BERT-MTL: Multi-Task Learning Paradigm for Improved Emotion Classification using BERT Model

Abhay Kumar, Neal B Desai, and Priyavarshini Murugan
Dept of Computer Sciences
University of Wisconsin-Madison

Abstract

We propose a Multi-Task learning (MTL) on the baseline BERT model that trains on multiple related tasks in parallel for performing related tasks of emotion classification from text. We explore how implicit data augmentation, regularization, and representation bias properties of MTL impact the performance on individual tasks. We evaluate the performance of the single task and MTL settings on four datasets (GoEmotions, Twitter Sentiments, Reddit Suicide/depression, and SST-2). Given that GoEmotions has multi-label examples, we extend the single-label classification to multi-label classification models. Additionally, we experiment with different emotion taxonomies to show if the potential grouping of emotion labels into higher-level categories helps in the downstream tasks. We conduct transfer learning experiments to show that the MTL model generalizes well to novel tasks and emotion taxonomies. Different ablation studies regarding task sampling, weighting, and similarity of the task are also presented. [Note: We have presented results from some of the experiments in this report, other experiments are yet to be run.]

Code is available at https://github.com/abhayk1201/CS769_Project.

1 Introduction

In recent years, emotion detection in text has become more popular due to its vast potential applications in marketing, political science, psychology, human-computer interaction, and artificial intelligence. An AI system should be powerful enough at capturing a broad spectrum of emotions people experience and express in daily life to provide a seamless experience in these applications. To engage in more empathetic interactions, future AI must perform fine-grained emotion recognition, distinguishing between many more varied emotions.

Sentiment analysis, with thousands of articles written about its methods and applications, is a well-established field in natural language processing. It has proven very useful in several applications such as marketing, advertising, question answering systems, summarization, as part of recommendation systems or even improving information extraction. In today's information age with widespread social media use, capturing and understanding a broad spectrum of emotions from posts and comments becomes an important consideration for the above applications, particularly when seeking to prevent or intervene in the case of extreme negative emotions in the users.

Emotion analysis can be viewed as a natural evolution of sentiment analysis and its more fine-grained model (Seyeditabari et al., 2018). Traditionally, most of the work in emotion recognition from text focuses on recognizing just six "basic" emotions, namely happiness, surprise, sadness, anger, disgust, and fear. Not all negative or positive sentiments are created equally. This set clearly fails to capture the broad spectrum of emotions that people experience and express in daily life, including admiration, nervousness, pride, and hope.

Identifying emotions is a hard task because of two factors: firstly emotion detection is a multi-class classification task combining multiple problems of machine learning and natural language processing while the second is the elusive nature of emotion expression in text. The latter stems from the complex nature of the emotional language and also the complexity of human emotions. Recently, there have been efforts to focus on larger classes of emotions from text-based data with the introduction of the EmpatheticDialogues dataset, which consists of online conversations in 32 different emotion categories, and the GoEmotions dataset, which consists of Reddit comments labeled with 28 different classes. These recently proposed datasets are an important step in training fine-grained emotion classification models that can recognize more nuanced emotions.

2 Related Work

2.1 GoEmotions: A Dataset of Fine-Grained Emotions

GoEmotions (Demszky et al., 2020) is a dataset of 58k carefully selected Reddit comments, labeled with 27 emotion categories or Neutral with comments extracted from popular English subreddits. The emotion taxonomy was designed considering related work in psychology and coverage in the data. The taxonomy includes a large number of positive, negative, and ambiguous emotion categories, making it suitable for downstream conversation comprehension tasks that require a subtle understanding of emotion expression. Some pertinent examples include the analysis of customer feedback or the enhancement of chatbots.

Hierarchical clustering on the emotion judgments finds that emotions related in intensity cluster together closely and that the top-level clusters correspond to sentiment categories. These relationships between emotions allow for their potential grouping into higher-level categories, which has implications for downstream tasks.

A baseline for modeling fine-grained emotion classification over GoEmotions is provided. A comparison of performance between a fine-tuned BERT based model, Ekman grouping, and sentiment grouping is provided. There is a possibility that this data can be generalized to other taxonomies and domains such as tweets and personal narratives.

