpose of the confirmatory analysis is to test whether the model's prediction scores indeed identify sentiment-laden textual patterns. Specifically, we use the manually compiled list of sentiment constructions as a reference and compare their model's prediction scores to see if the prediction scores agree with the sentiment polarities with the constructions.

Confirmatory analyses show that construction score reflects the sentiment polarities in the data (Figure 3). The predicted score of each construction instance is extracted from the closest partition in the text. The score significantly correlates with the sequence sentiment in which the construction resides, $r = .74$, $t(83) = 10.01$, $p < .0001$. Furthermore, a logistic regression model is used to model the relation between the construction scores and construction polarities. The estimated coefficient of construction scores in the model is statistically significant, $b = 0.30$, $z = 2.90$, $p < 0.005$. The results indicate the construction score computed by the model are consistent with human rating. Specifically, the higher the construction score, the more likely the construction being rated as positive.

It is worth noting that the framework itself has no clue of what construction is, yet it can significantly predict the sentiment of them by model scores. The model itself is trained on sequence-level annotation. It has no supervised signals, at least explicitly, to the inner sentential structure of the text. However, when guided with the information of word boundaries and the help of Shapley values, the model can generate consistent scores on human-labeled constructions. This result, while interesting, is also consistent with what previous studies found on pretrained language models. Researchers found they can probe lexical or phrase level structures in the model internal representations (**?**; Manning et al., 2020).

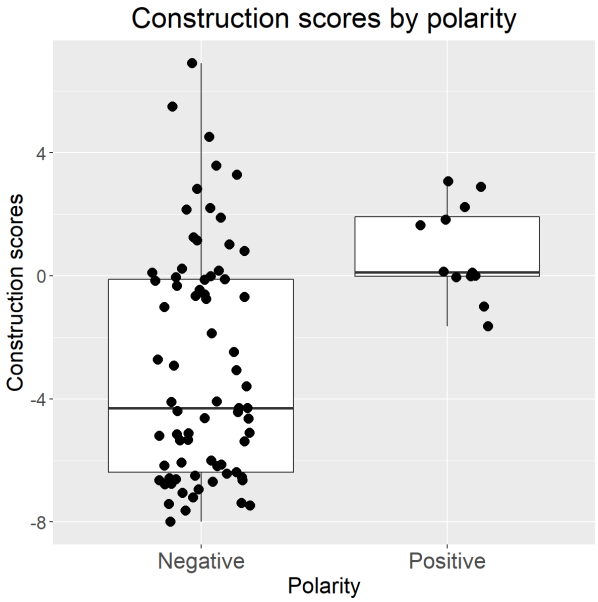

Figure 3: The distribution of manually edited constructions scores. Higher scores indicates more positive contribution in the classification model.

4.2 Exploratory analysis

Following the confirmatory analysis, the exploratory analysis further investigates whether the model can identify potential constructions. In the previous confirmatory analysis, we present the model with a human-made construction and compare their prediction scores. In contrast, the exploratory one asks the model what is the most probable constructions in the text, given it had *learned* the relations between the texts and their sentiment polarities.

In the exploratory analysis, we use 2,429 sequences, which the model correctly classifies the sentiment, to explore potential construction patterns. By computing the model prediction scores of each partition in each text, we extracted 3,297 candidate patterns. These patterns are further filtered by the following criteria: (1) their length (length of 3 to 10 characters

Pattern	Translateion	Model Score	Rating
推薦 [ENT]	*recommend* `[ENT]`	7.27	`Positive`
比較期待 [ENT]	*(be) more looking forward to* `[ENT]`	6.06	`Positive`
很不錯	*very good*	4.63	`Positive`
有夠爛	*(being) extremely bad*	-5.77	`Negative`
[ENT] 網速整個悲劇	*network speed of* `[ENT]` is awful	-6.18	`Negative`
[ENT] 很爛	`[ENT]` *is terrible*	-6.82	`Negative`

Table 1: Excerpt of construction patterns discovered by the framework. Three positive and three negative constructions are listed as examples. `[ENT]` is the placeholder for the proper noun (e.g. names of the network providers) used in the text.

are included), (2) the absolute value of model scores (only absolute values larger than the median are included), and (3) their pattern weights (values larger than the median are included).

