Prompt Distillation for Emotion Analysis

Andrew L. Mackey, Susan Gauch and Israel Cuevas
Deparment of Electrical Engineering and Computer Science, University of Arkansas , Fayetteville, Arkansas, U.S.A.
{*almackey, sgauch, ibcuevas*}*@uark.edu*

Keywords: Emotion Analysis, Natural Language Processing.

Abstract: Emotion Analysis (EA) is a field of study closely aligned with sentiment analysis whereby a discrete set of emotions are extracted from a given document. Existing methods of EA have traditionally explored both lexicon and machine learning techniques for this task. Recent advancements in large language models have achieved success in a wide range of tasks, including language, images, speech, and videos. In this work, we construct a model that applies knowledge distillation techniques to extract information from a large language model which instructs a lightweight student model to improve its performance with the EA task. Specifically, the teacher model, which is much larger in terms of parameters and training inputs, performs an analysis of the document and shares this information with the student model to predict the target emotions for a given document. Experimental results demonstrate the efficacy of our proposed prompt-based knowledge distillation approach for EA.

1 INTRODUCTION

Sentiment analysis (SA) is a prominent subfield of natural language processing (NLP) with the goal of analyzing text documents from which the document's polarity is obtained. Emotion analysis (EA) establishes additional granularity for classes beyond polarity from SA by focusing on the alignment of language with various emotional categories. For example, the Paul Ekman model for emotions defines six primary emotion categories: anger, disgust, fear, joy, sadness, and surprise (Ekman and Friesen, 1971). Another approach to illustrate the various emotional dimensions was proposed as the Robert Plutchik model with eight primary bipolar emotions: anger versus fear, joy versus sadness, anticipation versus surprise, and trust versus disgust (Plutchik, 1982). Additional models have been proposed that projects emotions into a dimensional space, such as for valence, arousal, and dominance (Russell and Mehrabian, 1977).

Various techniques have been proposed for the task of emotion analysis. The first major area of emotion analysis involves lexicon-based techniques where the techniques are focused on aligning the emotional categories of language with the specific words that were used (Baccianella et al., 2010) (Staiano and Guerini, 2014). The next major area of emotion analysis includes various machine learning techniques that discover latent patterns or representations for the detection of different emotional categories (Agrawal and An, 2012) (Calefato et al., 2018) (Hasan et al., 2019). Some researchers have investigated emotion representations that seek to achieve emotion representations that transcend multiple lexicons and datasets (Buechel et al., 2020). Some work in emotion classification has concentrated on aligning transformer-based architectures with emotional categories through deep contextual representations. Pretrained language models (PLM) have demonstrated various successes in outperforming many state-of-the-art techniques in the field. As the parameters and training data continued to scale for PLMs, large language models (LLM) emerged and demonstrated capabilities not seen in prior work, such as prompt-based learning and reasoning.

In this paper, we introduce a prompt-based knowledge distillation model for emotion analysis where the prompt serves as source of knowledge through which we distill that information for a student model under the supervision of a much larger teacher model. The first phase of our model involves a prompt-based teacher model followed by a knowledge distillation student training model. The teacher model uses prompt-based techniques to extract information from the LLM. The student model uses a transformer-based PLM where probabilities from both teacher and student models are aligned so that the student model is capable of generating similar probability distributions as the teacher model.

Mackey, A., Gauch, S. and Cuevas, I.
Prompt Distillation for Emotion Analysis.
DOI: 10.5220/0012951200003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 328-334
ISBN: 978-989-758-716-0; ISSN: 2184-3228

