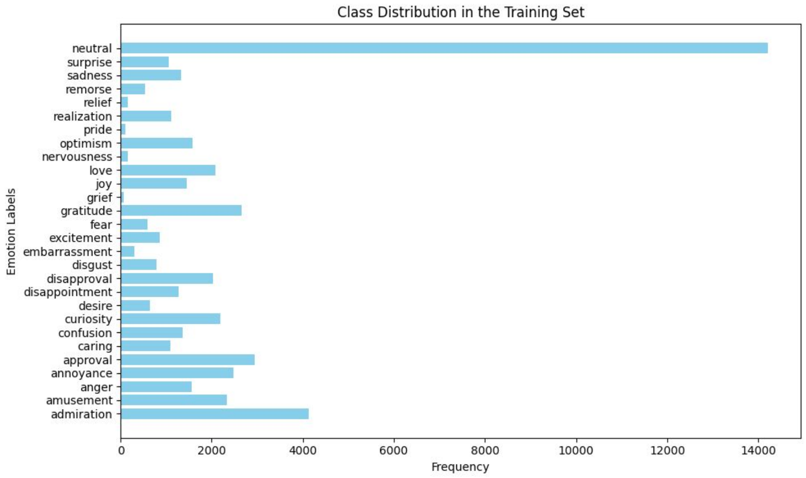
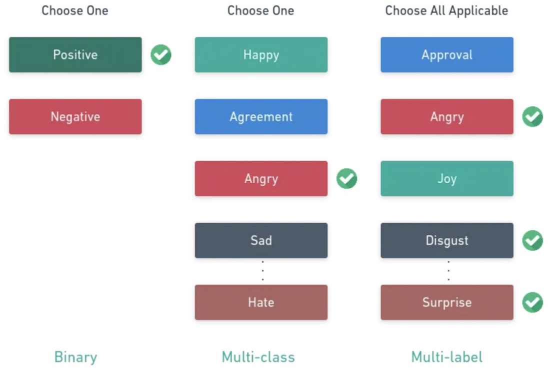
Multi-Label Emotion Detection Using Transformer Models

Diego Toribio
ECE-467: Natural Language Processing

Overview

- Goal:
  - Detect multiple emotions in text using multi-label classification.
- Dataset:
  - GoEmotions (58k text samples, 27 labels).
- Models Explored:
  - BERT
  - RoBERTa
  - DistilBERT
  - SqueezeBERT
- Tools:
  - Google Colab *Pro* & Kaggle (compute)
  - Weights & Biases (experiment tracking)





Dataset & Preprocessing

- Preprocessing:
  - Tokenization: Max 64 tokens
    - + padding/truncation
  - Label Encoding: Multi-hot binary vectors
  - **Challenge:** Class imbalance and overlapping emotions.

Dataset Split	Number of Samples	Number of Classes
Train	43410	28
Validation	5426	28
Test	5427	28

Table 1. Number of samples in the training, validation, test sets, along with the total number of emotion classes.

Category	Example Text
Neutral	My favourite food is anything I didn't have to cook myself.
Joy	Happy to be able to help.
Realization	Maybe that's what happened to the great white at Houston zoo.
Pride	I am just like this! Glad to know I'm not imagining it.
Excitement	It's crazy how far Photoshop has come. Underwater bridges?!! NEVER!!!

Table 2. Example Text Samples with Emotion Labels

Transformer Architecture

- Transformer Architectures:
  - **BERT**: Bidirectional model, baseline.
  - **RoBERTa:** Dynamic masking, no NSP → improved performance.
  - **DistilBERT**: Lightweight BERT with faster inference.
  - **SqueezeBERT**: Efficiency-focused with low computational cost.
- Experiment Details
  - Hyperparameters:
    - LR: 3e-5, 5e-5
    - Dropout: 0.3, 0.4, 0.5
    - Batch Sizes: 32, 64

Results Overview

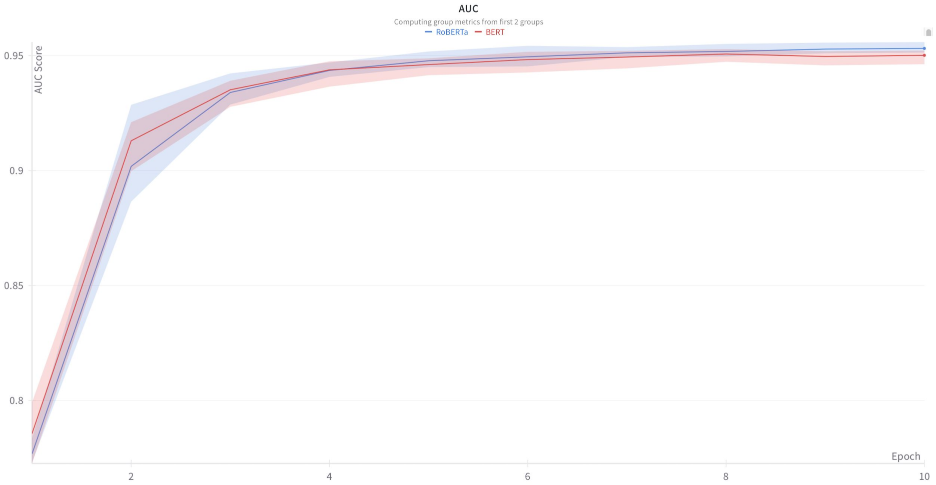
- RoBERTa outperformed all models (AUC, Training/Validation Loss)
- Lightweight models: faster but less accurate.

