Automated Expansion of Emotion Category Labelling and sentiment analysis in Natural Language Processing.

# Abstract

Machine Learning methods of natural language processing categorisation require the training text data to be labelled to allow the model to make associations between content and labels, which is currently a manual process. Additionally, existing labelled datasets may have few broad label categories where there would be benefit in having more numerous and specific categories which more accurately represent the content of the associated text. In this instance, we use the Emotion dataset on Kaggle, which is comprised of ~400,000 text samples from twitter, each with a label from 0-5 reflecting the emotional sentiment. Research and intuition suggest that the spectrum of human emotion in text cannot be accurately described by six categories, both since more than six emotions exist, and that a text sample may reflect multiple emotional sentiments. Rather than manually relabelling each text sample to qualitatively relabel each text, this research aims to utilise Convolutional Neural Networks (CNN) to provide primary and secondary emotional labels, then retrain based on these emotional combination labels as a method of automatically relabelling text data with more nuanced and specific labels. Further, these new data will be analysed using Recurrent Neural Networks (RNN), and Bidirectional Encoder Representations from Transformers (BERT) for comparison. Finally, an application based on BERT model was built to analyse the emotion distribution in each text given. The overall aim was to automatically generate new labels, then utilise state-of-the-art machine learning methods to assess whether accurate models can be built based on the new categories and choose a model with best performance to make a real-life application. Overall, the method was successful in providing new predictive labels, and application was useful in sentiment analysis. The outcome may provide insight into alternative methods of labelling training data for machine learning models which are less labour intensive and more specific and contribute to the auto-detection of texts containing specific emotions.

# Introduction

## Emotion Labels

### The Dataset Categories

The dataset used for this project is titled “Emotions” and contains ~400,000 tweets each with a corresponding categorical emotional label between 0 and 5, indicating Sadness, Joy, Love, Anger, Fear, and Surprise. No justification is provided for the use of these specific labels, and existing psychological research suggests that there would be gaps in the ability of these labels to encompass emotion accurately. For instance, Ekman (1992) finds there are six basic facially recognisable emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. These are similar but not identical to the provided emotional categories in the dataset, as we can assume “Happiness” is the same as “Joy”, but even then, “Disgust” is not represented in the data, whereas “Love” is when it may not be considered as an intrinsically recognisable emotion. In this case, question is called to the appropriate categories for emotion labelling.

The identification and classification of unique emotions has been theorised in numerous psychological research papers, many of which conclude different results. However, the theme of primary emotions is recurrent, with other emotions being regarded as combinations of or different intensities of primary emotions. Plutchik (1980) suggests 8 primary emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. Emotions outside these categories can be found as either a combination of emotions or a different intensity. For instance, extreme Joy may be classed as ecstasy, and Love as a combination of Joy and Trust. This framework provides a reasonable ground for the expansion of emotional categories in this context, as many of the primary emotions are present in the labelling system.

## Alternative Work

Tromp & Pechenizkiy (2014) utilised the Plutchik classification style and used a Rule-Based Emission Model (RBEM) auto-encoding method, though this still required the manual labelling of a dataset in Plutchik’s style to function. However, despite utilizing Plutchik’s model, the study does not produce any findings related to combinations of emotions, and only assign a single label relating to the highest probability score for the label. To this extent, utilising combinations of emotions is missing.

# Method

## Initial Dataset Analysis

The task of relabelling first required the construction of an accurate CNN model in which the original dataset can be analysed to identify primary categories. This CNN model was built using TensorFlow on Python, using rectified linear unit (ReLU) and softmax activation functions to convert the output into a vector of probabilities for each label, and standard kernel and filter sizes of 3 and 64 respectively. Additionally, the optimisation algorithm used was Adaptive Movement Estimation as much recent research suggests this outperforms traditional stochastic gradient descent for machine learning tasks (Hassan, et al (2022); Kingma & Ba, 2014).

## Construction of Relabelled Dataset

The same CNN model was run on training datasets where one emotion was removed from the labels, therefore forcing the model to select the next best label based on the labels it had access to. Doing this for each emotion removed allowed the creation of a new dataset containing the two best labels and their respective probabilities for each text. To distinguish between combinations of emotions and genuine primary emotions, a probability threshold was set for the secondary label at 70%. For texts where the secondary label probability was above the threshold, the label was assigned the combination of the primary and secondary label. The original CNN model was run on the new dataset, as well as an RNN and BERT model for accuracy comparisons.

## Relabelled Dataset Analysis

Finally, the CNN model was run using the entire new label dataset as training data, and the remaining data from the original dataset as the test data to provide predictive new labels to the old dataset.

