# Emotion Classification

**Kuan-Yu Lin** **Tzu-Ju Lin**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
{st180665,st180315}@stud.uni-stuttgart.de

## Abstract

We implement single-layer perceptron to perform emotion classification with the ISEAR dataset as our baseline model. However, one of the most efficient way to deal with NLP tasks is by fine-tuning BERT on the training data and add an extra layer to fit downstream tasks. We therefore propose BERT$_{\text{pure}}$, which is a model obtained through such fashion. However, the issue of machine-learning based models is that performances rely heavily on the quality and quantity of the training data. We therefore employ data augmentation and integrate lexicon-based approaches with the machine-learning based models in our experiment. All the proposed models outperform the baseline model significantly. However, we find that BERT performs well without any additional features or pre-processing. The code of this project has been released on this repository.

## 1 Introduction

Emotion classification aims to recognize and detect emotion of human behavioral data. With the growing popularity of social media, text-based emotion classification shows a promising future with its wide application on fields such as education, robotics, marketing and entertainment. However, the complexity of human emotion makes the task one of the toughest in the NLP realm. Fortunately, BERT (Devlin et al. (2018)) provides a fast and cheap way of obtaining outstanding results on this challenging topic.

Though BERT performs quite well on several emotion classification tasks ,but high performance often rely on the size of the training data. Another issue with machine-learning approach is that it is more sensitive to the quality of the training data and may fail when the data is biased, while lexicon-based relies more on the word-level of the given data and is less affected by the training data itself(Zhang et al. (2014)). We therefore hypothesize that the two approaches combined will perform better than ML models alone.

## 2 Methods

### 2.1 Perceptron

Perceptron is a supervised linear algorithm that can be used for binary classification. The algorithm learns a threshold function w˙x + b that serves as a linear decision boundary that separates two classes.

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

### 2.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a multi-layer bidirectional transformer proposed by Devlin et al. (2018). The model can provide language representation that extracts deep language features on both token level and sentence level from their bidirectional context and has been proved that it outperforms the state-of-the-art on various NLP tasks.

We apply BERT-base, which contains 12 layers of transformers block, 12 attention heads with a hidden size of 768 and has around 110M trainable parameters. The model can either be used with a fine-tuning approach, where a simple classification layer is added on top and parameters are fine-tuned on the downstream task, or with a feature-based approach, where fixed features are extracted from the pre-trained model.

### 2.3 Multilayer Perceptron

Multilayer Perceptron is a fully connected feed-forward neural network. It consists an input layer, one or multiple hidden layer of users' choice, and an output layer. In our implementation, except for the input layer, each node has a ReLU activation function and the model learns the weights between each connection through back propagation. Unlike Perceptron, Multilayer Perceptron can be used for multi-class classification.

### 2.4 T5

The Text-to-Text Transfer Transformer (T5) model is built to unified all text processing problem as "text-to-text" problem (Raffel et al., 2020). The model can perform English-to-English translation, such as "A father helping his kid to fight other kids." to "A father helping his kid to fight other children." We use it as a paraphrasing feature(Bird et al., 2021) to solve data scarcity. Hou et al. (2018) had proved that data augmentation by paraphrasing improved the results of NLP tasks.

## 3 Experiments

We briefly describe the ISEAR and NRC-Emotion-Lexicon dataset, baseline model, multi-class classification modelling approaches of our proposed models in this section.

### 3.1 Data Sets

**ISEAR** International Survey on Emotion Antecedents and Reactions is an English data set that includes text description of situations in which the respondents experienced 7 emotions (joy, anger, shame, sadness, fear, disgust, and guilt) respectively. The original data set includes 7666 examples from almost 3000 respondents. After removing invalid examples, 7509 of them were used in our experiment. 70% is used for training, 15% for validation and 15% for testing.

**NRC-Emotion Lexicon** NRC dataset, which is created by (Mohammad and Turney, 2010), is a word-emotion lexicon data set which includes 14,182 words and each word is associated with one or more of the following eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust and two sentiments: positive, negative. The lexicons are used for deriving emotions from the sentences in the ISEAR dataset.

### 3.2 Evaluation Matrices

To evaluate the performance of the models, we test the models on 1118 examples to derive precision, recall and f1-scores.

