Investigating Contextual Representations for fine-grained Emotion Classification

Anirudh Sundara Rajan *
University of Wisconsin – Madison
asundararaj2@wisc.edu

Karthik Suresh *
University of Wisconsin – Madison
ksuresh6@wisc.edu

Abstract

Multi-label emotion classification involves assigning one or more emotion categories to each input. Fine-grained emotion classification deals with classifying emotions from an input sentence (or any other modality) into one of many categories. This study uses Seq2Emo as the baseline and examines how various contextual embeddings affect the results. Seq2Emo uses ELMo embeddings derived using a BiLSTM trained using the next word prediction objective. With the advent of Transformer-based representation models such as BERT that uses techniques such as MLM (Masked Language Modelling), we wanted to investigate the effect of incorporating such models in tasks involving fine-grained emotion classification. Using RoBERTa-large, we run tests to demonstrate a significant improvement over the baseline. We finally go over potential future work for the final report and discuss techniques we can incorporate to potentially gain significant performance improvements.

1 Introduction

Emotion detection and classification are starting to become an important part of Machine learning research with varied inputs from different modalities. Audio (Soleymani et al., 2013), Visual (Randhavane et al., 2019; Javadi and Lim, 2021), Audio-Visual (Livingstone and Russo, 2018), and Textual (Chen et al., 2018; Demszky et al., 2020) inputs are some of the important modalities where researchers hope to capture behavioral cues and perform appropriate emotion classification.

With the increased use of the internet and, in turn, social media platforms, there is an increased need to leverage automated methods for the classification of textual data. Emotion detection, in particular, can be used to detect hate speech, cyberbullying, etc., and combat them effectively. Businesses are also adopting these methods to survey the large corpus of customer feedback/data they receive.

Traditional emotion classification was treated as a multi-class classification task (Scherer and Wallbott, 1994; Mohammad, 2012), where each input sentence belongs to one and only one class (i.e., emotion). Recent works (Demszky et al., 2020; Mohammad et al., 2018) have argued for multi-label classification of emotions where each data instance may have one or more target emotions. This can be seen to make sense as a particular sentence may exhibit multiple emotions at once (ex: 'confusion' and 'curiosity')

Most notable traditional works on categorizing emotion include Ekman's six primary categories (Ekman, 1992) and eight primary emotions in Plutchik's wheel of emotions (Plutchik, 1980). Works such as CrowdFlower (2016) have extended this work by curating a corpus of 40k tweets, with each one having its label as one of 13 emotions. The need for fine-grained analysis of textual data coupled with labeled corpora that aided in multi-label classification prompted more recent advancements, such as the development of the GoEmotions dataset (Demszky et al., 2020), which has 27 emotion categories labeled for 60k Reddit comments. Datasets such as EmoInt (Mohammad et al., 2018) were also curated to serve a similar purpose.

For this work, we aim to build on the Seq2Emo architecture (Huang et al., 2021) based on certain intuitions to improve its performance. Seq2Emo is a Seq2Seq-like framework that encodes the input sentence with an LSTM and utilizes a Bidirectional LSTM as a decoder. Hidden states are calculated as a concatenation of GloVe embeddings (Pennington et al., 2014) and ELMo contextualized embeddings (Peters et al., 2018).

Transformer-based models (Vaswani et al., 2017) have been used as building blocks to obtain state-of-the-art performance in numerous downstream tasks previously dominated by Traditional Deep

*Denotes Equal contribution

learning architectures. Recurrent networks that were used to solve NLP tasks suffer from a Lack of parallelizability of operations and limited ability to capture long-range dependencies. This was ameliorated to a good extent by the new Transformer models. Contextual Word embeddings derived from transformer-based representation models such as BERT (Devlin et al., 2018) proved to have improved representation power compared to traditional static embedding methods that came before it.

In this work, we leverage the above facts and perturb the embeddings used to see if improved representation power reflects in the results and by how much. We follow this up with detailed analyses involving the tested models. Toward the end, we also list the techniques that we hope to explore to gain further performance improvements.

2 Related Work

2.1 Emotion models

The most popular model of emotion is often cited as Paul Ekman's categorization (Ekman, 1992) of emotion into six basic and discrete categories - happiness, sadness, anger, disgust, surprise, and fear. These emotions are posited to be independent of each other and can combine to produce more complex combinations. The Plutchik model (Plutchik, 1980) argued that there were actually eight primary emotion categories that exist in opposite pairs but still held onto some of the postulates of Ekman's model, like the idea of primary emotions combining to form more complex ones.