### Build RNN model and trained on two datasets

In that work we proposed an improved RNN model which was based on the novel Long Short-Term Memory (LSTM) units and used to analyze the relabelled dataset with the expanded emotional categories. The LSTM which was incorporated in the stacked architecture model, in particular has been designed to capture the complexities of emotional sentiments conveyed in linguistic sequences. The model's schema had multiple dropout layers to adjust the overfitting and the final softmax layer was used to classify different emotional categories.

### Build BERT model and trained on two datasets

The BERT (Bidirectional Encoder Representations from Transformers) model is one of the most powerful language models to solve text classification tasks. In this project, we selected the “bert-base-uncased" model which is a basic pre-trained BERT model and is sufficient to solve our emotion classification task. The basic training steps for two datasets are the same. For data preparing part, the original data was cleaned by removing noise from them, such as HTML tags, URLs, and special characters. Then dataset was split into training and testing set. For text tokenization part, we converted raw text into tokens that BERT can understand by using BERT tokenizer. Each input sequence is of same length for batch processing. This was achieved by truncating long texts and padding short ones. We also created attention masks to tell the model which parts of the input are actual data, and which are padding. Before we started training, we first loaded pre-trained “bert-base-uncased" BERT model which has a final layer adapted to classify inputs into one of emotion categories. For training of the model, we set the epoch to three and set AdamW as optimizer. We used cross-entropy loss function to calculate the loss and the backpropagation was used to update the model’s weights.

### Merge of results

After getting new relabelled dataset, we trained both of original dataset and relabelled dataset on three different models which are CNN, RNN and BERT. So, for each dataset, we can evaluate which model have best performance in overall based on precision, recall, F1-score, etc. And we can also compare the advantage and disadvantages of three models. According to the results, we chose the model with best overall performance in both original and relabelled dataset when building the application in next step.

### Application

Based on the performance of CNN, RNN and BERT on both original and relabelled dataset, we chose a model with best overall accuracy. A user interface based on that best model was designed which allows user to type their customized sentences, and the output will be the probability distribution of six emotions including Sadness, Joy, Love, Anger, Fear, and Surprise. The output is in percentages rather than a single emotion label, which provides more details about the emotion combinations of a sentence, so if a sentence contains a secondary or even thirdly emotion, users are allows to find it.

# Results

## Convolutional Neural Network

### Original Emotion Labels Model

The overall accuracy of the CNN model for the original dataset including all emotional labels was 0.93, with the metrics of each emotion achieving precision and recall of 0.97 and 0.96 for sadness, 0.98 and 0.91 for joy, 0.77 and 0.95 for love, 0.92 and 0.95 for anger, 0.92 and 0.86 for fear, and 0.73 and 0.94 for surprise. The confusion matrix (Figure 1) provides some insight into the lower precision for love and surprise, where the model regularly mislabels joy as love and fear as surprise.

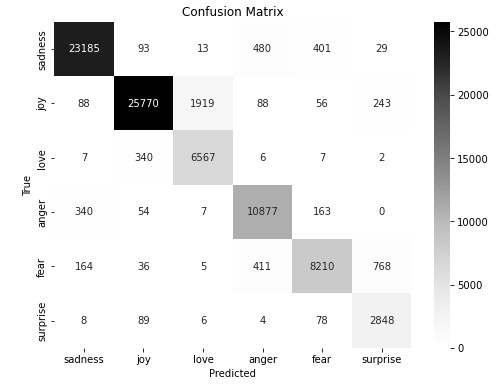


Figure 1 - Confusion Matrix for the CNN model on the original dataset.

The Receiver Operating Characteristic (ROC) curve demonstrates an area under curve (AUC) of ~1 for each emotion (figure 2), demonstrating a high ability to distinguish between classes.

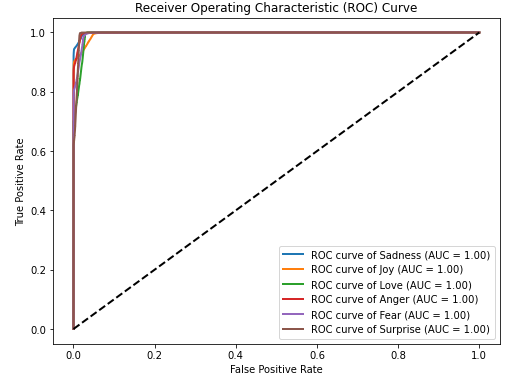


Figure 2 - ROC Curve per emotion for the CNN model on the original dataset.

### CNN Model Relabelled Emotion

The new dataset with the relabelled emotion shows a wide range of emotional combinations with varying counts. The primary three emotions represented are Joy, Sadness and Joy, and Sadness (figure 3), which is representative of the fact that Joy and Sadness were the most represented emotions in the original dataset.

