Representing affect information in word embeddings

Yuhan Zhang, Wenqi Chen, Ruihan Zhang & Xiajie Zhang*

Abstract. A growing body of research in natural language processing (NLP) and natural language understanding (NLU) is investigating human-like knowledge learned or encoded in the word embeddings from large language models. This is a step towards understanding what knowledge language models capture that resembles human understanding of language and communication. Here, we investigated whether and how the affect meaning of a word (i.e., *valence*, *arousal*, *dominance*) is encoded in word embeddings pre-trained in large neural networks. We used the human-labeled dataset (Mohammad 2018) as the ground truth and performed various correlational and classification tests on four types of word embeddings. The embeddings varied in being static or contextualized, and how much affect specific information was prioritized during the pre-training and fine-tuning phase. Our analyses show that word embedding from the vanilla BERT model (Devlin et al. 2019) did not saliently encode the affect information of English words. Only when the BERT model was fine-tuned on emotion related tasks or contained extra contextualized information from emotion-rich contexts could the corresponding embedding encode more relevant affect information.

Keywords. Language models; word embeddings; affect meaning; lexical semantics

1. Introduction. With the success of large neural network models in completing complicated language tasks, evaluating the models' interpretability and intrinsic capabilities has become a heated research trend (e.g., Manning et al. 2020, Mikolov et al. 2013). The evaluation work could be roughly classified into two types: one that relies on the output of the language models (LMs) to infer the model's linguistic ability and the other that looks into the components of LMs (e.g., word embeddings) for such inspiration. While previous evaluation tasks have focused on testing LMs' explicit linguistic knowledge (e.g., syntactic knowledge such as islands, semantic knowledge such as compositionality, word-level knowledge such as polysemy), we pick a piece of knowledge that is less studied but essential to intelligence. Specifically, we studied whether and how word embeddings learned via supervised methods in large neural networks encode the affect information of a word (e.g., valence, arousal, dominance). Our work shows that even though contextualized word embeddings were in general better at capturing intricate affect meanings compared to static word embeddings, especially after being fine-tuned on emotion related tasks, word embeddings from vanilla BERT did not attain salient affect knowledge.

In Section 2, we detailed relevant work that led us to our investigation. In Section 3, we laid out the unsupervised and supervised methodologies we took and in Section 4, we spelled out the findings. In Section 5, we discussed implications of our research and possible future directions.

*We would like to thank Mycal Tucker, Roger Levy, and the audience at ELM 2022 for their generous feedback. All mistakes are ours. Authors: Yuhan Zhang, Harvard University (yuz551@g.harvard.edu); Wenqi Chen, Harvard University (wenqichen@g.harvard.edu); Ruihan Zhang, Massachusetts Institute of Technology (ruihanz@mit.edu); Xiajie Zhang, Massachusetts Institute of Technology (xiajie@mit.edu).

2. Related work. Three major aspects in the current NLP world are related to our work: The linguistic knowledge revealed to be grasped by language models, the relationship between what word embeddings have achieved to represent and the actual lexical semantics, and how affect information is studied in natural language processing.

2.1. WHAT LANGUAGE MODELS CAN LEARN. A growing body of research has investigated what linguistic knowledge large artificial neural networks can learn while they are trained in a supervised way to predict the next word. In the syntactic category, scholars have shown that large LMs have varying grammatical knowledge ranging from the local subject-verb agreement to long-distance filler-gap dependencies (e.g., Hu et al. 2020, Linzen et al. 2016, Warstadt et al. 2020, Wilcox et al. 2018). Aside from looking at predicted words to infer LMs' linguistic knowledge, Manning et al. (2020) show that a linear transformation of word embeddings from BERT (Devlin et al. 2019) captures linguistic hierarchical structures. There is also increasing attention to understanding LMs' abilities to represent meanings. In semantics and pragmatics, promising and positive results seem to support LMs' increasingly sophisticated abilities such as doing natural language inference (e.g., Poliak et al. 2018, Wang et al. 2018). Instead of relying on direct output of LMs, finding the relationship between human understanding of language and representations in word embeddings have also been fruitful.

