USC University of Southern California

USC Viterbi School of Engineering

GoEmoX: Multi-Label Emotion Classification with DistilBERT

Shipeng Sun

DSCI 510

Introduction

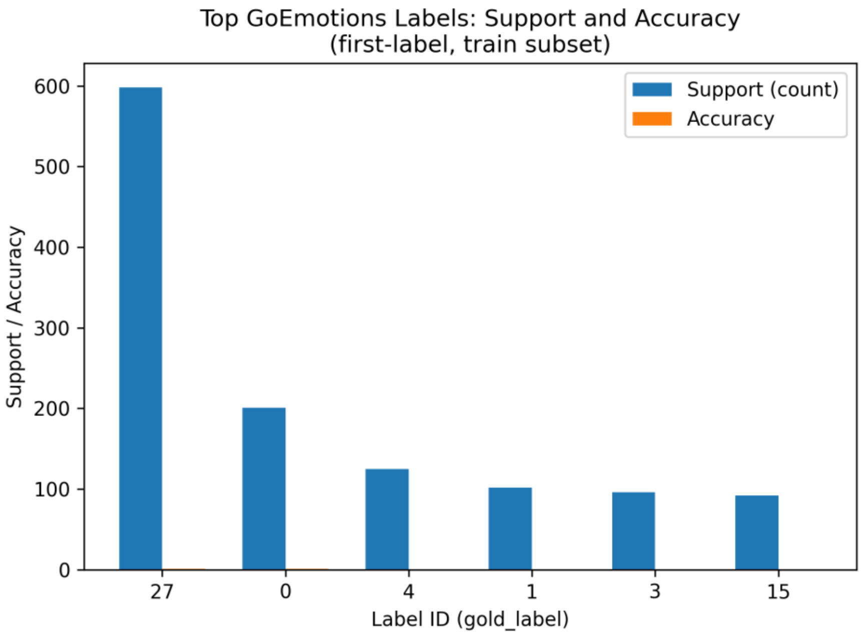
In this project I fine-tune a DistilBERT transformer model on a subset of the GoEmotions dataset to predict one of 28 emotion labels from short English texts. The in-domain data comes from Reddit comments, while the out-of-domain data includes HackerNews story titles and descriptions as well as GitHub Flask issue titles, which differ strongly in style and topic. The main goal is not only to obtain a reasonable emotion classifier on GoEmotions, but also to analyze what emotions the model actually learns and how its predictions change when it is applied to these out-of-domain texts. By comparing performance, prediction distributions, and confidence across datasets, I aim to understand the model’s robustness and its tendency to be over-confident on unfamiliar inputs.

Data Sources

Dataset	Domain / Text Type	Size Used	Access Method
GoEmotions (simplified)	Reddit comments, 28 emotion labels	2,000	HuggingFace API
HackerNews top stories	News headlines & short descriptions	300	HackerNews REST API
GitHub Flask issues	Bug reports & feature requests	4	GitHub REST API

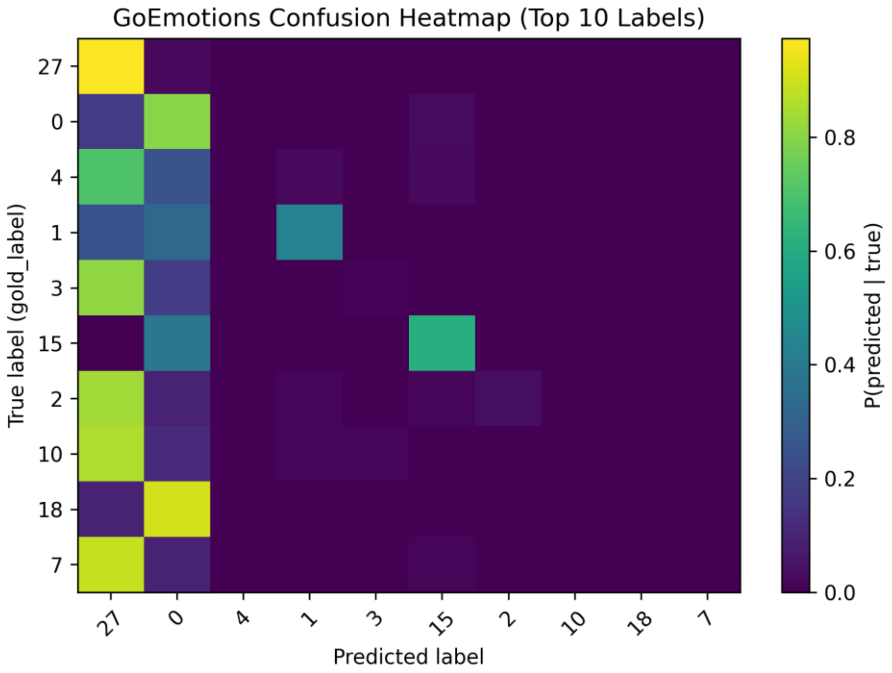
Summary of the results: In-Domain Performance