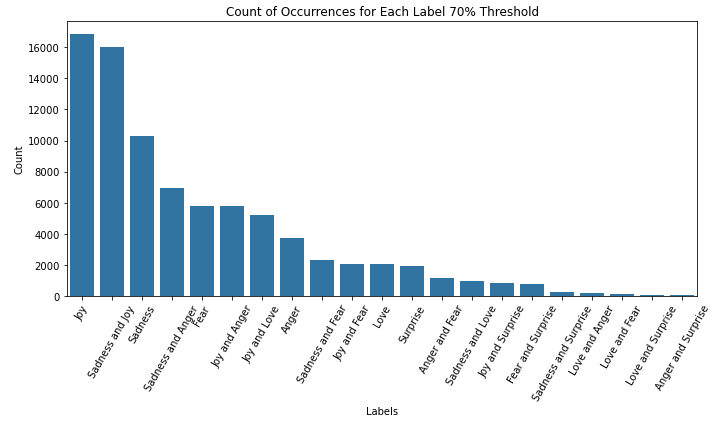


Figure 3 - Bar Chart demonstrating the counts of each new labelled emotion.

The overall accuracy of the CNN model for the relabelled dataset including all emotional labels was 0.80. The precision and recall metrics for each individual emotional label vary widely (Figure 3)

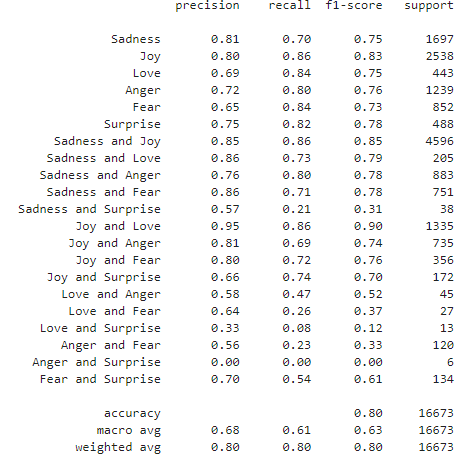


Figure 4 - Precision, Recall, f1-score, and support for each emotion label in the new dataset.

The ROC curve for the relabelled dataset demonstrates AUCs consistently above 0.99, with only one combination emotion reaching as low as 0.97 (Sadness and Joy). This is prior indicated to by the confusion matrix for the model run with no Joy label prediction, as the most often alternative prediction is for Sadness (Figure 6).

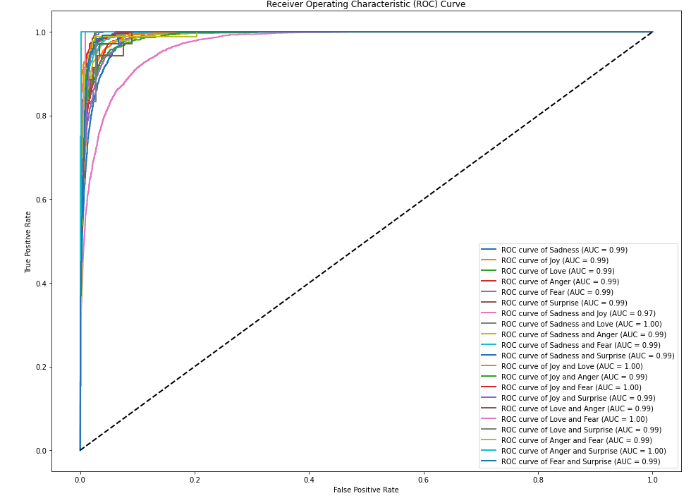


Figure 5 - ROC Curve per emotion for the relabelled dataset CNN model

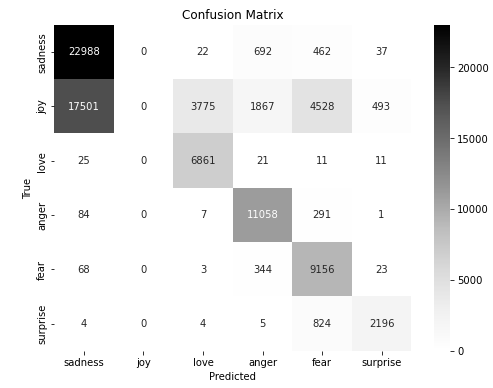
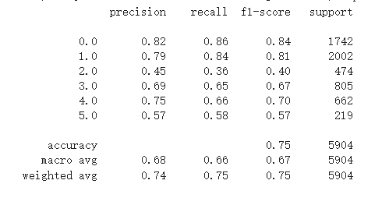