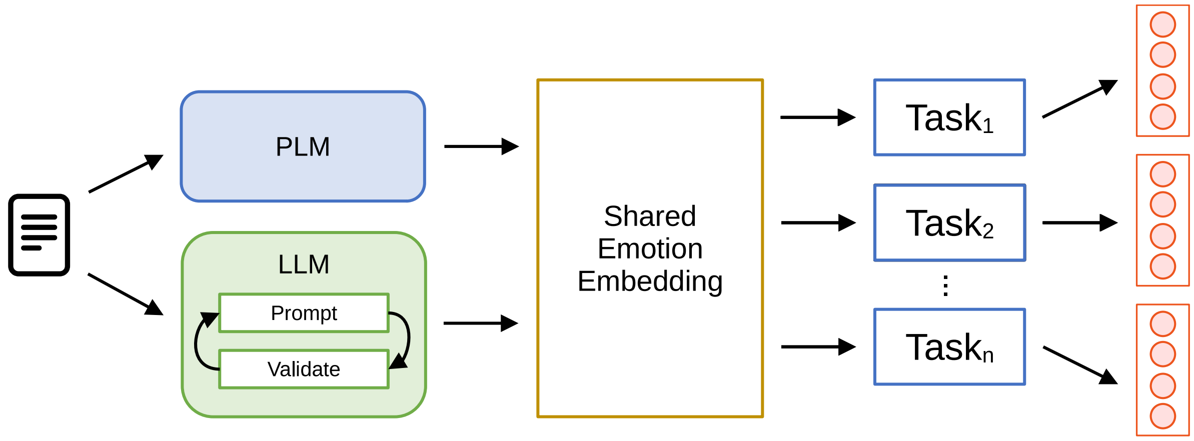

Figure 1: Overview architecture of the model. We combine a pre-trained language model with a large language model to extract the emotion embeddings cross-corpus to perform a classification of the emotions. For the final prediction y, we localize the classification head to a set of possible classes for the respective datatset.

2 RELATED WORK

Recent work in the research community has focused on tasks involving emotion analysis has concentrated primarily on PLMs for learning contextual representations using neural networks (Demszky et al., 2020) (Turcan et al., 2021) (Alhuzali and Ananiadou, 2021) (Wullach et al., 2021) (Mackey et al., 2021) (Toraman et al., 2022) (Rahman et al., 2024). PLMs undergo various training methods which enables them to learn latent contextual representations of text. These models are generally fine-tuned in order to adapt to task-specific objectives, such as emotion classification. Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based architecture that bidirectionally encoded embeddings to learn contextual information in textual data where the model was pre-trained simultaneously on the tasks of masked language modeling (MLM) and next sentence prediction (NSP) (Devlin et al., 2019). XL-Net improves upon BERT by introducing permutation language modeling where tokens are predicted in a random order (Yang et al., 2019). RoBERTa improved upon BERT by modifying the training approach where the NSP task was removed and dynamic masking was introduced, and increasing the amount of training data that was used (Liu et al., 2019).

Other work with LMs has resulted in different techniques for training methodologies. Knowledge distillation techniques, where a teacher model transfers knowledge from a complex model to a much simpler model, how shown promising results across different studies (Hinton et al., 2015) (Lukasik et al., 2022). Brown et al. define various levels of data used for in-context learning, such as fine-tuning (updating weights of a pretrained model), few-shot (models are provided a few demonstrations of a task with no additional weight updates to the model), one-shot (only one demonstration is permitted), and zero-shot (no demonstrations are permitted) (Brown et al., 2020). Brown et al. also demonstrate that as LMs increase in scale, their task-agnostic few-shot performance also increases (Brown et al., 2020). In addition, Halder et al. acknowledged that tranformer-based LMs fine-tuned to task-specific objectives curtail their ability to perform well in zero-shot, one-shot, or few-shot scenarios (Halder et al., 2020).

Work involving large language models continues to demonstrate their task-agnostic capabilities. One study demonstrated a technique of applying a series of reasoning steps named *chain of thought* where an LLM utilized chain-of-thought prompting that demonstrated reasoning abilities provided the LLM is adequately large (Wei et al., 2022). Adversarial distillation frameworks have also been proposed in research literature for improved knowledge distillation and transfer learning (Jiang et al., 2023).