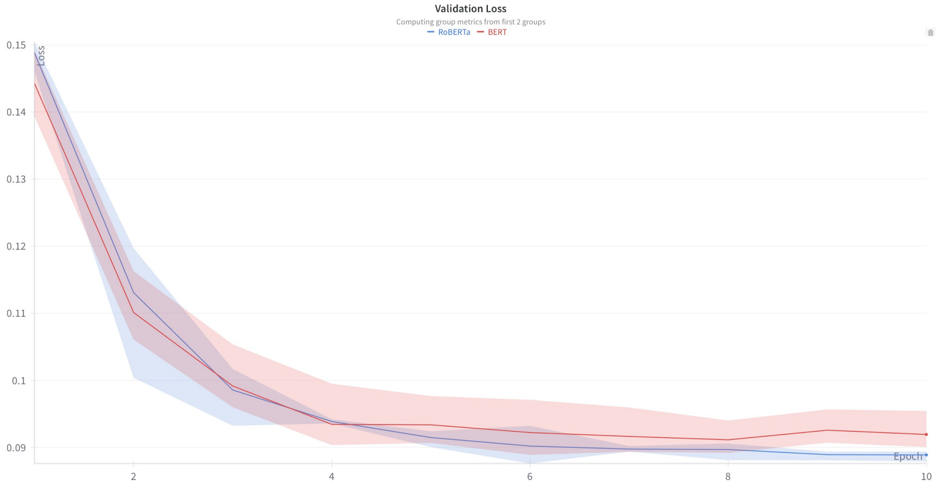
Model	Learning Rate	Batch Size	Dropout	AUC Score	Training Loss	Validation Loss
BERT-base	3e-5	32	0.3	0.952	0.073	0.090
RoBERTa	3e-5	32	0.3	0.956	0.081	0.088
BERT Large	3e-5	32	0.3	0.953	0.060	0.097
RoBERTa Large	3e-5	64	0.3	**0.957**	0.078	0.089
DistilBERT	3e-5	64	0.4	0.948	0.052	0.096
SqueezeBERT	5e-5	32	0.3	0.944	0.092	0.093

Table 4. Best-performing results for each model with optimal hyperparameters and corresponding AUC scores.

Performance Comparison

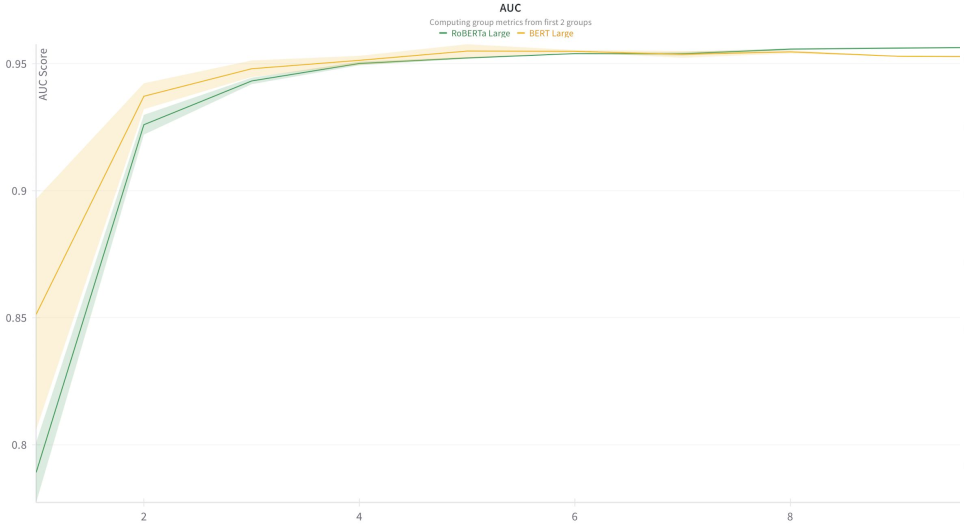
- RoBERTa outperforms BERT in both AUC (left) and validation loss (right).
  - Faster convergence and lower loss demonstrates better optimization and generalization in RoBERTa