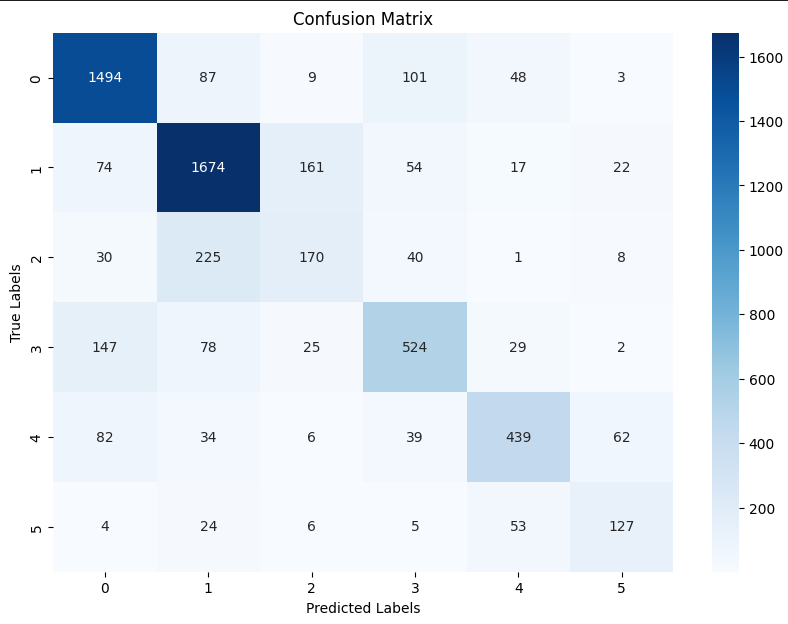
Figure 6 - Confusion matrix for original CNN model without Joy prediction.

## Recurrent Neural Network

The application of a Recurrent Neural Network (RNN) to the original dataset containing tweets labeled with six distinct emotions provided insightful outcomes. The overall accuracy of the RNN model reached 75%, demonstrating a robust capability to classify emotional content (Figure 7). In terms of precision and recall, the RNN performed best with the emotions of Sadness and Joy, achieving precision scores of 0.82 and 0.79, and recall scores of 0.86 and 0.84, respectively. These results indicate a high reliability in identifying tweets expressing these emotions. Conversely, the emotion of Love presented challenges, with notably lower precision and recall at 0.45 and 0.36, suggesting difficulties in accurately identifying this less frequently expressed emotion in the dataset.

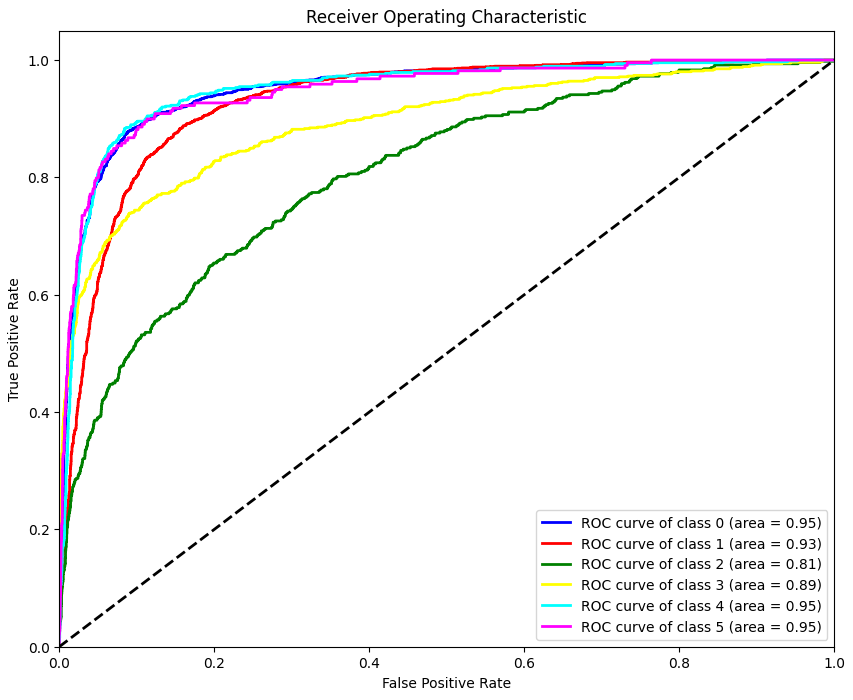


*Figure 7 - - Precision, Recall, f1-score, and support for original RNN model on the original dataset*