3 PROBLEM DEFINITION

Let $\mathcal{D}$ represent a dataset comprised of N documents, where each document in $\mathcal{D}$ consists of textual information and emotion labels. We observe the following for each $\mathcal{D}$: **(1)** the set of text documents in dataset $\mathcal{D}$ is represented as $\mathcal{X}_\mathcal{D}$ such that $|\mathcal{X}_\mathcal{D}| = N$; **(2)** the set of possible target labels for dataset $\mathcal{D}$ is represented as $\mathcal{Y}_\mathcal{D}$ where $|\mathcal{Y}_\mathcal{D}| = C$ different emotions; and **(3)** $\mathcal{D}$ is represented as the following set in the single-label setting:

$$\mathcal{D} = \{(x,y) \mid x \in \mathcal{X}_D \text{ and } y \in \mathcal{Y}_D\} \tag{1}$$

and the following serves as the representation for a multi-label setting:

$$\mathcal{D} = \{(x,y) \mid x \in \mathcal{X}_D \text{ and } y \in \mathcal{P}(\mathcal{Y}_D)\} \tag{2}$$

Let $\mathbf{D}$ represent the input text corpora where $\mathbf{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$. The task presented in this work is to train and align a model to recognize the latent emotion representations in a cross-corpus setting using $\mathbf{D}$ for the purpose of single-class and multi-class emotion classification of an emotion label (or set) y from a given input document x:

$$\hat{\mathbf{y}} = \arg\max_c \left[\Pr(y = c \mid x; \Theta) \right] \tag{3}$$

4 PROPOSED APPROACH

We present our proposed solution in this section for the single-class and multi-class cross-corpora emotion classification task. In Figure 2, we provide an overview of our framework for learning the latent emotion distribution of text documents. There are three major components to our approach: **(1)** a prompt-based knowledge distillation paradigm for extracting information from an LLM to facilitate the alignment of a task-specific model; **(2)** a task-specific, emotion classification model that leverages a pre-trained, transformer-based language model, which is fine-tuned for the emotion classification task; and **(3)** a cross-corpora framework for learning latent emotion representations.

4.1 Prompt-Based Methodology

For a given dataset $\mathcal{D} = (\mathcal{X}_D, \mathcal{Y}_D)$, each input and target is represented as $(\mathbf{x}^{\text{doc}}, y^{\text{emo}})$ such that $(\mathbf{x}^{\text{doc}}, y^{\text{emo}}) \in \mathcal{D}$. The target y^{emo} of the model is the emotion class for each document where $y^{\text{emo}} \in \mathcal{Y}_D$ (i.e. anger, grief, disgust, etc.) in the respective dataset $\mathcal{D}$. To facilitate knowledge distillation from an LLM, we define $(\mathbf{x}^{\text{prompt}}, \mathbf{y}^{\text{llm}})$ to represent the prompt-based input and label generated from an LLM for each $(\mathbf{x}^{\text{doc}}, y^{\text{emo}}) \in \mathcal{D}$.

Prompt Template. The following template is used to the generate each $\mathbf{x}^{\text{prompt}}$:

You will be given a human written sentence. Classify the sentence into one of the following categories: $\langle \mathbf{y}_0^{emo}, \mathbf{y}_1^{emo}, ...\rangle$. Return the following format only for each category as a probability distribution (the sum should be 1): $\langle \mathbf{y}_i^{emo}$, probability$\rangle$.

The following is the document: $\mathbf{x}_i$.

The target $\mathbf{y}^{\text{llm}}$ represents the emotion distribution produced by the LLM for the given input prompt $\mathbf{x}^{\text{prompt}}$, which is modeled as follows:

$$\hat{\mathbf{y}}_i^{\text{llm}} = \Pr(y^{\text{emo}} \mid \mathbf{x}_i^{\text{prompt}}) \tag{4}$$

$$= \text{LLM}(\mathbf{x}^{\text{prompt}}) \tag{5}$$

Hallucinations are a known problem in research literature where an LLM produces a response that is either factually incorrect or unaligned with the input prompt it was provided (Farquhar et al., 2024). To address the problem of hallucinations, we conduct a validation step for $\hat{\mathbf{y}}^{\text{llm}}$ to ensure the format of the output is aligned with the targets in the training data. Documents failing the validation step will undergo a fixed interval of reprompting where the input and interactions are returned to the LLM for further processing in the form:

$$\hat{\mathbf{y}}^{\text{llm}'} = \text{LLM}(\ \langle \mathbf{x}^{\text{prompt}'}, \langle \mathbf{x}^{\text{prompt}}, \mathbf{y}^{\text{llm}} \rangle \rangle\ ) \tag{6}$$