2.2. WORD EMBEDDING AND LEXICAL SEMANTICS. Since deep contextual language models provide contextualized word embeddings which naturally encode the distance between word tokens in a vector space, this property can be utilized to test whether the trained distance in word embeddings reflects the natural way words group together according to our lexical semantic knowledge. Existing studies have shown that the pre-trained BERT model is able to place polysemous words that appear in different contexts into distinct regions of the shared vector space (Wiedemann et al. 2019) and the word sense distances are correlated with human judgments (Nair et al. 2020). There is also the general claim that contextualized word embedding in BERT is good at word sense disambiguation (Loureiro et al. 2021). But in addition to this line of research on word sense disambiguation, investigation into other aspects of lexical semantics is limited. We made an attempt in this work to bring together studies of word embedding representations and another aspect of lexically encoded meaning – the affect information – as a novel case study of what LMs can learn from supervised training.

2.3. AFFECT INFORMATION IN NLP. Affect information refers to any explicit or implicit emotion related information. We use the word *affect* instead of *emotion* because we do not rely on common emotion words such as *happy* and *sad* as baselines of comparison; rather, we rely on three primary independent dimensions of emotions as scales to quantify the meaning of emotion. The three dimensions are *valence* (positiveness-negativeness/pleasantness-unpleasantness), *arousal* (active-passive, some people also take it to mean the intensity of the emotion invoked by the word), and *dominance* (dominant-submissive, or the level of control exerted by the word) (Osgood et al. 1957, Russell 1980, 2003). Another reason to choose this terminology is due to its validity and consistency: there are already multiple human-labeled datasets that are based on these scales as the quantitative ground truth (Bradley & Lang 1999, Mohammad 2018, Warriner et al. 2013).

In the fields that are related to emotion recognition, sentiment analysis, and affective com-

puting, transformer-based models and more fine-grained annotation of emotion categories have gathered the attention of the NLP/NLU community (e.g., Alswaidan & Menai 2020, Demszky et al. 2020, Rashkin et al. 2018, Suresh & Ong 2021). Zooming into existing studies relevant to word embeddings and affect information, it has been shown that multi-dimensional embeddings of emotional words achieved higher performance in sentiment analysis tasks than human annotated emotion vectors (Li et al. 2017). Furthermore, the multi-dimensional word embeddings for emotional words are often added onto base static word embeddings for downstream sentiment analysis tasks (Mao et al. 2019). While existing research focused more on the application of emotional word embeddings onto sentiment analysis tasks, seldom has explored whether state-of-the-art contextualized word embeddings have learned the affect meaning of all kinds of words. Our study is set up to fill the gap.

3. Methodology. Our methods focus on constructing the relationship between word embeddings and human cognition. On the human cognition side, we adapt the dataset from the affect ratings of humans from cognitive studies. On the word embedding side, we adopted both contextualized and static word embeddings. In order to probe the relationships, we take both the unsupervised and supervised methods to reflect the relationships.

3.1. AFFECT RATINGS OF HUMAN AS THE GROUND TRUTH. We selected the NRC VAD dataset (National Research Council (Canada) Valence Arousal Dominance) curated by Mohammad (2018) as the ground truth for affect meanings of English words.[1] There are 20,000 English words annotated along the three independent dimensions of affect. In each dimension, the rating starts from 0 to 1 where 0 means the most negative, passive, or submissive in the V, A, or D dimension and 1 means the most positive, active, or dominant in their respective dimension. This is the largest manually created VAD corpus in any language.

3.2. WORD EMBEDDINGS UNDER STUDY. In this paper, we studied four kinds of word embeddings as described below. We took from each word embedding the same subset of words that are compatible with the NRC VAD corpus. The embeddings differ from each other by how much and what kinds of contextualized information are captured from training. We took GloVe as a representative of existing static word embeddings in the NLP community and compared it with more contextualized ones. We also varied the kinds of contextualized information by prioritizing emotion related and sentiment analysis related information.