*Figure 8 - Confusion matrix for original RNN model on the original dataset*

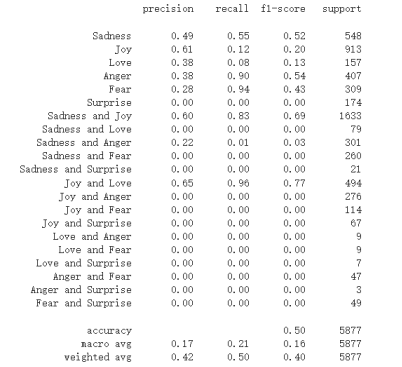
The confusion matrix (Figure 8), further elucidates the model's performance, highlighting areas of strength and weakness. For instance, the matrix shows a tendency of the model to confuse Love with other emotions, possibly due to its less distinctive expression in text or lower representation in the training data. The Receiver Operating Characteristic (ROC) curves (Figure 9) for each emotion underscore the model's diagnostic ability, with AUC scores impressively above 0.85 for all emotions, peaking at 0.96 for Sadness. These curves visually represent the trade-off between true positive and false positive rates at various threshold settings, affirming the model’s effectiveness across different emotional categories.



*Figure 9 – Roc curve for original RNN model on the original dataset*

Enhanced Recurrent Neural Network (RNN) Results with Relabelled Dataset

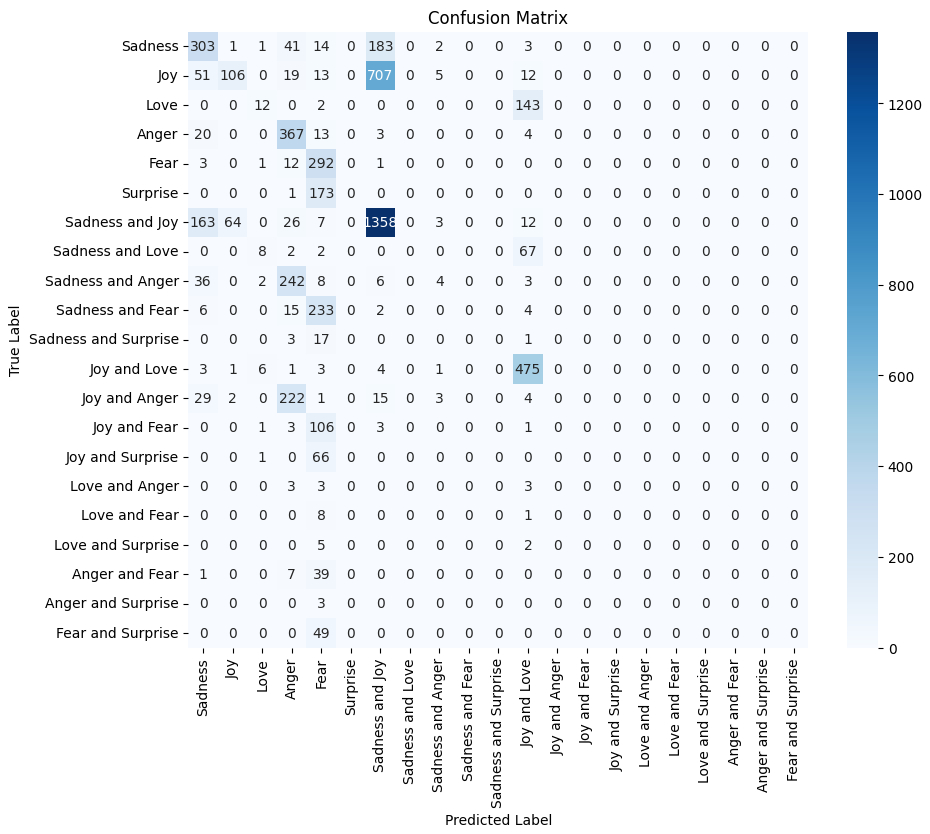
Although the evaluation of the advanced RNN on the relabelled dataset was not as favourable as it should be, and reflected that the wider scope of emotional expressions are challenging, it was still worth performing the enhanced RNN. The overall accuracy of the model was 50%, indicating a moderate capability to classify complex emotional states correctly (Figure 10). Notably, traditional emotions such as Sadness and Anger showed reasonable detection rates, with specific combinations like Sadness and Joy, and Joy and Love achieving higher precision and recall, which suggests some emotional combinations are more distinctively expressed in text than others. For instance, the precision and recall for 'Sadness and Joy' were notably high at 0.60 and 0.83 respectively, and 'Joy and Love' demonstrated a precision of 0.65 and a recall of 0.96, indicating a strong model performance for these specific emotional blends.