More recent works like The Hourglass of Emotions revisited model (Susanto et al., 2020) showed the highest scores when tested with Blitzer, Pang and Lee, and Amazon datasets. Demszky et al. (2020) take a different approach by using Principal Preserved Component Analysis (Cowen et al., 2019) and show numerical evidence for categorizing emotion into 27 categories.

2.2 Datasets

Affective Text (Strapparava and Mihalcea, 2007) was one of the first textual datasets in emotion classification that provided 250 news headlines to be categorized into one of seven emotion categories. Social media platforms have become a popular source for these datasets, with data instances curated and annotated from websites like Twitter (Wang et al., 2012; Abdul-Mageed and Ungar, 2017) and Reddit (Demszky et al., 2020). Datasets reflecting emotion are also curated from short-form dialogues (Chen et al., 2018), movie subtitles (Öhman et al., 2018), or self-reported experiences (Scherer and Wallbott, 1994)

2.3 Methods and Techniques

Seq2Emo (Huang et al., 2021) utilizes an encoder-decoder architecture to classify emotion categories. They use a concatenation of static GloVe (Pennington et al., 2014) word embeddings and ELMo contextualized embeddings (**?**). The hidden states from the encoder are calculated using an attention mechanism (luo) and are passed to the decoder, which has a BiLSTM architecture. Learnable emotion embeddings are passed to each step of the decoder, and the model is to perform binary classification on each one of these emotion categories. Outputs of previous units of the decoder are not fed into the future ones so as to avoid exposure bias (Bengio et al., 2015).

Sentiment Knowledge Enhanced Pre-training (SKEP) (Tian et al., 2020) is a pretraining objective proposed to incorporate sentiment knowledge through self-supervised training. This domain-specific training paradigm consists of two parts: (1) Sentiment masking: Based on automatically mined sentiment knowledge, they first recognize the sentiment information from the input sentence and then produce a corrupted version by removing this information. (2) They then propose three sentiment pre-training objectives requiring transformers to recover clean information about the sentiment from the corrupted version.

SentiBERT (Yin et al., 2020) is a variant of BERT that sought to predict emotion categories by capturing compositional sentiment semantics. Using a binary constituency parse tree in combination with contextualized representations captured using models like BERT, they were not only able to perform emotion classification on datasets like SST (Socher et al., 2013) but were also able to use the compositional sentiment semantics learned to tackle associated tasks successfully. Other methods have been proposed that use Span-prediction (Alhuzali and Ananiadou, 2021), Graph based networks (Xu et al., 2020), Domain Specific pre-training objectives (Sosea and Caragea, 2021), etc.

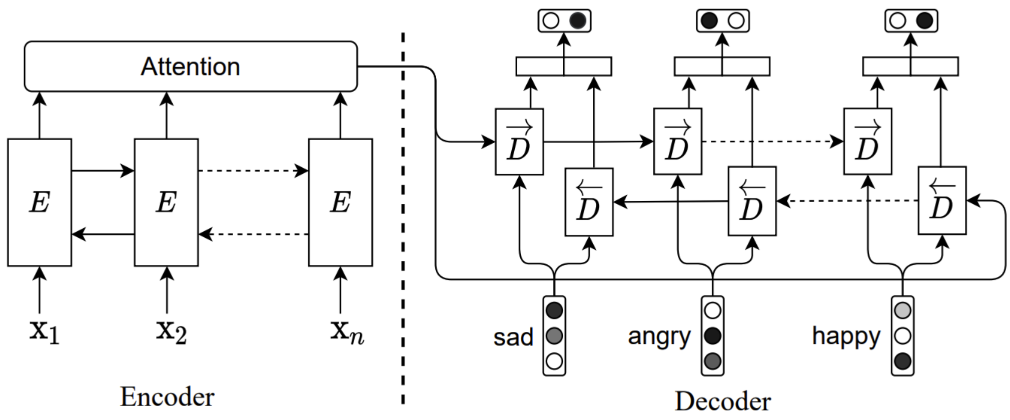

Figure 1: Model Architecture that we use as adapted from Huang et al. (2021)

3 Background: ELMo, BERT, and RoBERTa

ELMo (Embeddings from Language Models) (Peters et al., 2018) was one of the first to introduce the idea of contextual word embeddings as opposed to static embeddings that existed previously. This gave stronger representation power as each word token can have different meanings depending on context. They utilized a BiLSTM to perform Language modeling. What this means is that they fed in unlabeled data containing large amounts of sentences and trained it for next-word prediction i.e., predict the next word in a sequence of words.