2.2 Fine-Grained Emotion Prediction by Modeling Emotion Definitions

Singh et al. propose a new transformer-based framework for fine-grained emotion classification that leverages semantic knowledge of the emotion classes (Singh et al., 2021). BERT is used as the base model and an attempt is made to model the semantic meaning of emotion classes through their definitions while training the model for emotion classification. A multi-task framework with proportional sampling between emotion classification and definition modeling is employed. Experiments were conducted with three setups for definition modeling: 1. Class Definition Prediction (CDP) 2. Masked Language Modeling (MLM) and 3. both Class Definition Prediction and Masked Language Modeling (CDP+MLM). The model gives an overall improvement in F1 score in all the three setups on the fine-grained GoEmotions dataset, while the best score is obtained from the setup of CDP.

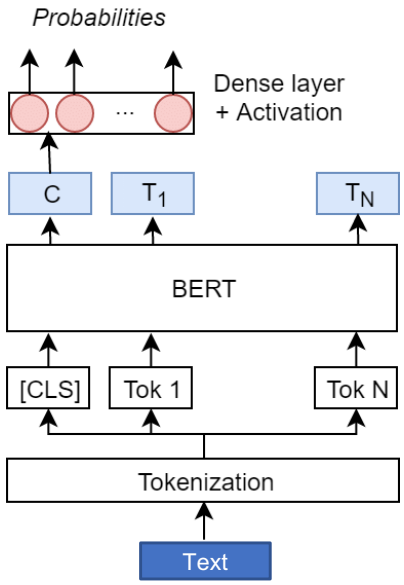
2.3 Using Knowledge-Embedded Attention to Augment Pre-trained Language Models for Fine-Grained Emotion Recognition

Pre-trained language models such as ELECTRA and BERT have achieved state-of-the-art performance in various text-classification tasks (Suresh and Ong, 2021). In this work, Knowledge-Embedded Attention (KEA) is introduced. KEA is a knowledge-augmented attention mechanism that enriches the contextual representation provided by pre-trained language models using emotional information obtained from external knowledge sources. This is achieved by incorporating the encoded emotional knowledge with the contextual representations to form a modified key matrix. This key matrix is then used to attend to the contextual representations to construct a more emotionally aware representation of the input text that can be used to recognize emotions. The main focus is on incorporating emotional knowledge to the contextual embeddings produced by pre-trained language models via late fusion and fine-tuning them to aid in the task of emotion recognition.

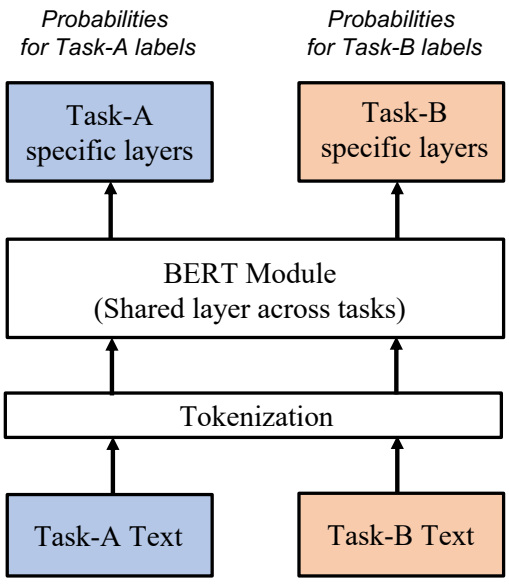
2.4 Multi-task learning:

Multi-task learning (MTL) has proven to be effective in a variety of machine learning applications, ranging from Natural Language Processing (NLP) and speech recognition to computer vision and drug discovery. In the literature, MTL has been referred to by a variety of titles, including joint learning, learning to learn, and learning with auxiliary tasks etc. According to Caruana, 1997, MTL increases generalization by using domain-specific information included in related tasks training data. The most widely used technique to MTL in neural networks is *hard parameter sharing*, often implemented by sharing the hidden layers across all tasks while maintaining numerous task-specific output layers. Another technique is *soft parameter sharing* (Duong et al., 2015) , where each task has its own model with its own parameters and the distance between model's parameters is regularized in order to encourage the parameters to be similar.