*Figure 10 - - Precision, Recall, f1-score, and support for Enhanced RNN with Relabelled Dataset*

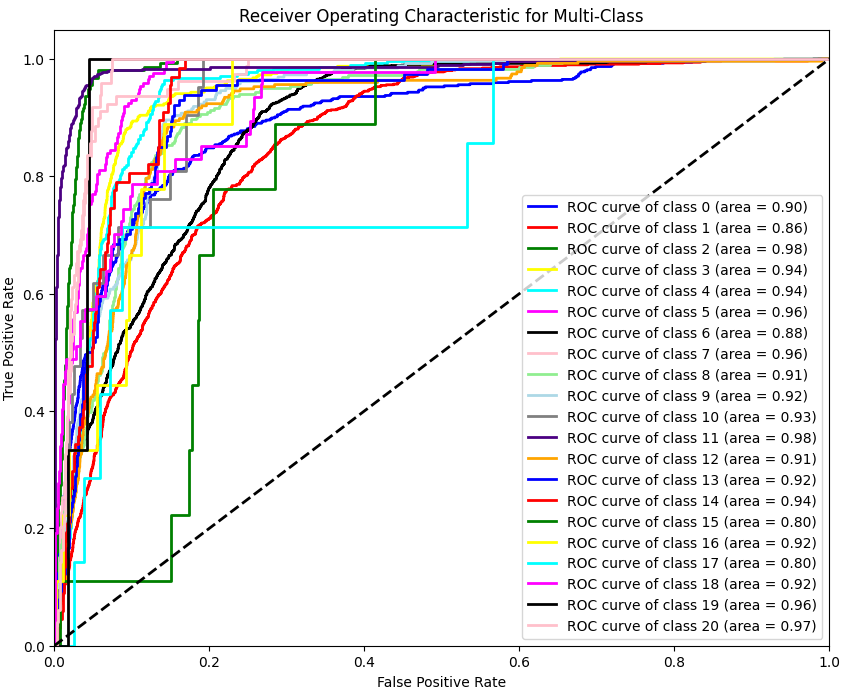
However, the model struggled with less frequent and more nuanced emotional combinations such as 'Sadness and Fear' and 'Love and Anger,' where both precision and recall were extremely low, pointing to the difficulties in distinguishing these subtler emotional states. This variability in model performance underscores the complexity of emotional categorization in natural language processing, particularly when expanding the spectrum of recognized emotions.

The confusion matrix provided further insight into the model's performance, illustrating substantial confusion between certain emotions, particularly where the model frequently misclassified less distinct emotions. This aspect of the analysis highlights the challenges in training models on complex emotional data where some categories may not be as clearly or frequently defined.



*Figure 11 - Confusion matrix for Enhanced RNN with Relabelled Dataset*

Most classes showed AUC values significantly above the no-skill line (diagonal), but the effectiveness varied significantly across classes, echoing the precision and recall metrics. Classes with higher AUC values corresponded to emotions or combinations thereof that were better represented or more distinct in the dataset.

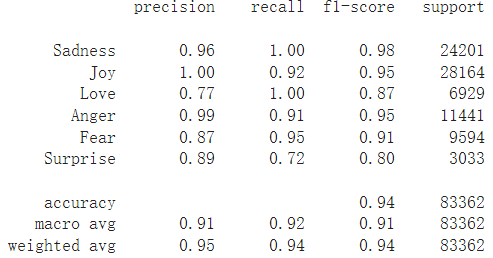


*Figure 12 – Roc curve for Enhanced RNN with Relabelled Dataset*

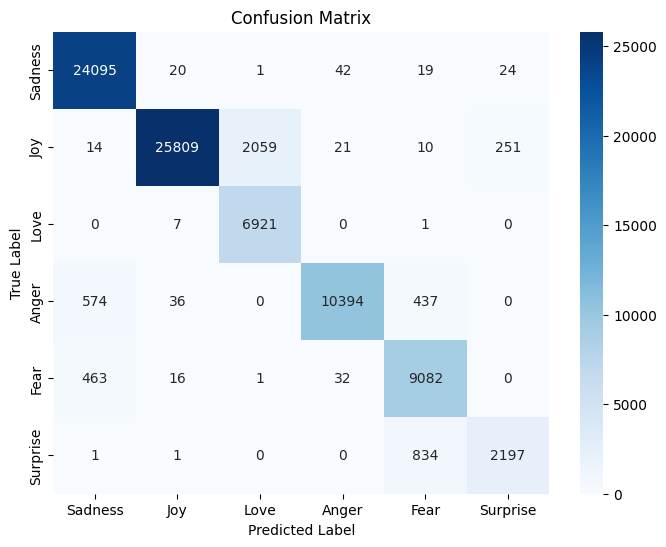
## Bidirectional Encoder Representations from Transformers