- Pre-trained word embeddings using the **GloVe** algorithm (Pennington et al. 2014) which we refer to as the GloVe embedding in this paper. The word embeddings were downloaded from Stanford NLP website[2] which were trained on Wikipedia and Gigaword 5 data.
- Pre-trained word embeddings that were retrieved from the last hidden layer of the base BERT model (Devlin et al. 2019) after we directly ran the model over the NRC word list. We refer to this type of word embedding representation as **base BERT** throughout the paper.
- Pre-trained word embeddings retrieved from the last hidden layer of the BERT model fined tuned on the GoEmotion dataset (Demszky et al. 2020). The GoEmotion dataset contains 58k English Reddit comments labeled by humans on 27 emotion categories. We refer to this

[1] Data can be downloaded from http://saifmohammad.com/WebPages/nrc-vad.html
[2] https://nlp.stanford.edu/projects/glove/

as **GoEmotion BERT** in the paper.

- Given that the BERT-based word embeddings are supposed to be contextualized because they could be functions of the entire input sentence (Ethayarajh 2019), we utilized this contextual sensitivity feature and output **contextualized BERT** word embeddings by deriving them from running IMDB movie reviews (Maas et al. 2011). For each NRC word, we first extracted all its embeddings by retrieving the last hidden layer of the BERT model, we then computed the first principal component to derive the single word vector for each individual word.

3.3. UNSUPERVISED PROBE: PRINCIPAL COMPONENT ANALYSIS. Firstly, in order to investigate whether the types of word embeddings under investigation encode affect information, we conducted dimensionality reduction via Principal Component Analysis (PCA from the *sklearn* package (Buitinck et al. 2013, Pedregosa et al. 2011)) over each of the studied word embeddings. PCA is a convenient tool for such visualization without us going deep into deciphering the architecture of multi-dimensional word embeddings. Specifically, we sought for the correlation between each principal component from the embeddings and the human ratings of each VAD dimension. Higher correlation indicates that the high-dimensional information in that word embedding is more likely to saliently encode affect information. This unsupervised approach provides a rudimentary insight into whether affect meanings are well captured by the word embeddings being investigated.

3.4. UNSUPERVISED PROBE: COSINE DISTANCE FOR SEMANTIC SIMILARITY. It has been shown that cosine distances between vector representations of words indicate the words' semantic relatedness (Dumais et al. 1988, Mikolov et al. 2013, Bojanowski et al. 2017). The literature we learned from is Nair et al. (2020). They investigated whether contextualized word embeddings – BERT embeddings in this case – capture human-like distinctions between English word senses, such as polysemy (chicken as animal vs. chicken as meat) and homonymy (bat as mammal vs. bat as sports equipment). For each pair of target word senses, they measured (1) the cosine distance of the two embeddings and (2) the human judgment of the relatedness of the word senses in a 2-dimensional spatial arrangement task. They then applied Spearman rank correlation analysis to these two measurements. The results turn out that the distance of word senses in BERT's embedding space correlated with human judgments and that the correlation of homonyms was higher.

Learning from Nair et al. (2020), we took a small sample of affect words and ran correlation tests to see the relationship between the word embedding space and the vector space of human judgments (i.e., the VAD 3-dimensional space). Similar to the PCA results, a stronger correlation means a more salient encoding of real-world affect information in the investigated word embeddings. More about the sampling method: Shaver et al. (1987) defined six basic emotion categories (i.e., love, surprise, sadness, anger, joy, fear) and a list of affect words in each category, totalling 132 affect words. We took 80 affect words from Shaver et al. (1987) (e.g., *disgust*, *envy*, *enjoyment*, *desire*) and calculated the cosine similarity between each and another word, resulting in 80×79 similarity scores. We did the same thing by iterating over the kind of vector space, from the NRC VAD 3-dimensional human judgment space, to the four kinds of word embeddings introduced in section 3.2, resulting in $80 \times 79 \times 5$ pairwise similarity scores. Then, we did Spearman rank correlation analysis between one and another embedding representations. We adopted a nonparametric correlation like Nair et al. (2020) because, first, the numerical distribution of a given type

of word embedding does not satisfy the normality assumption and, second, we cared more about whether the two embedding representations share a monotonic relationship, instead of a strictly linear relationship.