Multiple works on MTL (Ruder, 2017) show its advantages including *implicit data augmentation*, *attention focusing*, *representation bias*, and *regularization*. Learning on a single task could lead to overfitting, but jointly learning multiple tasks



(a) Overview of BERT classification model for a single task.



(b) Overview of Multi-Task Learning (MTL) framework.

Figure 1: Overview of BERT type text classification model in (a) Single Task and (b) Multi-Task Learning settings.

enables the model to learn a better representation by averaging the task dependent noise patterns and effectively offering *implicit data augmentation*. It might be difficult for a model to distinguish between relevant and irrelevant features in task with noisy or limited and high-dimensional data. Other tasks in MTL setting will offer more evidence for the relevance or irrelevance of those features, thus enabling the model to focus its *attention* on the relevant features. MTL enables the model to prefer generalized representations (*representation bias*, Baxter, 2000), which aid the model's future generalization to new tasks. By applying an inductive bias, MTL functions as a *regularizer* and reduces the risk of overfitting. The Multi-Task Deep Neural Network (MT-DNN, Liu et al., 2019) is a deep neural network that can train representations for a variety of natural language understanding (NLU) tasks. In order to adapt to new tasks and domains, MT-DNN not only uses large amounts of cross-task data, but also benefits from a regularization effect that leads to more generic representations. It incorporates a pre-trained bidirectional transformer language model BERT as its shared text encoding layers. MT-DNN achieves improved results on eight out of nine GLUE (Wang et al., 2018) NLP tasks.

3 Modeling

3.1 Architecture

Our model incorporates BERT (Devlin et al., 2018) transformer blocks as its shared text encoding layers followed by task specific classification layers. Firstly, the input text is tokenized using WordPiece embedding with 30,522 token vocabulary. The token sequence is represented as a sequence of embedding vectors. For our work, we need only one segment/sentence unlike pair of segments/sentences (like <Question, Answer>) needed in some tasks. The transformer encoder blocks captures the contextual information and generates the shared contextual embedding vectors. We keep most of the hyperparameters from Devlin et al., 2018 paper, i.e. *BERT-BASE, cased (12-layer, 768-hidden, 12-heads , 110M parameters)*. Finally, there are task-specific dense layers to learn task-specific representations followed by classification layers. The overview of the single task and MTL settings for emotion classification is shown in Figure-1.

3.2 Datasets

3.2.1 GoEmotions

GoEmotions (Demszky et al., 2020) is a dataset of 58k carefully selected Reddit comments, labeled with 27 emotion categories or Neutral, with com-

ments extracted from popular English subreddits. The emotion taxonomy was designed considering related work in psychology and coverage in the data. The taxonomy includes a large number of positive, negative, and ambiguous emotion categories, making it suitable for downstream conversation comprehension tasks that require a subtle understanding of emotion expression, such as the analysis of customer feedback or the enhancement of chatbots. The dataset summary is provided in Table-1.

Hierarchical clustering on the emotion judgments is performed, finding that emotions related in intensity cluster together closely and that the top-level clusters correspond to sentiment categories. These relations among emotions allow for their potential grouping into higher-level categories if desired for a downstream task. Ekman grouping (Ekman, 1992), and sentiment grouping are provided in the Demszky et al., 2020 paper. Some representative dataset examples are shown in Table-4.

3.2.2 Twitter Sentiment

Sentiment140 dataset (Go et al., 2009) allows you to detect sentiment of a brand, product, or topic on Twitter. It contains 1,600,000 tweets extracted using the twitter API. Rather of having human manually annotate tweets, training data was generated automatically. Tweet containing positive emoticons, such as *:)*, was annotated positive and tweet with negative emoticons, such as *:(*, was annotated negative.

3.2.3 Suicide and Depression Detection

The dataset (kaggle dataset) is a collection of posts from *SuicideWatch* and *depression* subreddits of the Reddit platform. All posts to *SuicideWatch* from its creation on December 16, 2008 until January 2, 2021 were collected, while *depression* posts were collected from January 1, 2009 to January 2, 2021. It contains 232074 texts with *suicide* or *non-suicide* labels.