### Model for original dataset

The results for BERT model trained with original dataset are as follows, from Figure 13 we found that BERT model has a high accuracy of 0.94, which is higher than CNN (0.93) and RNN (0.75), meaning that BERT model has best performance across most emotions, especially for “Sadness,” “Joy” and “Anger,” with f1-score of 0.98, 0.95 and 0.95. However, we found that “Love” still has a low precision of 0.77, which means that model sometimes classify other emotions such as “Joy” into “Love,” which can be found in confusion matrix in Figure 14.The recall of “Surprise” is 0.72 which means that model missed more actual instances of “Surprise” than other 5 emotions, the reason maybe that the total number of “Surprise” is significantly lower than other emotions.

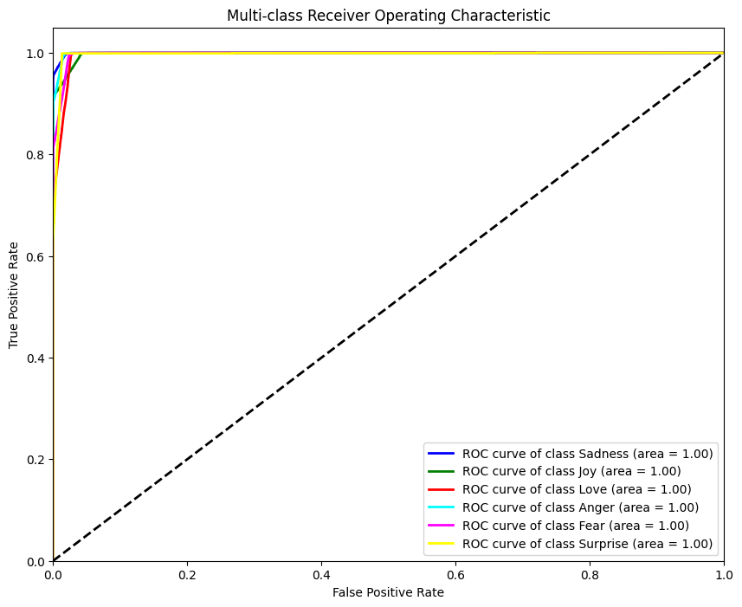


*Figure 13 - - Precision, Recall, f1-score, and support for BERT with Original Dataset*



*Figure 13– Confusion matrix for BERT with Original Dataset*

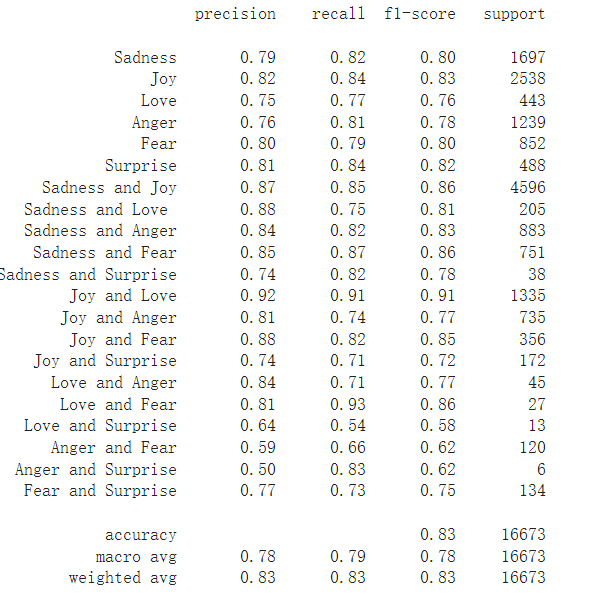
The Receiver Operating Characteristic (ROC) curve demonstrates an area under curve (AUC) for each emotion in Figure 14, the area for 6 emotions is all 1, demonstrating a high ability to distinguish between classes.



*Figure 14– ROC for BERT with Original Dataset*

### Model for relabelled dataset

The results for BERT model trained on relabelled dataset with mixed emotion labels are as follows. The overall accuracy is 0.83, which is significantly higher than RNN’s accuracy (0.50) and better than CNNs (0.80), indicating that BERT still has the best performance in recognizing mixed emotional labels. The main advantage for BERT over other two models is that BERT is good at detecting the labels with fewer presences, which leads to higher recalls accordingly. For example, label “Anger and Surprise” only appears 6 times in dataset and both CNN and RNN fails to detect them (recall = 0) while BERT has a recall of 0.83. Also, for label “Anger and Fear,” RNN fails to detect them, and CNN has a recall of 0.23 while BERT has a recall of 0.66.