- Validation metrics on 28 emotions:
  – Accuracy ≈ 0.395 (random baseline ≈ 0.036).
  – Macro F1 ≈ 0.077.
- Training accuracy (first label on 2,000 examples): ≈ 0.423.
- Frequent emotions dominate performance → need to inspect per-label behavior.



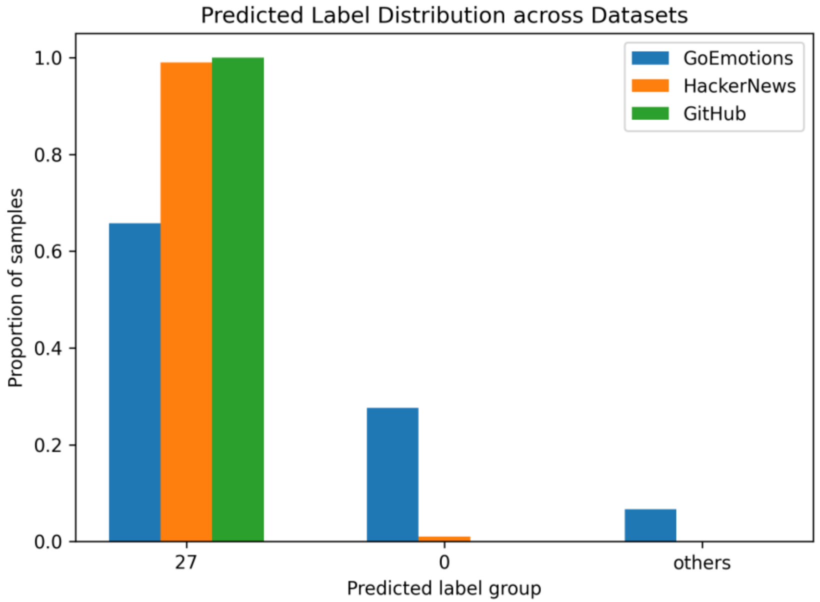
Summary of Results: Per-Label Confusions