Following the effectiveness of Transformers (Vaswani et al., 2017) and taking inspiration from ELMo, BERT (Devlin et al., 2018) was introduced as a self-supervised learning approach that pretrains a transformer encoder for it to be used on a variety of downstream tasks. The training objective that makes this paradigm powerful is termed Masked Language Modelling (MLM), and it relies on the availability of a large corpus of unlabelled data for the task at hand. MLM samples (randomly) parts of the input sentence to be substituted. It uniformly samples 15% of the input tokens to be substituted. Of these sampled tokens, 80% are replaced with a special [MASK] token, 10% are replaced with a random token, and the rest 10% are left unchanged. The task of the encoder is to predict the words in these slots and in, by doing so, build strong contextual representations of the tokens in the corpus. The architecture can then be fine-tuned with labeled data to solve other downstream tasks. The pretraining input to BERT was two continuous chunks of text and the task was to predict whether one chunk of text immediately followed the other. this was termed as Next Sentence Prediction (NSP) task.

RoBERTa (Liu et al., 2019) improves upon BERT by removing NSP from its pretraining approach and dynamically change the masked tokens from epoch to epoch. The encoder also models much more data than in the original BERT paper. This resulted in superior performance as compared to BERT.

4 Methodology

Problem statement: In this work, we are dealing with multi-label emotion classification. That is, each data instance has associated with one or more emotion labels. More formally, if we have K predefined candidate emotions, we can represent the target label for each input sentence **x** as **y** = $(y_1, ..., y_K) \in \{0, 1\}^K$. $y_i = 1$ implies that the ith emotion is "on" for that particular input sentence, and $y_i = 0$ means that the emotion is "off".

We are making modifications to the Seq2Emo architecture, which are described in detail below:

Encoder: On the encoder side, we use a combination of static and contextual word embeddings to capture the meaning of the input sentence. The static embedding used is GloVe (Pennington et al., 2014), and the contextual embeddings are one of ELMo, BERT, RoBERTa. ELMo uses a BiLSTM to encode the representations, while BERT and RoBERTa use a Transformer based architecture for the same.

$$\overrightarrow{h_t^E} = LSTM([\text{GloVe}(x_t); M(\mathbf{x})_t], \overrightarrow{h_{t-1}^E})$$
$$\overleftarrow{h_t^E} = LSTM([\text{GloVe}(x_t); M(\mathbf{x})_t], \overleftarrow{h_{t-1}^E})$$
$$h_t^E = [\overrightarrow{h_t^E}; \overleftarrow{h_t^E}]$$

Where $M \in \{\textbf{ELMo}, \textbf{BERT}, \textbf{RoBERTa}\}$ and h_t^E is the hidden state output from the encoder to the decoder

Decoder: The decoder is LSTM based and will be used to make sequential predictions on candidate emotions. Learnable embeddings for the emotion categories are fed into each step of the decoder, and binary classification (0 (“off”) or 1(“on”)) is performed for each emotion category. The learnable embeddings help the decoder in deciding what emotion to predict at that particular step. The input to each decoder step is calculated as the attention-weighted sum of the encoder hidden states concatenated with the decoder hidden state at that step.

5 Experiments and Results

5.1 Experimental Setup

We perform our experiments on the same setup as the Seq2Emo paper. Experiments are run on an **NVIDIA GeForce RTX 2080 Ti**. A batch size of 32 is used. We employ 5-Fold Cross Validation to assess the approach in a manner similar to the initial setup. On the GoEmotions dataset, we run the tests and evaluate results based on standard metrics such as **Precision, Recall, F1-Score**, and **Jaccard Coefficient**.

5.2 Results

Table 1 contains the results, which are presented. **Micro/macro** averages are used to give the precision, recall, and F1 scores. We outline each approach’s breakdown by emotion in the parts that follow.