### 3.3 Baseline

Single layer perceptron is employed as a baseline for the experiment. We experiment with two statistical representation approaches, including Bag-of-Words and Tf-idf. Since single-layer perceptron can only perform binary classification, each of the seven emotion has its own classifier. Therefore, we have seven classifiers for each method and fourteen in total for the baseline. Table 1 shows the result of the baseline implementation.

| Emotions | F1-score | |
|---|---|---|
| | Tf-idf | BOW |
| Joy | 0.62997 | 0.60510 |
| Fear | 0.62136 | 0.60265 |
| Shame | 0.46840 | 0.44 |
| Sadness | 0.53602 | 0.54849 |
| Disgust | 0.53846 | 0.47896 |
| Guilt | 0.39867 | 0.43919 |
| Anger | 0.33813 | 0.32028 |

Table 1: The f1-score of the baseline models for each emotion.

### 3.4 Neural Models

We train four models in our experiment. We fine-tune BERT-base and acquire BERT$_{\text{pure}}$ to see how much better does transformer model perform compared to the baseline model. Then, to improve BERT$_{\text{pure}}$, a machine-learning based model , we propose one model to deal with data scarcity and two models to see whether emotion lexicons help improve model performances.

#### 3.4.1 BERT$_{\text{pure}}$

We first apply BERT Model with a classification layer on top. For BERT$_{\text{pure}}$, we train the model with the original sentences from the ISEAR dataset without any modification.

#### 3.4.2 BERT$_{\text{T5}}$

We apply data augmentation to the training data by creating paraphrased examples with T5. Therefore, the size of the training data is five-times bigger than BERT$_{\text{pure}}$, with a total of 26,305 examples. The remaining training procedure is the same as BERT$_{\text{pure}}$. Figure 1 shows the workflow of this model.

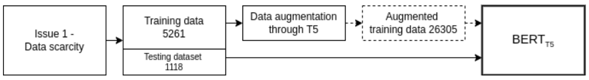

Figure 1: The workflow of BERT$_{\text{T5}}$.

#### 3.4.3 BERT$_{\text{emo}}$

In BERT$_{\text{emo}}$, we first check whether each token in the example sentences is associated with one of the emotions in the NRC emotion lexicon. Then, the related emotions are added to the end of the original sentences. For example, for the original sentence "When I [failed$_{\text{sadness}}$] an exam", the sentence is converted to "When I failed an exam. sadness." The training procedure remains the same as in BERT$_{\text{pure}}$.

#### 3.4.4 MLP$_{\text{BERT+emo}}$

We first implement the same fine-tuning procedure as in BERT$_{\text{pure}}$. Then we extract the last four hidden states associated with the [CLS] tokens, since the last four layers has produced the best results with the feature-based approach in Devlin et al. (2018). We then check whether each example includes words that are associated with the ten features in the NRC dataset. We concatenate the hidden states from BERT and ten binary features from the emotion lexicons together and use the concatenation to train a multi-layer perceptron. Figure 2 shows the workflow of BERT$_{\text{emo}}$ and MLP$_{\text{BERT+emo}}$.

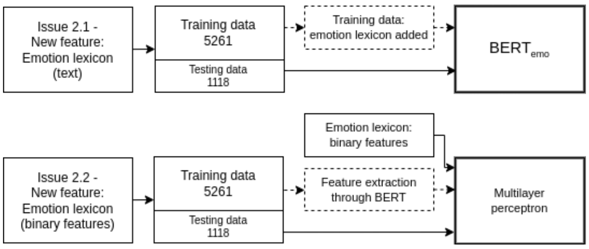

Figure 2: The workflow of BERT$_{\text{emo}}$ and MLP$_{\text{BERT+emo}}$.

### 3.5 Results

In this section, we show the results of the four proposed models. The statistics are shown both across emotions and within a single emotion.

Table 2 shows the overall results. BERT$_{\text{pure}}$ acquires a score of 0.71, followed by BERT$_{\text{emo}}$ with 0.7. BERT$_{\text{T5}}$ also obtains a similar score of 0.69, while MLP$_{\text{BERT+emo}}$ gets a score of 0.6.

| Models | Precision | Recall | F1-score |
|---|---|---|---|
| BERT$_{\text{pure}}$ | 0.71 | 0.71 | 0.71 |
| BERT$_{\text{T5}}$ | 0.69 | 0.70 | 0.69 |
| BERT$_{\text{emo}}$ | 0.70 | 0.71 | 0.71 |
| MLP$_{\text{BERT+emo}}$ | 0.60 | 0.61 | 0.61 |

Table 2: The performance of the four proposed models.