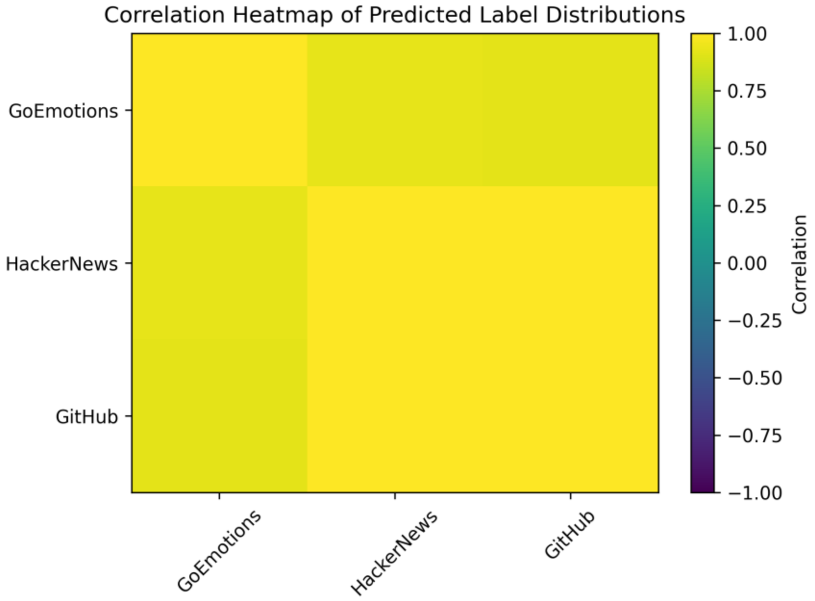
- Confusion heatmap for the 10 most frequent emotions in GoEmotions.
- Neutral (27) and admiration (0) mostly stay on the diagonal.
- Other labels (anger, approval, curiosity, etc.) are often mapped to neutral/admiration → many rare emotions are never predicted correctly.



Summary of Results: OOD Label Distributions

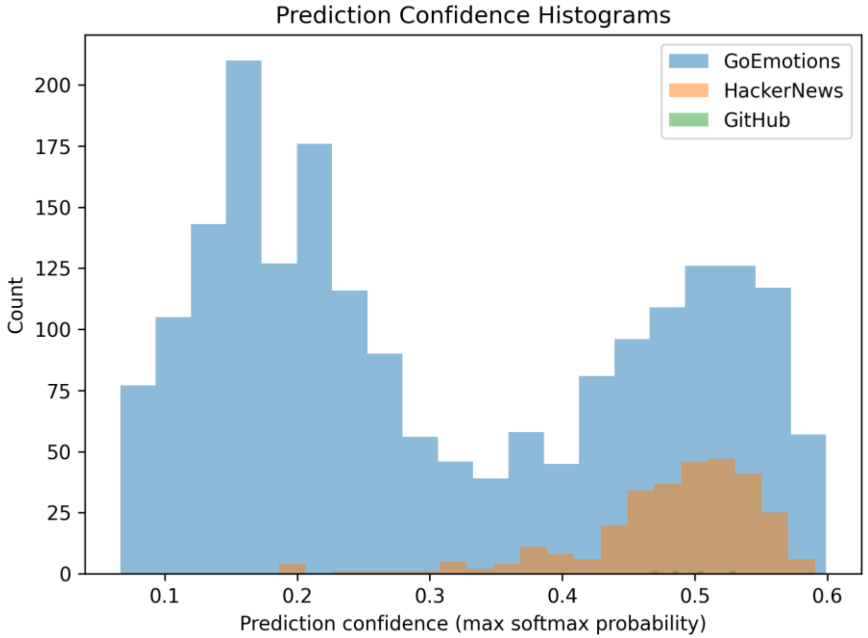
- Predicted label groups (27, 0, others) across datasets.
- GoEmotions still has some variety, but OOD datasets almost collapse to neutral.
- HackerNews: ~99% neutral; GitHub: 100% neutral.
- The correlation heatmap also shows that label distributions are almost identical across domains despite very different text.


Predicted Label Distribution across Datasets


Correlation Heatmap of Predicted Label Distributions

Summary of Results: Prediction Confidence

- Histograms of prediction confidence (max softmax probability).
- GoEmotions: more mid-range confidence values.
- HackerNews & GitHub: many predictions around 0.5–0.6 but still mostly neutral.
- Model is over-confident on unfamiliar OOD text even when it predicts almost only the majority emotion.



Challenges

Data and labels:

- – Only 2,000 GoEmotions examples; strongly imbalanced label distribution.
- – Multi-label dataset simplified to first label only.

Engineering issues:

- – API rate limits and HTTP errors when downloading data.
- – PyTorch / Transformers installation problems on Windows.

Modeling issues:

- – Model collapses many emotions into neutral/admiration, especially on OOD text.
- – Difficult to evaluate calibration and uncertainty with limited time.

Thank you!