2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 978-1-6654-5490-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACIIW57231.2022.10085993

Context-Dependent Deep Learning for Affective Computing

Varsha Suresh
Department of Computer Science
National University of Singapore
Singapore
varshasuresh@u.nus.edu

***Abstract*—Deep-learning models have been widely employed for recognizing emotions from various modalities. Yet these models face a number of challenges such as generalizing to different test conditions and predicting fine-grained emotions to name a few. One possible way to tackle these challenges is to provide these deep-learning models with additional *context* which can be in the form of domain knowledge from external knowledge sources or inherent task properties in the form of task-specific auxiliary losses. We hypothesize that incorporating *context* can better guide deep-learning models to look at the right features. In this extended abstract, we specifically focus on the problem of fine-grained emotion recognition using text-based data. We explore how to augment state-of-the-art NLP models with *context* to improve their performance in detecting fine-grained classes. We also discuss the implications of our research and future directions.**

***Index Terms*—Deep learning, Context-dependent models, Fine-grained classification, Emotion Recognition**

I. Introduction

Automatic emotion recognition systems are widely used in many commercial settings such as using emotions detected from job candidates' videos to make employment decisions and for designing chatbots that support mental health applications. As most of these applications directly impact people, these systems must be robust in the way they model and detect emotions.

Current emotion recognition systems rely on deep-learning-based models, where raw input is directly passed to pre-trained models such as transformers (in NLP) [3] or ResNet-based models (in Computer Vision) [4] that are fine-tuned to predict the target emotions. However, these systems face several challenges. For instance, take the case of text-based emotion recognition systems where numerous approaches have been proposed to date [5], [6], however, most of these works focus on detecting a limited number of emotions comprising of 6 to 8 emotions. Detecting fine-grained emotions involve distinguishing between highly semantically similar emotions that have very subtle variations between them such as sad vs. devastated. The ability to detect such nuanced emotions, akin to what we as humans experience in our daily lives, is critical to the development of AI chatbots that can offer more empathetic responses during interactions. How can we enable our existing models to generalize better and scale to such settings?

A possible way to achieve this is to augment existing models with additional knowledge regarding the task and/or input which complements the knowledge provided by the raw input data – which is generally the only source of learning aid for these models. We hypothesize that augmenting models with additional *context* can enable current deep-learning systems to be more aware of the task and/or input data thereby guiding them to learn the right features. *Context* can be acquired from external knowledge bases that provide domain knowledge or inherent task-specific properties such as inter-label relationships that can provide additional insights about the task/input to constructively improve the model's prediction. There are plenty of ways we can obtain relevant auxiliary information, ranging from straightforward ways to use already available emotion-specific knowledge bases such as lexicons, and knowledge graphs to more advanced ways of using evidence and theories from emotion research conducted in other fields such as psychology.

In this extended abstract, we focus on the task of fine-grained emotion recognition in the text domain. We hypothesize that using context-dependent deep learning models can help the model distinguish the subtle variations that exist in the presence of fine-grained emotions. Currently, pre-trained language models such as BERT are the state-of-the-art approaches in text-based classification including emotion recognition [5]–[7]. Therefore, we explore how to augment pre-trained language models with useful *context* which can improve their performance in detecting fine-grained classes. To summarise the contributions of the proposed thesis,

- In Study 1 [1], we examine how knowledge (in the form of lexicon knowledge) can improve the classification of fine-grained emotions.
- In Study 2 [2], we explore how innate task properties such as inter-label relationships can help in fine-grained classification tasks.
- In Study 3, we plan to explore the utility of context-dependent models in other domains in Affective Computing such as facial expression recognition.

A quick summary of our work towards fine-grained emotion recognition is depicted in Fig. 1. The rest of the paper consists of related work discussing text-based emotion recognition models with a specific focus on fine-grained classification and

978-1-6654-5490-2/22/$31.00 ©2022 IEEE

the relevant technical background related to our approaches. Next, we discuss our approaches in detail followed by the methodology. Finally, we conclude by elaborating on our contributions to the field of Affective Computing.

II. Background and Related Work

In this section, we first discuss related work on fine-grained classification specifically in the NLP domain. Then we discuss existing literature relevant to the technical background.

A. Fine-grained classification

Fine-grained classification is a challenging problem that has been explored in various domains such as object recognition [8], [9] and cancer detection [10], [11]. In NLP, detection of fine-grained classes has been explored in sentiment classification tasks where instead of binary sentiment classification (positive and negative), models are trained to distinguish between different levels of positive such as positive and very positive. Research has shown different ways to approach fine-grained sentiment classification. [12] used multi-task learning where they showed simultaneously predicting coarse- and fine-grained sentiment help in the performance of fine-grained classification. Another set of works modified pre-trained language models' pre-training objective by incorporating sentiment semantics [13] or by using auxiliary sentiment-related objectives such as sentiment word masking and sentiment word prediction [14] to improve fine-grained sentiment classification.