*Figure 15- - Precision, Recall, f1-score, and support for BERT with Relabelled Dataset*

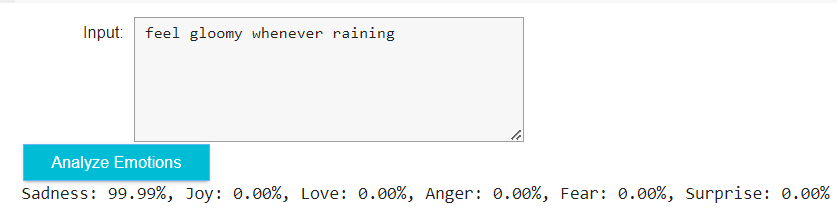
## Discussion of results from three models

To make a comparison of 3 models' performance based on above result analysis, we can conclude that BERT model has best performance in the classification of texts with both single and multiple labels – its accuracy in original and relabelled dataset is 0.94 and 0.83, both the highest among three models. For CNN model, the accuracy in original dataset is 0.93 which is extremely close to BERT model, however its accuracy in relabelled dataset is a bit lower than BERTs (0.80) due to lower recalls for some labels. Although the overall accuracy for BERT is better than CNNs, the CNN has its advantage that it is computationally cheaper. The RNN model has a mediocre performance with accuracy of 0.75 in original dataset and 0.5 in relabelled dataset, the likely reason is that RNN is not good at text classification tasks, or the model we built is not complicated enough to carry out this task. In conclusion, we chose BERT trained with original dataset as our model for application, and in future works, auto-expansion of emotion labels can be done by BERT model, while currently we use CNN.

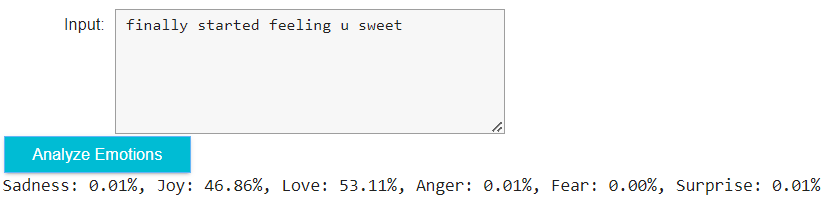
## Application

To test how our application works, we used some text from relabelled dataset which contains both single-emotion label and mixed emotion label, so model’s ability of recognizing two kinds of labels can be tested. For instance, text “feel gloomy whenever raining” is classified as “Sadness,” and the output we got from model is 99.99% of emotion in this text is sadness as in Figure 16. And for text “finally started feeling u sweet” which is classified as “love and joy,” the output tells that text contains 46.86% love and 53.11% joy, which matches its label perfectly.

More than 100 texts were tested, and the model application has a satisfying performance overall. In feature works, this model can be applied in diverse areas including auto-detection of comments with negative emotion such as anger and sadness, and it can also be used to modify other emotion datasets with single emotion label for each text by classifying emotion text more specifically (e.g. change original label of “Anger” to “Fear and Anger”.)



*Figure 16- - Output of the model when input is “feel gloomy whenever raining”*



*Figure 17- - Output of the model when input is “finally started feeling u sweet”*

# Conclusion

The method of expanding emotion data labels automatically was successful, though consideration must be given to ways of analysing predictions on data that does not have validation labels, otherwise accuracy of predicted labels can only be drawn from the intermediate sample of tested rebelling data, which is derived from the test data of the original sample. However, the method could be applied in scenarios such as product labelling, whereby each product in a catalogue has a text-based description and some rudimentary tags, which could be expanded upon to make more niche categories for product searching.

Additionally, the most successful model was the Bidirectional Encoder Representations from Transformers model, so if this method were to be repeated it may be beneficial to utilise BERT at every stage in the process. However, this may increase computation times, as the CNN was able to produce competitive results with short computational times.

Also, an application is built based on BERT model which allow users to input their desired text, and they will get the sentiment analysis of that text including emotion distribution of six emotions.

# References

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3-4), 169–200. <https://doi.org/10.1080/02699939208411068>

Hassan, E., Shams, M. Y., Hikal, N. A., & Elmougy, S. (2022). The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Multimedia Tools and Applications. <https://doi.org/10.1007/s11042-022-13820-0>

Kingma, D. P., & Ba, J. (2014, December 22). Adam: A Method for Stochastic Optimization. ArXiv.org. https://arxiv.org/abs/1412.6980

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), Emotion: Theory, research, and experience: Vol. 1. Theories of emotion (pp. 3-33). New York: Academic.

Tromp, E., & Pechenizkiy, M. (2014). Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel. ArXiv, abs/1412.4682.