Table 3 shows the f1-scores of the four models across emotions. Model 1, which represents BERT$_{\text{pure}}$, scores the highest on joy at 0.87, followed by fear at 0.78 and the lowest 0.61 on shame. BERT$_{\text{T5}}$ also obtains a high score of 0.88 on joy and a lowest 0.59 on shame. BERT$_{\text{emo}}$ acquires the highest joy score (0.9) among all models. However, both shame and anger acquire the lowest score of 0.61. As for MLP$_{\text{BERT+emo}}$, the model sequentially gets also a nice score of 0.81 in terms of joy. However, the lowest score comes from anger instead of shame, with only 0.47.

| Emotions | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Joy | 0.87 | 0.88 | 0.90 | 0.81 |
| Fear | 0.78 | 0.74 | 0.76 | 0.69 |
| Shame | 0.61 | 0.59 | 0.61 | 0.51 |
| Sadness | 0.70 | 0.73 | 0.71 | 0.60 |
| Disgust | 0.68 | 0.66 | 0.70 | 0.61 |
| Guilt | 0.69 | 0.62 | 0.67 | 0.53 |
| Anger | 0.63 | 0.61 | 0.61 | 0.47 |

Table 3: The emotion f1-score of BERT$_{\text{pure}}$, BERT$_{\text{T5}}$, BERT$_{\text{emo}}$ and MLP$_{\text{BERT+emo}}$. Model 1, Model 2, Model 3, and Model 4 in the table refer to the models mentioned above respectively.

### 3.6 Error Analysis

We find an unbalanced performance among emotions in all models. 'Joy' outperforms all other emotions in all models (see Table 3), which indicates that this phenomenon is not caused by algorithms or methods but the nature of the data. A possible explanation can be found in Plutchik's wheel of emotions, proposed by (Plutchik and Kellerman, 2013). In the wheel, other emotions that are included both in the dataset and the wheel seem to be more related to each other than to joy. Being the only positive emotion in the ISEAR dataset, joy is intuitively and technically easier to be classified correctly.

In terms of the other six emotions, the positions of anger, fear, sadness and disgust are closer to each other in Plutchik's wheel, which indicates that they are more similar to each other and dissimilar to joy. In this case, it is difficult to distinguish them without additional information. One example from our experiment is: "Failed to be elected to be a class leader." The true label is anger, but it is classified as sadness by BERT$_{\text{pure}}$. Without being given other clues such as the facial expression and personality of the speaker, it is in fact difficult for even human to predict the emotion correctly.

Table 4 shows the confusion matrix of BERT$_{\text{pure}}$. We find that some emotions confuses our models. For example, 32 examples of disgust are mis-classified as anger and 16 from the other way around.

| | | Predicted Classes | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Joy | Fear | Shame | Sadness | Disgust | Guilt | Anger |
| Actual Classes | Joy | 143 | 2 | 7 | 5 | 1 | 2 | 1 |
| | Fear | 5 | 131 | 3 | 4 | 10 | 4 | 6 |
| | Shame | 6 | 11 | 89 | 7 | 8 | 24 | 10 |
| | Sadness | 5 | 8 | 9 | 106 | 7 | 5 | 10 |
| | Disgust | 3 | 9 | 4 | 8 | 109 | 1 | 32 |
| | Guilt | 1 | 5 | 15 | 9 | 3 | 103 | 14 |
| | Anger | 4 | 9 | 8 | 12 | 16 | 10 | 114 |

Table 4: The emotion confusion matrix of BERT$_{\text{pure}}$.

## 4 Summary and Conclusion

We propose four multi-class classification models: BERT$_{\text{pure}}$, BERT$_{\text{T5}}$, BERT$_{\text{emo}}$ and MLP$_{\text{BERT+emo}}$ for performing emotion classification on ISEAR dataset. The first three models are adapted from BERT-base with a fine-tuning fashion, while the forth is trained on text representation extracted from Bert and NRC emotion-lexicon features. All four models outperform the baseline model. Bert$_{\text{pure}}$ obtains the best f1-score, followed by BERT$_{\text{emo}}$ and BERT$_{\text{T5}}$, while MLP$_{\text{BERT+emo}}$ do not perform as well as the fine-tuning models. With a limited size and unstable quality of data, BERT seems to be still, one of the most efficient way of dealing with NLP tasks, without the need of much modification or additional features.

## 5 Future Work

For future work, we consider to modify our method of paraphrasing. Instead of performing English to English translation, we want to apply the procedure to multiple languages, such as German, Spanish, Mandarin, etc., and then translate the sentences back into English to obtain sentences with a higher variability. This might be helpful in terms of creating more data and enhancing the quality of them.

## References

Jordan J Bird, Anikó Ekárt, and Diego R Faria. 2021. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–16.

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*.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv preprint arXiv:1807.01554*.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*, volume 1. Academic Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Hailong Zhang, Wenyan Gan, and Bo Jiang. 2014. Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th Web Information System and Application Conference*, pages 262–265.