3.2.4 Stanford Sentiment Treebank

Stanford Sentiment Treebank (SST-2) dataset contains positive and negative sentiment sentences extracted from movie reviews. Socher et al., 2013. We pre-processed the data as per the script provided with code released by Stickland and Murray, 2019 to get data splits with 67349 train, 1821 test, and 872 dev examples.

3.3 Metrics

Given the disparate distribution of emotion labels in GoEmotion dataset, i.e. large disparity in terms of emotion frequencies, macro-F1 score is more apt for comparative study. A macro-average computes the metric separately for each class and then takes the average (thereby treating all classes equally), whereas a micro-average aggregates all class contributions to compute the average metric. For completeness and consistency on a dataset, we chose the suitable metric used in previous line on works for the dataset.

3.4 Training Settings

In GoEmotions dataset, proper names referring to people have been masked with a [NAME] token and religion terms with a [RELIGION] token. Two unused tokens of vocab files are replaced by [NAME] and [RELIGION] tokens. For GoEmotions task, we use 50 as max token sequence length.

Single Task Setting: Empirically, we find that the best results are obtained with a small batch size of 16 and a learning rate of 5e-5 with roughly 5 epochs of training. As mentioned in Table-1, roughly 17% of training examples of GoEmotions dataset have two or more labels. To facilitate multi-label classification, we use a sigmoid cross entropy loss for GoEmotion tasks. However, other tasks like SST-2, twitter sentiment and depression detection tasks have single-label classification.

MTL Setting: We use Adam with learning rate of 2e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warm-up over the first 10% of steps, and subsequent linear decay of the learning rate, going down to zero at the end of training (Stickland and Murray, 2019).

4 Experiments

This section summarizes the different experiments run on different datasets in single task and multi-task learning settings.

4.1 Single Task Learning Setting

A baseline for modeling fine-grained emotion classification over GoEmotions is provided. We achieved comparable results to those obtained by Demszky et al (Demszky et al., 2020). Their state of the art fine-tuned BERT model on 27 discretely labeled emotion categories achieved an average F1 score of 0.46. Our model was trained using similar model architecture and hyperparameters to

Split	train (43,410), test (5,427), dev (5,426)
Labels (Original)	admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise and neutral
Sentiment taxonomy	*positive, negative, ambiguous* and *neutral*
Ekman taxonomy	*Neutral* label & 6 groups: *anger* (anger, annoyance, disapproval), *disgust* (disgust), *fear* (fear, nervousness), *joy* (all positive emotions), *sadness* (adness, disappointment, embarrassment, grief, remorse) and *surprise* (all ambiguous emotions)
Number of labels per example	1: 83%, 2: 15%, 3: 2%, 4+: 0.2% (rounded)

Table 1: Goemotion Dataset Summary

that of Demszky et al and trained using a NVIDIA Tesla P100 GPU. Our average F1 score for the original 27 labels is comparable at 0.45. The F1 score over the training checkpoints is shown in figure Figure 2. We also explored the other two label groups, namely Ekman and sentiment. The average F1 score for sentiment was 0.67 while the average score for Ekman was 0.60. Unsurprisingly, the F1 score increases as the number of labels decreases. More granular classification between the original 27 unique labels is naturally a more difficult task than the 4 groupings in the sentiment taxonomy. These results are strong baselines on which to improve upon for future experimentation with multi-task learning.

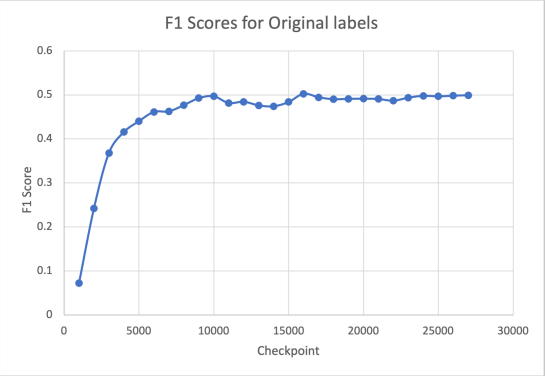


Figure 2: F1-score for original grouping over training checkpoints.

