Assignment 3: Training a Scikit-learn model

The dataset chosen for this assignment is the emotion dataset from HuggingFace, used for a text-based classification task. The emotion dataset contains tweets labelled with six emotions: love, anger, fear, joy, sadness, and surprise. The dataset is manageable, containing around 20,000 samples, and requires preprocessing (text to numerical features), which a TfidfVectorizer performs. The dataset was separated into training and test sets via an 80-20 distribution.

The machine learning model selected from Scikit-learn is the LogisticRegression model. This is mainly because it is well-suited for multi-class classification via Scikit-learn's multi_class = multinomial parameter, which this task requires given that text must be classified into one of six emotions. Although other models could also work, LogisticRegression is much simpler to implement and much more efficient, allowing for a relatively large dataset to run on moderate hardware. It is also very interpretable; the influence of each feature on predicting an emotion can easily be observed since each feature gets a coefficient for each class. Finally, the model is compatible with the dataset. We are taking input text from the dataset and preprocessing into features using TfidfVectorizer, resulting in a sparse, high-dimensional matrix. LogisticRegression assumes a linear relationship between features and the log of odds of the classes, which works effectively for text classification. It is also less prone to overfitting on sparse data sets compared to tree-based models or linear regression.
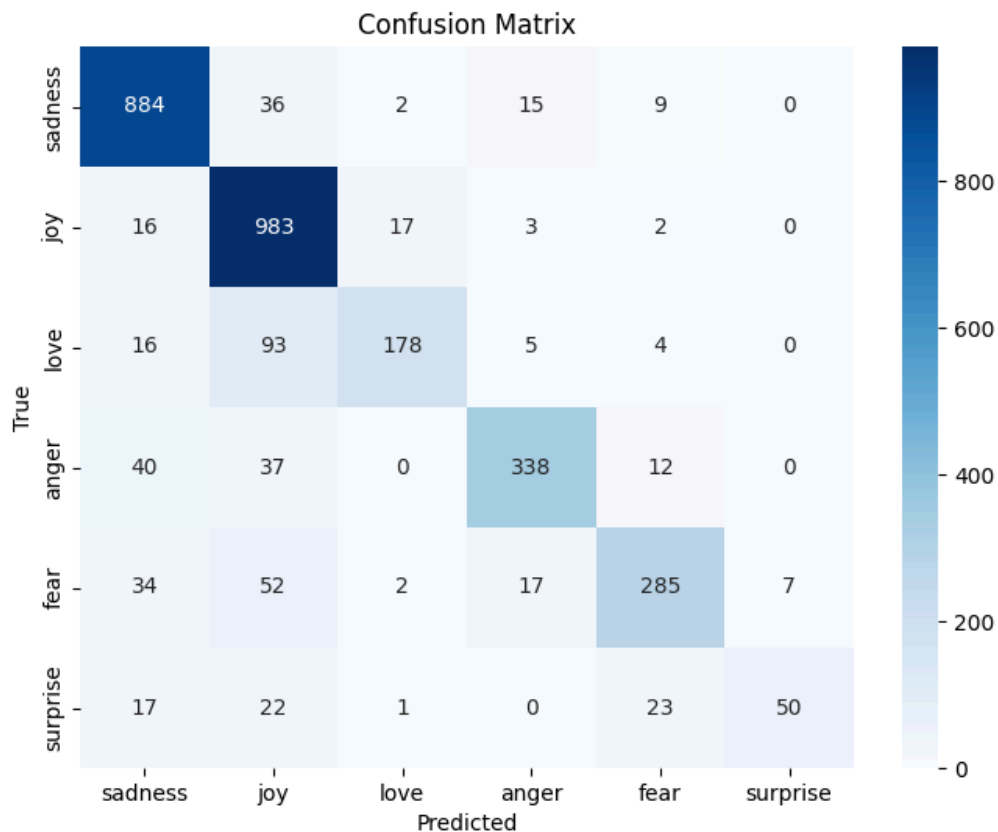
The performance metrics for the model's predictions are shown below:

Accuracy: 0.8494
Classification Report:

	Precision	Recall	F1-score	Support
Sadness	0.88	0.93	0.91	946
Joy	0.80	0.96	0.88	1021
Love	0.89	0.60	0.72	296
Anger	0.89	0.79	0.84	427
Fear	0.85	0.72	0.78	397
Surprise	0.88	0.44	0.59	113
Accuracy			0.85	3200
Macro avg	0.87	0.74	0.78	3200

Weighted avg	0.85	0.85	0.84	3200

Confusion Matrix

The overall accuracy of the model's predictions was 0.8494, meaning the model is very effective in making correct predictions. The F1-scores for each respective emotion show that it was better at predicting some emotions over others. This is likely because some emotions were more represented in the training data than others, giving the model a wider range of samples to train itself on. This is supported by the fact that sadness and joy, which were represented the most in the dataset, boasted the highest F1-scores.