In our work, we focus on the fine-grained classification task of emotion recognition. Emotion recognition has been performed using architectures such as LSTMs, RNNs, and GRUs [3], [15]. Nowadays, fine-tuning pre-trained models such as BERT [16] have achieved state-of-the-art performance for downstream tasks and consequently, they have been used for emotion recognition as well [5]–[7]. A common limitation of these works, however, is that they focus on a small set of emotions such as the basic set of Ekman's emotions. One reason could be due to the lack of fine-grained emotion datasets but recently two works have focused on expanding the set of emotions. [17] introduced Empathetic Dialogues, which contains text conversations labeled with 32 emotion labels, and [18] introduced GoEmotions, which contains Reddit comments labeled with 28 emotion labels. While the introduction of these datasets has helped us explore fine-grained emotions, much work is needed in improving fine-grained emotions which has its uses in many practical scenarios such as chatbots [19].

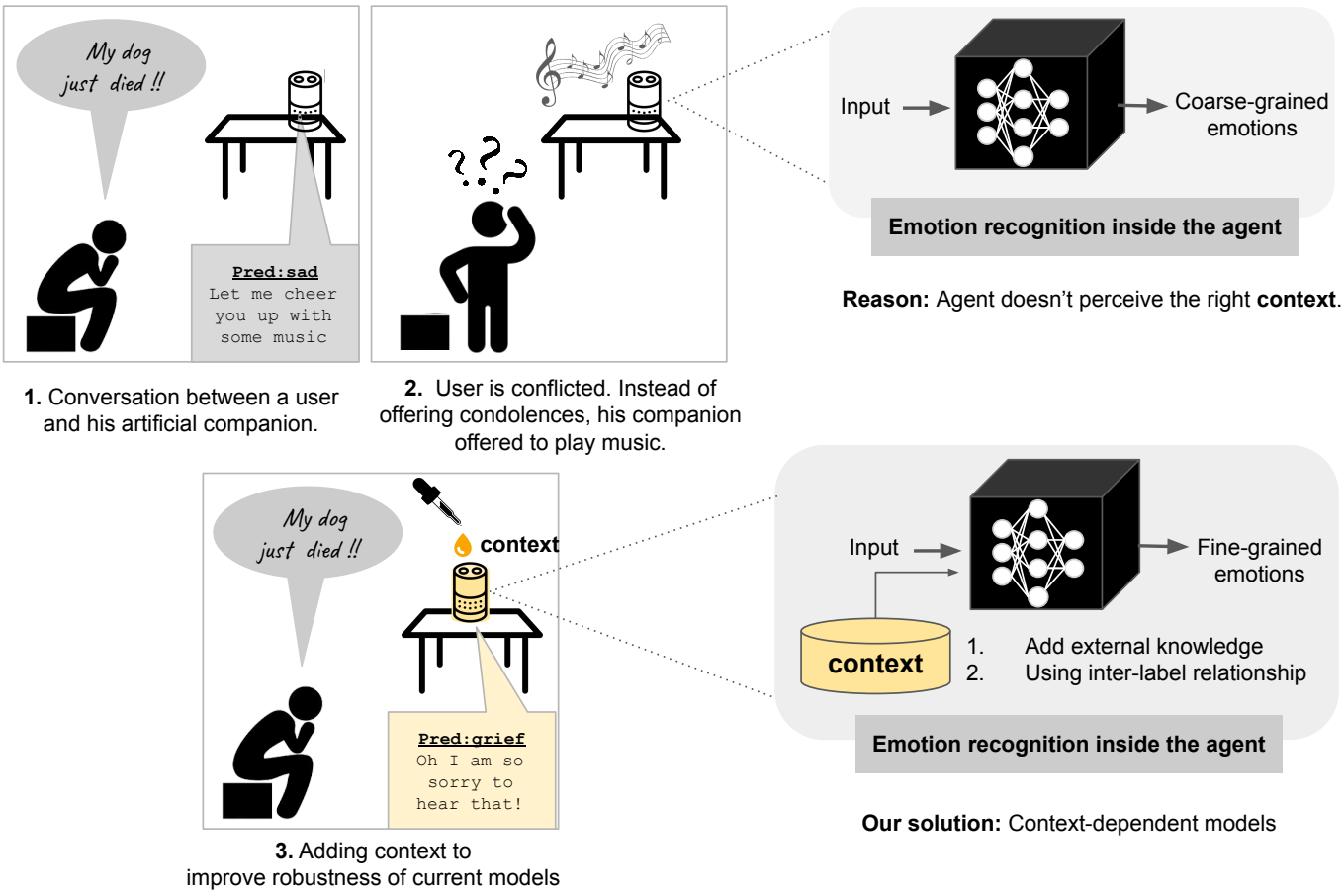

Fig. 1. Illustration of using context-dependant models for fine-grained emotion recognition. We hypothesize that incorporating additional *context* to our current models can help agents be more empathetic. We explore two ways of adding *context* for improving fine-grained emotion recognition, adding external knowledge [1] and inter-label relationships [2].

B. Knowledge-based approaches

Research has shown that incorporating external knowledge into deep-learning models help improve their performance [20]. Different ways such as concatenation [21] and attention-based mechanisms [22], [23] have been used to incorporate domain knowledge into standard architectures such as BiLSTMs, CNNs, and RNNs.

As we are particularly interested in augmenting pre-trained models, we focus the rest of this survey on works that integrate domain knowledge into representations learned by pre-trained language models. Some approaches augment the knowledge at the pre-training stage by changing the template of input to include the additional knowledge [24], [25] or by including extra task-specific objectives [14]. Another line of approaches augments the model's representation at the fine-tuning stage. In this set, some approaches fuse useful text-based information such as auxiliary sentences with the original raw input in the early stage [26], while other approaches combine the knowledge embedding with the pre-trained models' representation at the later stage via concatenation [21], [27] or by using separate using models to combined the knowledge embedding with the pre-trained representations [28].

We can see that combining the knowledge source at the fine-tuning stage helps add knowledge sources without the need to re-train the language model from scratch, given that the additional knowledge and the augmentation strategy does not lead to catastrophic forgetting of the pre-trained knowledge acquired by the language model. In our work, we aim to incorporate external emotion-specific knowledge in form of emotion lexicons [29] while fine-tuning the pre-trained language model.

C. Contrastive approaches