4.2 Multi Task Learning Settings

4.2.1 Tasks sampling and weighting

A straightforward technique to train a model on several tasks is *round-robin sampling*, which involves selecting a batch of training samples from each task and cycling through them in a fixed order. However, this may not work well especially when the tasks have different numbers of training examples because we may end up looping through another task's smaller dataset several times by the time we've seen every sample from the larger task (task with larger training set) and thereby overfitting to smaller tasks. Given this, we try following task sampling/weighting methods-

Proportional sampling: This method allows to observe more examples from tasks with larger datasets. Specifically, during each training step, we choose a batch of instances from i^{th} task with probability p_i, and set p_i proportional to N_i, the number of training examples for i^{th} task.

$$p_i \propto N_i \tag{1}$$

Square root sampling: This method is generalized version of the previous one and shown below-

$$p_i \propto N_i^{\alpha} \tag{2}$$

The disparity in the probabilities of choosing tasks is reduced if we pick $\alpha < 1$. In our experiments, we chose $\alpha = 0.5$ and call it square root sampling.

Annealed sampling: Stickland and Murray, 2019 notices that training on tasks more evenly towards the end of training is critical to avoid interference. They devise annealed sampling strategy where α changes with each epoch e (E is the total number of epochs). as shown below.

$$\alpha = 1 - 0.8\frac{e-1}{E-1} \tag{3}$$

Ablation Study: We experiment[1] with different sampling methods in MTL settings with two tasks-GoEmotions and SST-2. The ablation study results are shown in Table-2. Interestingly, we get same

[1]Colab link: https://colab.research.google.com/drive/1gxGwTN65cMEbvMpDNoi3T--MLkwuSSyF?usp=sharing

Sampling Method	GoEmotions eval	SST-2 eval
Proportioanl	0.63	0.92
Square root sampling	0.63	0.92
Annealed	0.63	0.92

Table 2: Evaluation accuracy for two tasks for different sampling methods in MTL setting.

evaluation accuracy for different sampling strategies on both tasks. This can be explained from the fact that both the GoEmotions and SST-2 dataset sizes are similar. We will be running experiments to see the impact on tasks with skewed dataset sizes.

4.2.2 Auxiliary Tasks

In this ablation study, we want to see the impact of auxiliary tasks. The ablation study results are shown in Table-3. MTL setting increases the evaluation accuracy for GoEmotions task by roughly 4%, whereas the evaluation accuracy for SST-2 task remains same. We will running more experiments with different combinations of auxiliary tasks/datasets along with the GoEmotions task.

Task	Training Settings	eval accuracy
GoEmotions	Standalone	0.59
	MTL setting	0.63
SST-2	Standalone	0.92
	MTL setting	0.92

Table 3: Ablation study for single-task vs MTL settings.

4.3 Ongoing Experiments

Case/Uncase Bert Pre-trained models: The GoEmotions dataset has cased sentences. We hypothesize that users usually write UPPER-case words or sentences to imply strong emotions (like anger, disapproval etc). We have shown few such examples from the GoEmotions dataset in Table-4. Note that some training examples have multiple labels. We will perform an ablation study to see if the Case vs Uncase variants of BERT models makes a difference in evaluation accuracy.

Multi-task Learning variants ablation study: MTL enables the model to infer representations that will produce a more favorable result for other tasks. This will also aid the model's future generalization to new tasks, since a hypothesis space that performs well for a large number of training tasks will also perform well for learning novel tasks. We will experiment with few novel tasks to see the representation bias of the MTL setting. We can also try different sized pre-trained bert-base (12-layer, 768-hidden, 12-heads , 110M parameters) and Bert-large (24-layer, 1024-hidden, 16-heads , 340M parameters), if our GPU setup (on Google Colab) allows training for bert-large models.

emotion	Training example
anger	YOU SHUT YOUR MOUTH WHEN YOU,ÄôRE TALKIN,Äô TO ME!
disapproval	I MEAN HE'S NOT WRONG
approval, desire	Give me [NAME]. We need some thump when we clear the benches.
curiosity	Is that proving Jewishness for Halakha or for the [RELIGION] Agency?