4.2 Emotion Classification Model

The task-specific emotion classification model begins by employing the use of a transformer-based language model to provide contextual representations $\mathbf{h}^{\text{emo}} = \langle \mathbf{h}_1^{\text{emo}}, \mathbf{h}_2^{\text{emo}}, ..., \mathbf{h}_k^{\text{emo}} \rangle$ for input tokens $\mathbf{x}^{\text{doc}}$ where k represents the number of time steps. The transformer-based encoder LM is parameterized with ϕ for all datasets $\mathcal{D} \in \mathbf{D}$ to generate the contextualized word representations $\mathbf{h}_i^{\text{emo}}$ for each time step i:

$$\mathbf{h}_i^{\text{emo}} = \text{LM}_\phi(\mathbf{x}_i^{\text{doc}}) \tag{7}$$

The last layer of $\mathbf{h}_i^{\text{emo}}$ is used to compute the distribution for the emotion classes, where it is parameterized by ϕ_d for each $\mathcal{D}_d \in \mathbf{D}$ to obtain the target prediction distribution $\hat{\mathbf{y}}_i^{\text{emo}}$ and the softmax layer is applied to normalize the logits:

$$\hat{\mathbf{y}}_i^{\text{emo}} = \Pr(y_i^{\text{emo}} \mid \mathbf{h}_i^{\text{emo}}) \tag{8}$$

$$= \text{Softmax}(\mathbf{W}_{\phi_d} \mathbf{h}_i^{\text{emo}} + b_{\phi_d}) \tag{9}$$

The model shares a common set of parameters ϕ between all members of $\mathbf{D}$ to facilitate latent emotion representation learning in a cross-domain environment, while the task-specific classification head maintains a specific set of a parameters ϕ_d.

4.3 Knowledge Distillation

The goal of a prompt-based teacher model is to extract knowledge from an LLM and transfer it to the task-specific student model, which is responsible for fine-grained emotion classification. The prompt-based model instructs the emotion classification model to enable the smaller model to generalize in a manner that resembles the teacher model. The student model minimizes a loss function which focuses on both correctly predicting the target label y^{emo} while simultaneously aligning the model with the teacher model's responses $\mathbf{y}^{\text{llm}}$.

The model utilizes a cross-entropy loss function for the single-class emotion classification task

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i}^{C} y_i^{\text{emo}} \log(\hat{y}_i^{\text{emo}}) \tag{10}$$

and a binary cross-entropy loss function for multi-class emotion classification

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i}^{C}\left[y_i^{\text{emo}} \log(\hat{y}_i^{\text{emo}}) + (1-y_i^{\text{emo}})\log(1-\hat{y}_i^{\text{emo}})\right] \tag{11}$$

for when there exists multiple emotion labels for a given document.

We use τ to represent the temperature rate hyperparameter to produce a softer probability distribution over all possible classes for class imbalances through knowledge distillation techniques. For these models, the losses from the emotion detection task and the prompt-based alignment model are summed together after each batch by using the adjustable hyperparameter λ, which balances the terms below:

$$\mathcal{L}_{\phi} = \lambda \mathcal{L}_{emo} + (1-\lambda)\tau^2 \mathcal{L}_{llm} \tag{12}$$

5 EXPERIMENTS

In this section, we provide an empirical analysis of our proposed model and investigate the following research questions:

- **RQ1:** What is the effectiveness of the proposed model for the emotion classification task in terms of model performance metrics?
- **RQ2:** Does the choice of LM contribute to the performance of the proposed model?
- **RQ3:** How does the knowledge distillation from an LLM to the proposed model contribute to the overall performance?

5.1 Data

Our experiments are conducted on two benchmark datasets: WASSA-21 dataset and Real World Worry dataset (Buechel et al., 2018) (Kleinberg et al., 2020). The WASSA-21 dataset was provided in the 11th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA) Shared Task: Empathy Detection and Emotion Classification (Tafreshi et al., 2021). The dataset consists of $n = 1860$ reactions to news stories indicating that there is harm to a person, group, or other. The labels for each record are mapped to seven emotion categories, which include a *neutral* category and Ekman's basic emotion categories: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. This label represents the dominant emotion for the text.