Context in knowledge-augmented approaches is heavily dependent on the availability and type of external knowledge. To reduce the reliance on external sources we explore the incorporation of *context* by tapping into the inherent task properties of fine-grained classification.

For this, we use contrastive learning. In traditional contrastive learning, direct comparison is done amongst the representations of the samples within a batch to pull the positives of the current sample closer and push the negatives farther. Positives denote samples with the same labels as the current sample, while the rest of the classes form their negatives. In fine-grained classification, there is a strong interplay amongst these classes, for instance, class sad is much closer to devastated than class angry. Our intuition suggests that including this property of inter-label relationships into existing contrastive objectives can help guide the model better in fine-grained classification.

Many computer vision approaches use contrastive learning particularly in self-supervised settings [30]–[32]. In the self-supervised version, the positive samples are the augmented versions of the same sample. Recently, [33] extended self-supervised contrastive loss to supervised settings where all samples with the same label as the current sample are regarded as positives.

Recently in NLP, the self-supervised contrastive loss has been used for pre-training language models such as BERT [34], [35]. [36] used a combination of cross entropy and supervised contrastive loss for fine-tuning pre-trained language models to improve performance in few-shot learning scenarios. In our work, we aim to fine-tune the pre-trained language models by augmenting contrastive loss with inter-label relationships between the target classes as *context* to improve fine-grained emotion recognition.

III. Approach

A. External Knowledge as Context

In the first work, we focus on improving fine-grained emotion recognition by using external knowledge sources as *context*. We introduce Knowledge-Embedded Attention (KEA) which uses knowledge from emotion lexicons to augment the contextual representations from pre-trained language models such as ELECTRA and BERT.

First, we embed an input text into two representations, the contextual representation from the last layer of the pre-trained language model and a latent emotional encoding obtained from an external knowledge source. We use lexicon data as an external knowledge source, which consists of association scores for emotional dimensions such as emotional intensity, valence, and arousal. [29].

Second, to combine both these representations we use an attention mechanism where the emotional information is used to attend to the contextual representations to construct a more emotionally-aware representation of the input. In [1] we evaluate the performance of KEA using different types of text-based emotion datasets ranging from tweets to conversations [17], [18], [37]. We find that KEA-incorporated models reduce the confusion between closely-confusable labels when compared to representative baselines. In addition, we also perform experiments to understand the generalisability of our approach to different types of pre-trained models and knowledge sources.