3.5. SUPERVISED PROBE: LINEAR CLASSIFIER. In addition, we experimented with supervised learning methods to probe whether the word embeddings can be linearly separable into the VAD dimensions. We trained a one-layer neural network (i.e., a logistic regression model) to predict the binary VAD labels[3] with the embeddings as inputs. The linear classifier was designed to be as simple as possible to reveal information encoded in the embeddings. If the prediction results match well with the human criterion, it is reasonable to conclude that the word embeddings have the capability to represent the affective information.

4. Analyses & Results.

4.1. PRINCIPAL COMPONENT ANALYSIS. Figure 1 shows the PCA result. We extracted 5,586 words that are in the vocabulary of GloVe, base BERT, GoEmotion BERT, and contextualized BERT embeddings, ran PCA over these word embeddings, and represented the numerical value of the first two principal components along the two axes. Each horizontal panel represents one dimension of VAD. Each column in Figure 1 represents a type of tested word embedding. Each point is color coded with human judgment. The darker the color, the higher the rating is along one of the VAD dimensions. Table 1 shows the Spearman correlation coefficient between the numerical values of human judgments along each affect dimension and the numerical values of the corresponding principal component given a type of word embedding.

From Table 1, we learn that every type of word embedding can capture some aspects of the 3-dimensional affect meaning based on the significance level of the correlation coefficients. Yet judging from the magnitude of correlation coefficients for both principal components as well as the number of correlation coefficients that are larger than 0.2 out of the three VAD dimensions, we see that the GoEmotion BERT embedding and the contextualized BERT embedding outperform the other two.

Dimension	PCA	GloVe	base BERT	GoEmotion BERT	contextualized BERT
Valence	1st	0.117 (.000)	0.015 (.247)	0.084 (.000)	-0.077 (.000)
	2nd	-0.006 (.672)	0.133 (.000)	**0.227** (.000)	**0.291** (.000)
Arousal	1st	-0.005 (.696)	0.011 (.412)	**0.285** (.000)	**0.251** (.000)
	2nd	**0.238** (.000)	-0.094 (.000)	-0.042 (.002)	-0.007 (.599)
Dominance	1st	**0.316** (.000)	0.028 (.037)	0.160 (.000)	0.086 (.000)
	2nd	0.182 (.000)	0.111 (.000)	**0.283** (.000)	**0.380** (.000)

Table 1: Spearman correlation coefficients (p values in the bracket) between human ratings (0-1) in each affect dimension and each corresponding principal component of four types of word embeddings (correlation coefficients in bold when larger than 0.2, also see each subplot in Figure 1)

In Table 2, we present the explained variance ratios to show what percentage of the variance in the whole embedding space was explained by the first two principal components for each type

[3]We transformed the numerical human rating (0-1) to a binary categorical variable with the threshold of 0.5.