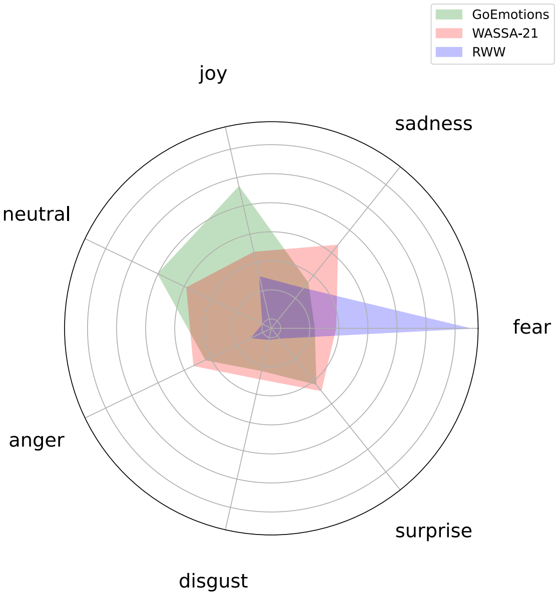

Figure 2: Distribution of the emotion labels by dataset. The RWW dataset emphasized *fear* and *sadness* labels. The GoEmotions dataset had a stronger presence of documents labeled as *neutral* and *joy*. The WASSA dataset contained more labels with the *sadness* and *surprise* labels in comparison to other datasets.

Table 1: GoEmotions emotion mapping to Ekman emotions.

Emotion	**Association**
anger	anger, annoyance, disapproval
disgust	disgust
fear	fear, nervousness
joy	joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring
sadness	sadness, disappointment, embarrassment, grief, remorse
surprise	surprise, realization, confusion, curiosity

The second dataset used in our experiments is the COVID-19 Real World Worry dataset (Kleinberg

Table 2: COVID-19 emotion mapping to Ekman emotions.

Emotion	Association
anger	anger
disgust	disgust
fear	fear, anxiety
joy	happiness, relaxation
sadness	sadness
surprise	desire

et al., 2020). The dataset contains $n = 2491$ records that were extracted by surveying participants and ask them to express their emotional feelings towards the COVID-19 pandemic. Participants were asked to construct two different forms of text. The first document they were asked to author included instructions to express their feelings towards the then current COVID-19 situation with a minimum of 500 characters. The second document expressed them to convey the same feelings in the form of a social media post that had a maximum of 240 characters. Participants were asked to rate their emotions toward the situation and select one of the following emotions that best represented their feelings: anger, anxiety, desire, disgust, fear, happiness, relaxation, and sadness. We used the emotion definitions from (Demszky et al., 2020) as indicated in Table 1 to map perform the emotion mappings as indicated in Table 2.

5.2 Baseline Experiments

To evaluate the efficacy of our proposed prompt-based knowledge distillation model, we use PLMs as the baseline for our experiments. We benchmark our model using the BERT, RoBERTa, and XLNet PLMs where the input will only be the document and target emotion(s). We evaluate the model performance of each dataset and report the mean precision, recall, and F_1-scores after 3 runs using macro averaging.

5.3 Experimental Settings

Our model was constructed using the PyTorch framework along with the HuggingFace `transformers` library for the pretrained language model implementations.[1] We followed similar experimental settings as provided in (Demszky et al., 2020). Our model uses the AdamW optimizer (Loshchilov and Hutter, 2017) while setting the learning rate to $5e^{-5}$, batch size to 16, and maximum sequence length of 512. Since previous research literature demonstrated overfitting beyond four epochs, we limited our the number of epochs during the fine-tuning step to four (Demszky et al., 2020). For the large language model, we utilized the `GPT-4o` model provided through the API.

5.4 Experimental Results

Table 3 reflects the results from the experiments conducted in this paper. Each experiment was executed independently of other datasets. The best results are indicated in bold. As reflected in the results, our method is able to demonstrate increased performance above the baseline methods for the WASSA-21 and RWW datasets. This demonstrates that the PLM acquires additional knowledge through transfer learning and knowledge distillation through this technique that it did not acquire through the data alone. Furthermore, we also discover that the RoBERTa PLM is able to achieve superior performance over the other PLMs evaluated in the tests we conducted. Despite the extreme differences in the distribution of the labels between the datasets as evidenced in Figure 2, we observe that the proposed technique is able to work given the task-agnostic knowledge provided from the teacher model. When RoBERTa was used as the underlying PLM for our technique, we were able to achieve a gain of $\Delta = +2.18$ increase in performance for the F_1 score for the WASSA-21 dataset and $\Delta = +1.86$ for the RWW dataset.