B. Incorporating Inter-label Relationship as Context

Next, we look at developing a more generalized approach, where we guide the model's learning using the inherent task property of inter-label relationships. We aim to embed these relationships into the contrastive objective to weigh the closely confusable classes such as angry and furious more than far-apart classes such as angry and sad.

Contrastive Loss (CL) aims to pull the positives closer together and push the negatives far apart. In the supervised version of contrastive loss, [33], all the samples belonging to the same class from a batch forms the positive set and the rest of the samples in the batch is the negative set. Let's denote the set of positives as $\mathcal{P}$ and the set of negatives as $\mathcal{N}$. Let us also denote a batch of sample and label pairs as $\{x_i, y_i\}_{i \in I}$, where $I = \{1, \cdots, K\}$ is the indices of the samples and K is the batch-size. Generally in contrastive losses, K samples and its corresponding K augmented data-points forms the batch, therefore the batch size is $2K$ and $I = \{1, \cdots, 2K\}$. The

positive set is given by $\mathcal{P} = \{p : p \in I, y_p = y_i \wedge p \neq i\}$. The supervised contrastive loss for a given sample x_i is given by:

$$L_{SCL_i} = \sum_{p \in \mathcal{P}} \log \frac{\exp(h_i \cdot h_p/\tau)}{\sum_{k \in \mathcal{I}/i} \exp(h_i \cdot h_k/\tau)} \quad (1)$$

Here, τ denotes the temperature hyper-parameter. From the above Eqn. 1 we can see that Supervised Contrastive Loss weighs negatives samples belonging to other classes equally to the current sample x_i. However, this does not happen in reality, especially in fine-grained classification, where classes such as sad are devastated are much closer than classes such as sad and angry.

We introduce Label-aware Contrastive Loss (LCL) where we aim to weigh the negative samples differently based on the label relationships between them and the current sample, thereby guiding the model to differentiate the harder negatives. For obtaining the weights, we use a simple weighting network which is another pre-trained language model trained simultaneously with the main network to predict emotion. The prediction probabilities obtained from the weighting network serve as the measure of the inter-label relationship between the current sample's class and the rest of the classes. In [2] we compare LCL with various representative baselines with four emotion datasets. We find LCL improves fine-grained classification, especially in presence of a large number of classes. To further test the efficacy of these models, we also carry out controlled experiments to understand the performance under varying difficulty levels of sets of fine-grained classes.

C. Extending Context-based Deep Learning Models to other modalities

In the above set of approaches, we primarily look at the problem of fine-grained classification in text-based emotion recognition. In the future, we aim to expand context-aware deep learning models to other modalities of emotion recognition such as facial expression recognition. A major challenge that current facial expression recognition models encounter is the issue of generalization to different test conditions. [38] showed that commercial facial expression recognition models are highly biased toward Black faces, similarly, several works highlight the generalisability issues of facial expression models across different datasets [39]. These issues primarily arise because the models learn irrelevant features such as skin color rather than task-relevant features such as facial landmarks. In our upcoming work, we aim to use *context* to tackle the ongoing challenge of generalisability in facial expression recognition.

IV. Conclusion and Expected Contributions

In this paper, we discussed our current research direction towards context-based deep learning models for emotion recognition. From our experiments and comparison with representative baselines in [1], we see that adding external knowledge to pre-trained language models improves emotion recognition, especially in fine-grained settings. In our second work, we re-think *context* by using the inherent property of inter-label relationships amongst fine-grained classes. In [2], we find that adding inter-label relationships into the fine-tuning objective of pre-trained language models helps improve fine-grained emotion recognition performance. We also perform a series of controlled experiments to test the efficacy of our approach by varying both the number of classes and the closeness between the classes. In our future work, we plan to expand context-dependent deep learning models to solve challenging problems in other modalities such as facial expression recognition.

We would like to end by re-iterating a thought experiment that is representative of the current deep-learning paradigm which is discussed in [40], [41]: If a model is trained to learn chess using the best movements indicated using red arrows, the model can either learn how to play chess or learn to follow the red arrows. Both approaches can work on games in this particular dataset but the first solution is the right way to learn chess. "With no external knowledge, the network typically learns the simpler solution" [40].

By incorporating *context*, we believe that we can guide deep-learning models to look at the right features, which is essential to building effective and robust emotion recognition models. Such models could help scale existing emotion predictions to challenging tasks such as fine-grained classification without the need for complex architectures and/or a large amount of data.

References

[1] V. Suresh and D. C. Ong, "Using knowledge-embedded attention to augment pre-trained language models for fine-grained emotion recognition," in *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII 2021)*, 2021.