Table 4: Sample examples from training dataset

We will be doing extensive experiments for ablation studies mentioned in subsections-4.2.1 and 4.2.2. for similar/related tasks (Twitter sentiment or depression/suicide datasets). We will study how sampling techniques could be exploited to avoid long-tail label distribution in a given task. In fact, we observe long-tail distribution in GoEmotions dataset. We see a large disparity in terms of emotion frequencies (e.g. admiration is 30 times more frequent than grief), despite our emotion and sentiment balancing steps taken during data selection. This is expected given the disparate frequencies of emotions in natural human expression.

Transfer Learning Experiments To demonstrate that GoEmotions data generalizes across domains and taxonomies, we perform transfer learning experiments using known emotion benchmarks. The purpose is to show that GoEmotions may be used as baseline emotion understanding data when there is minimal labeled data in a target domain. We test different amounts of training data from the target domain dataset, such as 100, 200, 500, 1000, and 80% of the dataset examples. For each train set size, we produce ten random splits, with the remaining examples serving as a test set. We compare three different finetuning setups mentioned in Table-5.

Multi-label classification: (Yang et al., 2018) has shown that Multi-label classification is more

BASELINE	finetune BERT only on target dataset
FREEZE	first finetune BERT on GoEmotions, then perform transfer learning by replacing the final dense layer
NOFREEZE	identical configuration as FREEZE, with the exception that the lower layers are not frozen.

Table 5: Different Transfer Learning setups

complex than single-label classification in that the labels tend to be correlated. Modeling the multi-label classification task as a sequence generation problem helps to capture the correlations between labels and automatically select the most informative words when predicting different labels. However, this sequence- to-sequence (Seq2Seq) model suffers from exposure bias (Bengio et al., 2015), i.e., an error at early steps may affect future predictions. We plan to explore how multi-label and multi-task settings are related. For example, multi-task learning could be used to combine polarity sentiment analysis and multi-label emotion classification. (Huang et al., 2021) proposed an approach to implicitly models the relationship of different emotions in its bi-directional decoder. We will perform different experiments and will present an comparative study of different approaches for multi-label classification, especially on fine-grained emotion labels.

5 Plans

The experiments to be conducted are described in Section-6. We plan to perform ablation studies by running experiments with different settings. We already have baseline single task model and MTL setting implementations. Additionally, we will explore the extension to multi-label classification and transfer learning experiments demonstrating generalization across domains and taxonomies. We will also study other strategies to focus on minority group (e.g. weighted loss or focal loss objectives). If time permits, we will also explore how label-shift or covariate-shift could be mitigated by the MTL setting or how it impacts the performance of MTL models.

5.1 Division of work

In the following Table-6, we have added a tentative plan for the remaining weeks and division of the work among the group members.

Time	Task
Before Assignment-3	Literature survey, identified training dataset for model building, set up the initial model pipeline, implemented MTL setting and standalone single task setting, Performed fine-grained analysis on the existing method and some ablation studies. Prepared assignment-3 report.
Week 1 (04/11 - 04/17)	(1) Extend the MTL model to support multi-class classification for GoEmotions dataset. [Abhay] (2) Do rigorous ablation studies and run MTL setting on different combination of tasks (3) Run single-task models on other dataset- twitter and depression dataset. [Neal]
Week 2 (04/18 - 04/24)	(1) Run transfer learning experiments on novel task to see how representation bias of MTL helps to get a generalized representation; (2) Explore other choices for MTL. [Priyavarshini] (3) Review strengths and limitations of proposed method. [Abhay]
Week 3 (04/25 - 05/02)	(1) Prepare slides for class paper presentation of 05/02 or 05/04. [Priyavarshini] (2) Start preparing the initial draft for the final report. [Neal]
Week 4 (till 05/06)	(1) Finalize project report and update github code repository. [Abhay, Neal, Priyavarshini]

Table 6: Tentative Plan and Work division

References

Jonathan Baxter. 2000. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

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.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850.

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

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

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.

kaggle dataset. Suicide and depression detection dataset.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2021. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

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.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.