The size of the contextual embedding for the "base" models is 768, whereas it is 1024 for the **ELMo** and "large" models. As demonstrated, despite having a relatively low-dimensional representation, the **RoBERTa-base** representations aid the model in producing results that are extremely similar to the baseline. Despite having the same number of dimensions, **RoBERTa-large** performs significantly better than the baseline, demonstrating the effectiveness of transformer-based encoders. We will shortly run tests with domain-specific encoders like **SKEP** and **SentiBERT**, and we anticipate even bigger improvements.

5.3 Analysis

We compute the emotion-wise precision, recall, and F1-score to examine the areas where each method falls short. These results for the basic Seq2Emo with the ELMo contextual representations are shown in Table 2. The F1-score is unsurprisingly low for emotions with low support (such as grief, pride, and relief), but it should be noted that despite having a relatively high support, abstract emotions like caring, confusion, and curiosity have extremely low F1-score. Furthermore, it seems to perform better on positive emotions in general.

Additionally, when we use **RoBERTa-large** as the text encoder (refer to Table 3) to compare and contrast the baseline with the model’s emotion-wise scores, we find that there is a roughly 10% rise in emotions like anger, caring, curiosity, fear, and nervousness. While some of the more abstract emotions show significant improvement, the overall accuracy attained is still not very good and requires improvement. Additionally, we observe that **Seq2Emo + RoBERTa-large** obtains an F1-score of zero for the emotion "relief," indicating that it completely misses that particular emotion. As a result, even though we get superior results than the baseline, **Seq2Emo + RoBERTa-large** still does not completely understand the subtle differences between certain emotions. We have not included the emotion specific metrics for the other 2 cases due to space issues, these can be viewed in the github repository.

6 Future Work

6.1 Further Extensions

Seq2Emo + RoBERTa-large considerably outperforms the baseline, but the encoders of these models, which were trained on the open domain, still have a poor understanding of the subtle characteristics that characterize each emotion. We contend that the absence of a sentiment-dependent goal prevents language models like BERT, RoBERTa, etc. from accurately capturing the subtle emotional information present in text.

Encoder	Precision	Recall	F1	Jaccard
`ELMo (baseline)`	`0.6584/0.6642`	`0.5350/0.4036`	`0.5903/0.4740`	`0.5415`
`BERT-base`	`0.6597/0.6120`	`0.4983/0.3489`	`0.5678/0.4094`	`0.5084`
`RoBERTa-base`	`0.6768/0.6445`	`0.5295/0.3934`	`0.5941/0.4528`	`0.5374`
`RoBERTa-large`	`0.6762/0.6467`	`0.5401/0.4276`	`0.6005/0.4822`	`0.5437`

Table 1: Ablation on **Seq2Emo** using different types of text encoders. Precision, recall and F1 score are reported as both micro/macro averages.

Emotion	Precision	Recall	F1-score	Support
admiration	0.7086	0.6369	0.6708	504
amusement	0.7944	0.8636	0.8276	264
anger	0.6395	0.2778	0.3873	198
annoyance	0.5577	0.1812	0.2736	320
approval	0.5283	0.2393	0.3294	351
caring	0.5000	0.2148	0.3005	135
confusion	0.5652	0.2549	0.3514	153
curiosity	0.5181	0.3521	0.4193	284
desire	0.6875	0.3976	0.5038	83
disappointment	0.5610	0.1523	0.2396	151
disapproval	0.5246	0.2397	0.3290	267
disgust	0.5930	0.4146	0.4880	123
embarrassment	0.8571	0.3243	0.4706	37
excitement	0.6512	0.2718	0.3836	103
fear	0.7143	0.5769	0.6383	78
gratitude	0.9631	0.8892	0.9247	352
grief	1.0000	0.1667	0.2857	6
joy	0.6522	0.5590	0.6020	161
love	0.7816	0.8571	0.8176	238
nervousness	0.5556	0.2174	0.3125	23
optimism	0.7477	0.4462	0.5589	186
pride	0.7143	0.3125	0.4348	16
realization	0.6667	0.1241	0.2093	145
relief	0.5000	0.0909	0.1538	11
remorse	0.5965	0.6071	0.6018	56
sadness	0.7419	0.4423	0.5542	156
surprise	0.6702	0.4468	0.5362	141
neutral	0.6068	0.7443	0.6685	1787

Table 2: Emotion-wise precision, recall and F1-scores for the baseline Seq2Emo model. The model struggles with many abstract emotions and does not perform well on negative/neutral emotions, according to F1-scores.