[2] ——, "Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[3] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowl. Inf. Syst.*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[5] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu, and Y.-S. Chen, "EmotionX-IDEA: Emotion bert–an affectional model for conversation," *arXiv:1908.06264*, 2019.

[6] L. Luo and Y. Wang, "EmotionX-HSU: Adopting pre-trained bert for emotion classification," *arXiv:1907.09669*, 2019.

[7] K. Yang, D. Lee, T. Whang, S. Lee, and H. Lim, "EmotionX-KU: Bert-max based contextual emotion classifier," *arXiv*, 2019.

[8] X.-S. Wei, J. Wu, and Q. Cui, "Deep learning for fine-grained image analysis: A survey," *arXiv preprint arXiv:1907.03069*, 2019.

[9] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017.

[10] L. Li, X. Pan, H. Yang, Z. Liu, Y. He, Z. Li, Y. Fan, Z. Cao, and L. Zhang, "Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images," *Multimedia Tools and Applications*, vol. 79, no. 21, pp. 14 509–14 528, 2020.

[11] R. Qin, Z. Wang, L. Jiang, K. Qiao, J. Hai, J. Chen, J. Xu, D. Shi, and B. Yan, "Fine-grained lung cancer classification from pet and ct images based on multidimensional attention mechanism," *Complexity*, vol. 2020, 2020.

[12] G. Balikas, S. Moura, and M.-R. Amini, "Multitask learning for fine-grained twitter sentiment analysis," in *Proceedings of the 40th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 1005–1008.

[13] D. Yin, T. Meng, and K.-W. Chang, "SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3695–3706.

[14] H. Tian, C. Gao, X. Xiao, H. Liu, B. He, H. Wu, H. Wang, and F. Wu, "SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4067–4076.

[15] D. Zhang, W. Zhang, S. Li, Q. Zhu, and G. Zhou, "Modeling both intra- and inter-modal influence for real-time emotion detection in conversations," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, New York, NY, USA, 2020, p. 503–511.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[17] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5370–5381.

[18] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[19] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau *et al.*, "Recipes for building an open-domain chatbot," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 300–325.

[20] A. Roy and S. Pan, "Incorporating extra knowledge to enhance word embedding," in *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI-20*, 2020, pp. 4929–4935.

[21] L. De Bruyne, P. Atanasova, and I. Augenstein, "Joint emotion label space modelling for affect lexica," *arXiv:1911.08782*, 2019.

[22] K. Margatina, C. Baziotis, and A. Potamianos, "Attention-based conditioning methods for external knowledge integration," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 19, pp. 14–182.

[23] B. Shin, T. Lee, and J. D. Choi, "Lexicon integrated cnn models with attention for sentiment analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 149–158.

[24] N. Poerner, U. Waltinger, and H. Schütze, "E-BERT: Efficient-yet-effective entity embeddings for bert," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 803–818.

[25] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1441–1451.

[26] Z. Wu and D. C. Ong, "Context-guided BERT for targeted aspect-based sentiment analysis," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[27] N. Babanejad, H. Davoudi, A. An, and M. Papagelis, "Affective and contextual embedding for sarcasm detection," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[28] R. Wang, D. Tang, N. Duan, Z. Wei, X.-J. Huang, J. Ji, G. Cao, D. Jiang, and M. Zhou, "K-Adapter: Infusing knowledge into pre-trained models with adapters," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1405–1418.

[29] S. Mohammad, "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)*, 2018, pp. 174–184.

[30] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, 2020.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[32] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020.

[33] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[34] H. Fang and P. Xie, "CERT: Contrastive self-supervised learning for language understanding," *arXiv preprint arXiv:2005.12766*, 2020.

[35] Y. Meng, C. Xiong, P. Bajaj, S. Tiwary, P. Bennett, J. Han, and X. Song, "Coco-lm: Correcting and contrasting text sequences for language model pretraining," *arXiv preprint arXiv:2102.08473*, 2021.

[36] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," in *International Conference on Learning Representations*, 2021.

[37] S. Mohammad and S. Kiritchenko, "Understanding Emotions: A dataset of tweets to study interactions between affect categories," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[38] L. Rhue, "Racial influence on automated perceptions of emotions," *Available at SSRN 3281765*, 2018.

[39] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, 2020.

[40] M. Pezeshki, O. Kaba, Y. Bengio, A. C. Courville, D. Precup, and G. Lajoie, "Gradient starvation: A learning proclivity in neural networks," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[41] G. Parascandolo, A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf, "Learning explanations that are hard to vary," in *Ninth International Conference on Learning Representations (ICLR 2021)*, 2021.