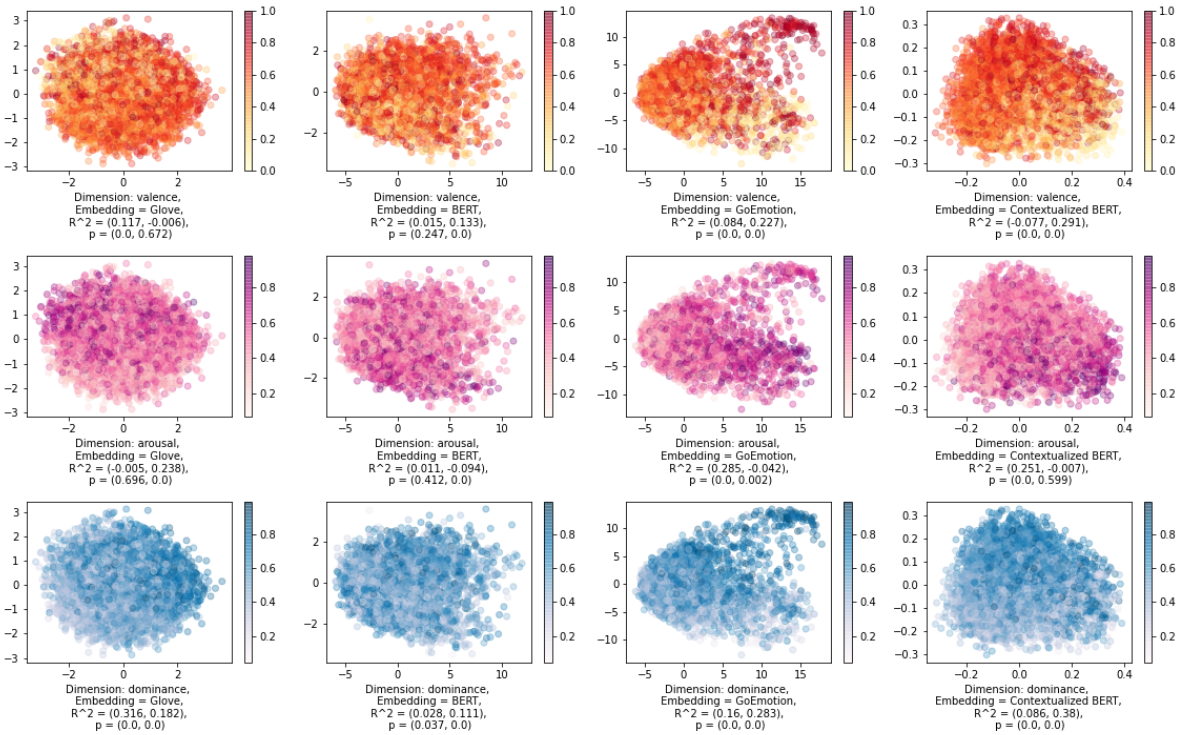

Figure 1: 2-components PCA representations of GloVe, base BERT, GoEmotion BERT, and contextualized BERT & human ratings for valence, arousal, and dominance (Human ratings were color-coded in a spectrum. Each dot represents a word. 5,586 words are represented. The darker the dot, the more prominent the human rating in the respective dimension.)

of word embedding. Two pieces of observations are worth interpreting. First, the relatively low ratio of explained variance across the eight values indicates that the information represented in the first two principal components might not be representative to generalize on the capacity of certain word embedding to encode certain affect meaning. There might be some affect meaning that has been captured in higher dimensions of the vector space but not projected in the PCA measure. We would need more sensitive probes other than PCA in future investigations. Second, assuming that the PCA analysis is a great approximant for affect encoding, the fact that the first component of base BERT accounts for 29.4% of the total variance and yet does not correlate with VAD as much as the other word embeddings shows that base BERT might not be optimized to distinguish intricate affect meaning of words. This offers a nice window to peek into what linguistic information is and isn't weighted the most for BERT.

Explained ratio	GloVe	base BERT	GoEmotion BERT	contextualized BERT
First	0.040	0.294	0.118	0.021
Second	0.030	0.040	0.070	0.016

Table 2: Explained variance ratio of the two principal components

4.2. COSINE SIMILARITY. Table 3 captures the Spearman correlation coefficients based on pairwise cosine similarity of 80 words whose embeddings were from each type of word embeddings. The main comparison is in the first column where human judgments on the semantic similarity of the 80×79 pairs of words are compared with the cosine distance of those word pairs based on different types of word embedding. The result shows that all the four types of word embeddings show a significant correlation with human affect judgment. The base BERT embedding has the weakest correlation of all, which also echoes with Figure 1. Then, in the order of more relatedness is the GoEmotion BERT embedding, GloVe, and the contextualized BERT. It is worthwhile to deliberate on why the static word embeddings like GloVe outperforms the base BERT. Also, it is interesting that even though the GoEmotion BERT is fined tuned on emotion specific task, the correlation is still worse than BERT-based embeddings with movie specific contextualized information.

CORR (p)	VAD	GloVe	base	GoEmotion	contextualized
VAD	1.000 (.000)				
GloVe	**0.272** (.000)	1.000 (.000)			
base	0.116 (.000)	0.148 (.000)	1.000 (.000)		
GoEmotion	**0.252** (.000)	0.172 (.000)	0.013 (.471)	1.000 (.000)	
contextualized	**0.314** (.000)	**0.710** (.000)	**0.240** (.000)	**0.204** (.000)	1.000 (.000)

Table 3: Spearman Correlation (p value) of pairwise cosine similarities between each and the rest of word embedding types (Coefficients larger than 0.2 and less than 1 are in bold.)