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
admiration	0.7152	0.6429	0.6771	504
amusement	0.7952	0.8826	0.8366	264
anger	0.5923	0.3889	0.4695	198
annoyance	0.6500	0.1625	0.2600	320
approval	0.5950	0.2051	0.3051	351
caring	0.6066	0.2741	0.3776	135
confusion	0.5570	0.2876	0.3793	153
curiosity	0.4770	0.5845	0.5253	284
desire	0.5714	0.3855	0.4604	83
disappointment	0.6207	0.1192	0.2000	151
disapproval	0.5979	0.2172	0.3187	267
disgust	0.6875	0.3577	0.4706	123
embarrassment	0.7500	0.4054	0.5263	37
excitement	0.6279	0.2621	0.3699	103
fear	0.7037	0.7308	0.7170	78
gratitude	0.9503	0.8693	0.9080	352
grief	1.0000	0.1667	0.2857	6
joy	0.6216	0.5714	0.5955	161
love	0.7824	0.8613	0.8200	238
nervousness	0.5833	0.3043	0.4000	23
optimism	0.7008	0.4785	0.5687	186
pride	0.7143	0.3125	0.4348	16
realization	0.7143	0.1034	0.1807	145
relief	0.0000	0.0000	0.0000	11
remorse	0.5479	0.7143	0.6202	56
sadness	0.6047	0.5000	0.5474	156
surprise	0.6800	0.4823	0.5643	141
neutral	0.6617	0.7029	0.6817	1787

Table 3: Emotion-wise precision, recall and F1-scores for the Seq2Emo model with **RoBERTa-large** text encoder. In most instances, the **RoBERTa-large** text encoder works significantly better than the baseline, but the F1-scores for a few emotions are still not very high, and the model comprehensively fails at predicting the emotion "relief".

To close this gap in the case of fine-grained sentiment analysis, we think language models that have already been trained on sentiment analysis could be used as the encoder. Our primary goal is to examine the performance of the transformer-based models SKEP and SentiBERT. SentiBERT, learns to predict local emotion at various nodes in the constituency tree, whereas SKEP employs emotional polarity prediction and the aspect-sentiment pair prediction as training tasks. The more nuanced emotional details in language must therefore be captured by these specialized language models.

So far, in order to gain an understanding of different challenges in fine-grained emotion classification, we performed ablation studies by simply switching out the representations. Additionally, we believe that fine-tuning these language models rather than just using the representations could further improve performance.

6.2 Work Delegation

The aforementioned tests will be carried out before the final report, as planned. Studying the effects of different representations in the context of this task is another one of the main objectives of this project. In order to determine what kind of emotional information is missed by these models and why, we first plan to perform experiments using the SentiBERT and SKEP language models. Then, we'll visualize attention maps for various examples. Datasets also suffer from an imbalance in terms of the number of data instances in each class and we hope to investigate methods to learn good classifications for even these emotion categories.

For the above set of tasks, **Anirudh Sundara Rajan** will perform experiments and collect the results from them and **Karthik Suresh** will work on Error and Data analysis of the results and look for areas of improvement. The tasks above are a very broad generalization of the work yet to be done and there may be significant overlap in the tasks that we choose to do. Apart from the above, we also wish to try out some domain-specific ideas that have worked out for related tasks and see if they would hold water in fine-grained emotion classification. The breadth that we choose to explore in this regard is yet to be explored by us and we will leave that as work for the final report. In addition, we also intend to fine-tune all the mentioned language models for the task in focus. We both will be working on this task and setup the experimental setup for this.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3(4):369–382.

CrowdFlower. 2016. Crowdflower.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

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*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar R Zaiane. 2021. Seq2emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4717–4724.

Roya Javadi and Angelica Lim. 2021. The many faces of anger: A multicultural video dataset of negative emotions in the wild (mfa-wild). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Saif Mohammad. 2012. # emotional tweets. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Emily ¨Ohman, Kaisla Kajava, J¨org Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. 2019. Learning perceived emotion using affective and deep features for mental health applications. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 395–399. IEEE.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and 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*, pages 1631–1642.

Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 2013. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6.

Tiberiu Sosea and Cornelia Caragea. 2021. emlm: A new pre-training objective for emotion related tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 286–293.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.

Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter"" big data"" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2020. Emograph: Capturing emotion correlations using graph networks. *arXiv preprint arXiv:2008.09378*.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.