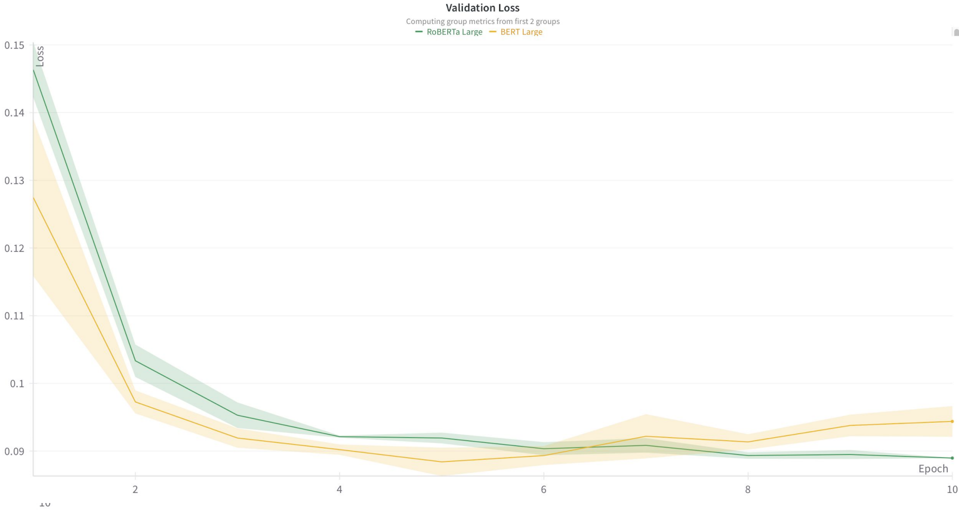




Scaling Up Performance

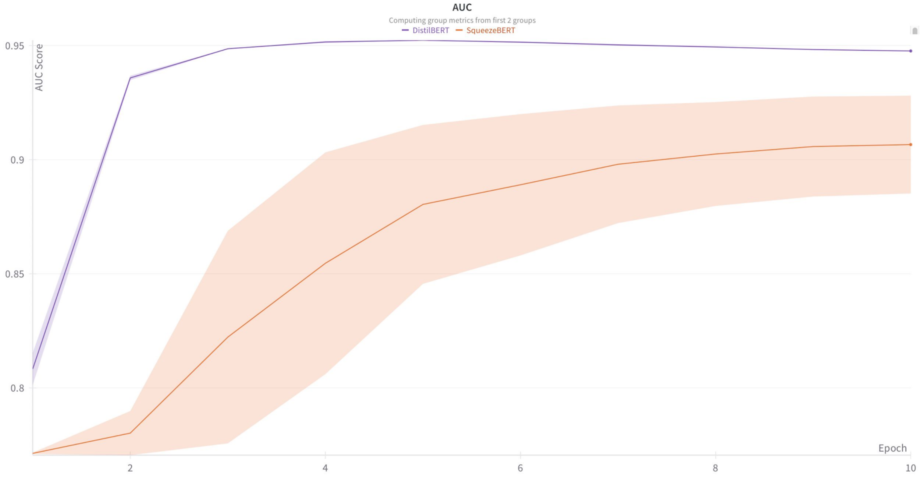
- Scaling up improves performance for both models.
- RoBERTa-Large outperforms BERT-Large with faster convergence, lower validation loss, and *slightly* higher AUC stability.