4.3. LINEAR CLASSIFIER. In this section, we report the classification results from the linear classifier probe. The training covariate matrix of the classification model was each type of word embeddings. The response variables were the binary categorical label representing lower or higher range of one of the VAD dimensions. The conversion was based on a 0.5 threshold. Each type of word embedding with each affect dimension constituted a linear classifier and together we constructed 12 classifiers. We used the NRC VAD vocabulary as the training and the validation data (with a 70/30 split). The test data comprised of 130 word with strong affect information randomly extracted from Shaver et al. (1987).

We compared the performance of GloVe, base BERT, GoEmotion BERT, and contextualized BERT embeddings on each of the three affect dimensions and present the result in Table 4. It is clear that the contexualized BERT embedding shows superior performance compared to the others. To be noted, our validation set included many words whose VAD score might not be as salient and intuitive as the ones in the test set. This might explain that the validation set received lower accuracy score than the test sample.

Besides, we observed that all of the model embeddings are best at predicting the valence dimension compared with arousal and dominance. This matches our prior knowledge because valence represents the positive or negative of the words and carries the most straightforward meaning for the context to capture. But arousal and dominance will be more complicated and subtle to capture depending on its distributional semantics. Interestingly, GoEmotion BERT does not perform very well in the dimensions of arousal and dominance. This phenomenon leaves ample space to investigate further how word embeddings are trained to capture certain meaning but not the others.

Embedding types	Validation accuracy			Affect word sample accuracy		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
GloVe	0.75	0.70	0.73	0.84	0.74	0.75
base BERT	0.76	0.73	0.74	0.93	0.85	0.82
GoEmotion BERT	0.68	0.68	0.70	0.92	0.76	0.76
contextualized BERT	**0.85**	**0.77**	**0.85**	**0.95**	**0.88**	**0.90**

Table 4: Performance of different word embeddings in predicting VAD labels.

5. Discussion & Conclusion. In this work, we investigated the capability of word embeddings to capture affect meaning. By combining results from unsupervised and supervised learning, we found that our contextualized BERT embedding, where the representation of each word is the first principal component of all embeddings of the same word in its multiple occurrences in the IMDB dataset. In comparison, the base BERT does not show prominent encoding of affect meaning and sometimes is even inferior to the static GloVe. The GoEmotion BERT embedding also does not perform as well as the contextualized BERT embedding. Out of the VAD dimensions, it is hard to know which dimension is relatively well represented or easy to capture: Based on the PCA results, all VAD dimensions could attain a correlation higher than 0.2 in some embedding representations; based on the linear classifier results, valence seems to be better represented than the other two. The discrepant pattern might result from the different sample sizes used in these tasks (i.e., 5,586 words for PCA and 130 words as the test set for the linear classifiers) and more controlled experiments are needed to pin down the fact.

5.1. IMPLICATIONS. Here we provide a novel aspect of testing word embedding capabilities that involve intricate emotion and affect related meaning. Based on what lexical and linguistic knowledge word embeddings can and cannot capture, especially for the renowned BERT embedding (Ettinger 2020, Klafka & Ettinger 2020, Manning et al. 2020, Pandia et al. 2021), we add one additional piece of information. The relatively poor encoding of affect information in BERT might suggest these lines of implications: (1) the affect information of English lexicon is by nature implicit and hard to capture from the distributional features of words in huge corpora; (2) for sentiment related NLP tasks or affect computing tasks (Maas et al. 2011, Socher et al. 2013, Zhang et al. 2015), we might consider using affect enriched embeddings or leveraging additional emotional capabilities of LMs from transfer learning.

5.2. LIMITATIONS & FUTURE WORK. As far as we know, this is the first attempt in the NLP community to investigate the capability of word embeddings to encode affect information in words. We lay three lines of research for future investigations. First, from an algorithmic point of view, we would like to ask what specific training mechanisms enable word embeddings to capture certain information, especially those intricate meanings that don't rely on co-occurrence and distributional semantics. Second, talking about meaning and semantics in a broader sense, there are still tons of unknowns that fall outside of the realm of formal semantics or compositionality. More rigorous identifications and descriptions of different kinds of meaning are needed for studying the knowledge of meaning in LMs. Third, to what extent can language carry affective information? This is an essential question in affective computing (Picard 2000, 2003, Tao & Tan 2005). A better under-

standing from both the algorithmic level and the linguistic level of affect encoding and recognition would provide valuable insights into this question.