It should also be noted that the largest gain in performance was achieved through the prompt-based knowledge distillation approach with the BERT PLM in the RWW dataset. We observe an increase of $\Delta = +2.37$ in the F_1 score under these settings.

6 CONCLUSIONS

Throughout our work in this paper, we investigated the task of emotion analysis under a prompt-based knowledge distillation setting where we trained a student model by aligning it with a teacher model which provides instruction on how to generate similar probability distributions in a task-specific objective. Future directions for this work can involve exploring other techniques, such as chain-of-thought or other reasoning approaches, or augmented LLM approaches to improve the teacher model through prompting strategies. The proposed methodology can be extended to consider additional modalities of data.

REFERENCES

Agrawal, A. and An, A. (2012). Unsupervised emotion detection from text using semantic and syntactic rela-

[1] https://huggingface.co/docs/transformers/en/index

Table 3: Comparison of baselines with experimental settings. Our proposed prompt-based knowledge distillation models outperform the baseline models.

Type	**Model**	**WASSA-21**			**RWW**		
		Precision	Recall	F1	Precision	Recall	F1
Baseline	BERT	68.52	68.67	67.70	18.81	19.33	18.80
	RoBERTa	72.39	73.84	71.74	23.20	21.97	20.61
	XLNet	60.74	63.18	60.92	20.40	20.26	18.52
Experiment	BERT+PKD	69.02	71.15	68.58	23.52	23.32	21.17
	RoBERTa+PKD	**73.85**	**75.16**	**73.92**	**24.63**	**22.69**	**22.47**
	XLNet+PKD	62.55	64.29	61.75	22.52	21.41	19.28
	Δ Change	+1.46	+1.32	+2.18	+1.43	+0.72	+1.86

tions. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 346–353.

Alhuzali, H. and Ananiadou, S. (2021). SpanEmo: Casting multi-label emotion classification as span-prediction. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., and Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Buechel, S., Modersohn, L., and Hahn, U. (2020). Towards label-agnostic emotion embeddings.

Calefato, F., Lanubile, F., and Novielli, N. (2018). Emotxt: A toolkit for emotion recognition from text.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Halder, K., Akbik, A., Krapac, J., and Vollgraf, R. (2020). Task-aware representation of sentences for generic text classification. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hasan, M., Rundensteiner, E., and Agu, E. (2019). Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jiang, Y., Chan, C., Chen, M., and Wang, W. (2023). Lion: Adversarial distillation of proprietary large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kleinberg, B., van der Vegt, I., and Mozes, M. (2020). Measuring emotions in the covid-19 real world worry dataset.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lukasik, M., Bhojanapalli, S., Menon, A. K., and Kumar, S. (2022). Teacher's pet: understanding and mitigating biases in distillation. *Transactions on Machine Learning Research*.

Mackey, A., Gauch, S., and Labille, K. (2021). Detecting fake news through emotion analysis.

Plutchik, R. (1982). A psychoevolutionary theory of emotions.

Rahman, A. B. S., Ta, H.-T., Najjar, L., Azadmanesh, A., and Gönül, A. S. (2024). Depressionemo: A novel dataset for multilabel classification of depression emotions.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Staiano, J. and Guerini, M. (2014). Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 427–433.

Tafreshi, S., De Clercq, O., Barriere, V., Buechel, S., Sedoc, J., and Balahur, A. (2021). WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In De Clercq, O., Balahur, A., Sedoc, J., Barriere, V., Tafreshi, S., Buechel, S., and Hoste, V., editors, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Toraman, C., Şahinuç, F., and Yilmaz, E. (2022). Large-scale hate speech detection with cross-domain transfer. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Turcan, E., Muresan, S., and McKeown, K. (2021). Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models.

Wullach, T., Adler, A., and Minkov, E. (2021). Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.