There are 248 construction patterns identified from the framework(see Table 1 for examples). Two annotators firstly rated the polarities of these patterns. Secondly, they are also asked if construction is embedded in the patterns and generate a regular expression if there is one. Annotators are both native speakers who are graduate students in the Graduate Institute of Linguistics, NTU.

Annotators have strong inter-rater agreement on the patterns' polarities, $\kappa = .80$. There are 110 patterns are rated as having positive (22 patterns) or negative (88 patterns) sentiments, and the model score could significantly predict sentiment ratings in the logistic regression model, $b = 1.50$, $z = 2.59$, $p < 0.01$. In addition, the annotators are able to extract 55 additional constructions from the discovered patterns.

Overall, the exploratory analysis shows encouraging results. The complete auto-generated constructions still show agreements in their model scores and human ratings. Furthermore, a fraction of those auto-generated patterns, although not in high proportion, can help human annotators construct constructions. Given the labor-intensive nature of handcrafting constructions, the exploratory results of the current model are still beneficial in identifying constructions.

5 Conclusion

In this paper, we aim to draw attention to a promising new direction in sentiment analysis. The construction grammar provides a natural ground to understand and extend the predictive results from deep learning models. The proposed framework, *SentiCon*, integrates the classification model, the Shapley values, NLP pipelines, and pattern weighting. Both confirmatory and exploratory analyses show promising results on the model's capacity to analyze and discover potential constructions. Furthermore, the proposed framework help address the issue of identifying constructions in the real-life text corpus, which is otherwise time-consuming, and may facilitate future theoretic and application studies on construction grammar.

Acknowledgements

This work was supported by Ministry of Science and Technology (MOST), Taiwan, Grant Number MOST 110-2634-F-001-011, and Chunghwa Telecom, Taiwan.

References

Monika Bednarek. 2008. *Emotion talk across corpora*. springer.

Giulia M.L Bencini and Adele E Goldberg. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language*, 43(4):640–651.

Madeline Clark, Najia Khaled, Miriam Kohn, and Solveiga Armoskaite. 2019. Let's talk emotions: A case study on affective grammar. *Glossa: a journal of general linguistics*, 4(1).

Norbert Corver. 2013. Colorful spleeny ideas speak furiously. *A passionate question at the interface of language and emotion. Ms. Utrecht OTS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jonathan Dunn. 2017. Learnability and falsifiability of construction grammars. *Proceedings of the Linguistic Society of America*, 2:1–1.

Charles J Fillmore. 1988. The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, volume 14, pages 35–55.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Adele E Goldberg. 2010. Constructions: A new theoretical approach to language. *Sprachwissenschaft*, pages 717–729.

Adele E Goldberg. 2019. *Explain me this*. Princeton University Press.

Thomas Hoffmann and Graeme Trousdale. 2013. Construction grammar: introduction. In *The Oxford handbook of construction grammar*. Oxford University Press.

S. K. Hsieh and Y. H. Tseng. 2020. Tutorial on sense-aware computing in chinese. In *Paper presented in 32nd conference on Computational Linguistics and Speech Processing (ROCLING 2020)*.

C.R. Huang, S.K. Hsieh, and K.J. Chen. 2017. *Mandarin Chinese Words and Parts of Speech: A Corpus-based Study*. Routledge Studies in Chinese Linguistics. Taylor & Francis.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online, July. Association for Computational Linguistics.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul.

Sophia Yat Mei Lee. 2018. *Emotion and cause: Linguistic theory and computational implementation*. Springer.

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.

S. Lundberg. 2021. Shap (shapley additive explanations). `https://github.com/slundberg/shap`.

Asifa Majid. 2012. Current emotion research in the language sciences. *Emotion Review*, 4(4):432–443.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Friedemann Pulvermüller, Bert Cappelle, and Yury Shtyrov. 2013. Brain basis of meaning, words, constructions, and grammar. In *The Oxford handbook of construction grammar*.

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, June. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.