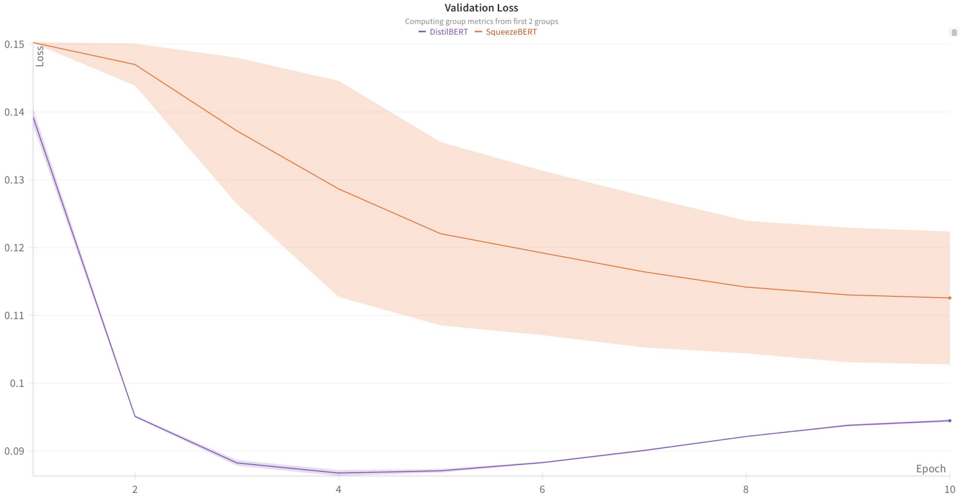




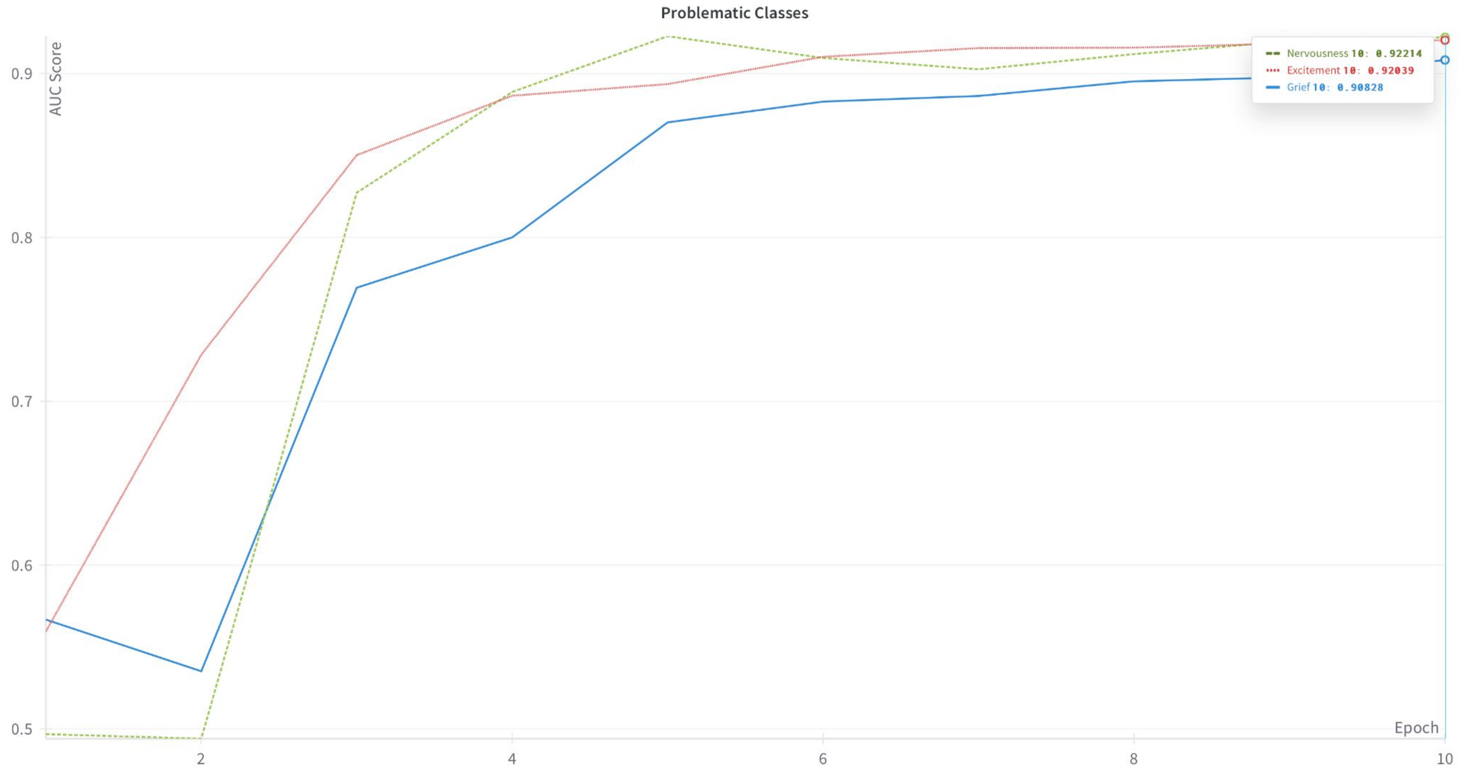
Lightweight Model Comparison

- DistilBERT outperforms SqueezeBERT in both AUC and validation loss, with faster convergence and better generalization.
    - SqueezeBERT prioritize efficiency over accircuy.





Problematic Classes



Conclusion

- RoBERTa consistently outperforms BERT with faster convergence, higher AUC, and lower validation loss.
- Lightweight Models:
    - DistilBERT balances efficiency and accuracy.
    - SqueezeBERT favors speed at the cost of performance.
- RoBERTa is the optimal choice for accuracy, while DistilBERT is ideal for resource-constrained settings.

Questions?