References

Alswaidan, Nourah & Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems* 62(8). 2937–2987.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *ACL Transactions of the Association for Computational Linguistics* 5. 135–146.

Bradley, Margaret M & Peter J Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Tech. rep. The Center for Research in Psychophysiology.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt & Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade & Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4040–4054. Online: Association for Computational Linguistics. 10.18653/v1/2020.acl-main.372. https://aclanthology.org/2020.acl-main.372.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Dumais, Susan T, George W Furnas, Thomas K Landauer, Scott Deerwester & Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–285.

Ethayarajh, Kawin. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 Embeddings. 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)*, 55–65. Hong Kong, China: Association for Computational Linguistics. 10.18653/v1/D19-1006. https://aclanthology.org/D19-1006.

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

Hu, Jennifer, Jon Gauthier, Peng Qian, Ethan Wilcox & Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1725–1744. On-

line: Association for Computational Linguistics. 10.18653/v1/2020.acl-main.158. https://aclanthology.org/2020.acl-main.158.

Klafka, Josef & Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. *arXiv preprint arXiv:2005.01810* .

Li, Minglei, Qin Lu, Yunfei Long & Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing* 8(4). 443–456. 10.1109/TAFFC.2017.2723012.

Linzen, Tal, Emmanuel Dupoux & Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4. 521–535.

Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar & Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics* 47(2). 387–443.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng & Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1015.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal & 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. https://doi.org/10.1073/pnas.1907367117.

Mao, Xingliang, Shuai Chang, Jinjing Shi, Fangfang Li & Ronghua Shi. 2019. Sentiment-aware word embedding for emotion classification. *Applied Sciences* 9(7). 10.3390/app9071334. https://www.mdpi.com/2076-3417/9/7/1334.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3111–3119.

Mohammad, Saif. 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)*, 174–184. Melbourne, Australia: Association for Computational Linguistics. 10.18653/v1/P18-1017. https://aclanthology.org/P18-1017.

Nair, Sathvik, Mahesh Srinivasan & Stephan C. Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense Knowledge. *Proceedings of the Cognitive Aspects of the Lexicon Workshop at the 28th International Conference on Computational Linguistics (COLING)* https://arxiv.org/abs/2010.13057.

Osgood, Charles Egerton, George J Suci & Percy H Tannenbaum. 1957. *The measurement of meaning* 47. University of Illinois press.

Pandia, Lalchand, Yan Cong & Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint:2109.12951* .

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. http://www.aclweb.org/anthology/D14-1162.

Picard, Rosalind W. 2000. *Affective computing*. MIT press.

Picard, Rosalind W. 2003. Affective computing: challenges. *International Journal of Human-Computer Studies* 59(1-2). 55–64.

Poliak, Adam, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White & Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 337–340. Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5441. https://aclanthology.org/W18-5441.

Rashkin, Hannah, Eric Michael Smith, Margaret Li & Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* .

Russell, James A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6). 1161.

Russell, James A. 2003. Core affect and the psychological construction of emotion. *Psychological Review* 110(1). 145.

Shaver, Phillip, Judith Schwartz, Donald Kirson & Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology* 52(6). 1061.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng & Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Suresh, Varsha & Desmond C Ong. 2021. Using knowledge-embedded attention to augment pretrained language models for fine-grained emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. IEEE.

Tao, Jianhua & Tieniu Tan. 2005. Affective computing: A review. In *International Conference on Affective Computing and Intelligent Interaction*, 981–995. Springer.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy & Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5446. https://aclanthology.org/W18-5446.

Warriner, Amy Beth, Victor Kuperman & Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4). 1191–1207.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang & Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392.

Wiedemann, Gregor, Steffen Remus, Avi Chawla & Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *Conference on Natural Language Processing* http://arxiv.org/abs/1909.10430.

Wilcox, Ethan, Roger Levy, Takashi Morita & Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. Brussels, Belgium: Association for Computational Linguistics. 10.18653/v1/W18-5423. https://aclanthology.org/W18-5423.

Zhang, Xiang, Junbo Zhao & Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems (NeurIPS)* 28.