CSCI 544 Project Final Report: Emotion recognition

Peiyao Zhu, Yanli Zhang, Henghui Bao, Yingjie Xu, Zitao He
Viterbi School of Engineering
University of Southern California
{peiyaoz, yanlizha, henghuib, yingjiex, zitaohe}@usc.edu

1 Abstract

In addressing the complex challenge of emotion recognition within text, our NLP project advances the field of sentiment analysis by developing a model that identifies 27 distinct emotional states, utilizing the GoEmotions dataset of 58,000 Reddit comments. By fine-tuning state-of-the-art language models such as ALBERT, and XLNet, our work meticulously balances precision and recall for each emotion, reflecting a deep understanding of the nuanced expression of human emotions in digital communication. Despite achieving high accuracy in prevalent emotions such as 'admiration', 'amusement', and 'gratitude', our results indicate the necessity for further refinement in recognizing rare emotions, such as 'grief'. This highlights the need for continuous improvement in feature engineering and model training, reaffirming our commitment to developing a trustworthy and precise emotion classification system that goes beyond traditional sentiment analysis.

2 Introduction

In the digital age, text-based interactions, driven by the prevalence of social media platforms and chatbots, have brought to the forefront the need to decipher underlying emotions within text. Our NLP project embarks on a comprehensive exploration of sentiment analysis, focusing on emotion recognition. Our overarching objective is to develop a highly proficient model capable of effectively detecting and categorizing up to 27 distinct emotional states derived from textual data. We meticulously fine-tune state-of-the-art language models to strike a precise balance between precision and accuracy in emotion classification, recognizing the limitations of traditional sentiment analysis methods.

Social media's transformation into a primary platform for emotional expression underscores the significance of our project. Accurate emotion extraction offers businesses invaluable insights, enabling tailored marketing strategies and enhancing chatbot responsiveness for an improved user experience. Leveraging the GoEmotions dataset, featuring approximately 58,000 Reddit comments, each annotated with multiple emotions, our project has developed and tested two distinct models, ALBERT, and XLNet, while strategically varying maximum sequence lengths to optimize performance. Join us in this journey of unlocking the nuanced language of emotions within textual data, employing the capabilities of Natural Language Processing in today's digital communication landscape.

3 Methods

3.1 Research Design

This project aims to study the performance of pretrained language models, XLNet and ALBERT, on the training results of the GoEmotions dataset.[1] It involves testing combinations of hyperparameter settings for each model to strike a balance in mitigating the varying effects on sentiment analysis introduced by hyperparameter tuning. Ultimately, the project seeks to analyze the associations and impacts of the two language model structures on performance in the context of sentiment analysis.

3.2 Dataset

GoEmotions is the largest manually annotated collection, comprising 58,000 English Reddit comments, meticulously labeled across 27 emotion categories, including Neutral(Demszky et al., 2020). The categories of emotions were identified by Google together with psychologists and include 12 positive, 11 negative, 4 ambiguous emotions, and 1 neutral, which makes the dataset suitable for solving tasks that require subtle differentiation between different emotions.

[1] Data and code available at https://github.com/TaoZiHe/544FinalProject

3.3 Pre-trained Models

3.3.1 XLNet

XLNet is a refined pretraining method that enhances bidirectional context modeling beyond the capabilities of traditional autoregressive language models like BERT(Yang et al., 2020). It employs a generalized autoregressive approach, maximizing likelihood over all factorization order permutations, thus addressing BERT's limitations related to masked positions and pretrain-finetune discrepancies. XLNet's integration of Transformer-XL concepts has demonstrated marked improvements over BERT in tasks like sentiment analysis and question answering.

3.3.2 ALBERT

ALBERT optimizes BERT's model scaling by introducing parameter-reduction techniques, reducing memory use, and speeding up training. It employs a self-supervised loss that improves inter-sentence coherence, enhancing multi-sentence task performance(Lan et al., 2020).

3.4 Experimental Configuration

The experiments were conducted using an NVIDIA GeForce RTX 4070 GPU. This hardware setup was complemented by a system memory of 64 GB. The experimental procedures were implemented using PyTorch. All coding and model execution were carried out within a Jupyter Notebook environment.

3.5 Evaluation Metrics

For the evaluation and analysis of the XLNet and ALBERT models, we utilized the evaluation framework provided by SamLowe(Lowe, 2023). We adapted these tools to conduct a parallel analysis for our XLNet and ALBERT models, ensuring a consistent and robust evaluation methodology. This approach allowed for a direct comparison of results across different models under similar testing conditions.

4 Experiments

4.1 Data Format

The dataset employed in our study utilizes a structured format, as illustrated in Table 1. Each textual input within the dataset is associated with variable labels, capturing the diversity and complexity of the information present in the samples.

4.2 Dataset Size

Our dataset comprises a substantial number of samples for training, validation, and testing purposes. Specifically, we have 43,410 samples designated for training, 5,426 for validation, and 5,427 for testing. This large-scale dataset allows for robust training and evaluation of the models, ensuring a comprehensive analysis of their performance.

4.3 Fine-tuning Experiments

To investigate the effectiveness of our models on the target dataset, we conducted fine-tuning experiments using XLNet and ALBERT architectures. Hugging Face's model was employed for fine-tuning, with the freezing of all layers except the last layer. This strategy aims to leverage pretrained representations while adapting the models to the nuances of the target dataset.

4.4 Model Selection

We opted for XLNet and ALBERT due to their proven capabilities in handling diverse natural language processing tasks. These models offer a balance between computational efficiency and performance, making them suitable candidates for our fine-tuning experiments.

4.5 Fine-tuning Setup

The fine-tuning process involved experimenting with various hyperparameters, including batch size, learning rate, and the number of epochs. We systematically explored different combinations to identify the optimal configuration for each model. The freezing of layers and selective fine-tuning allow us to tailor the models to the specific characteristics of our dataset.

4.6 Batch Size, Learning Rate, and Epochs

In our experiments, we explored the impact of batch size on training efficiency, learning rate on convergence speed, and the number of epochs on overall model performance. The variations in these hyperparameters were systematically tested to provide insights into their influence on the fine-tuning process.

5 Results & Discussion

5.1 Result

We employed two distinct strategies to determine the decision threshold: a consistent threshold of 0.5 and an adaptive threshold fine-tuned for each

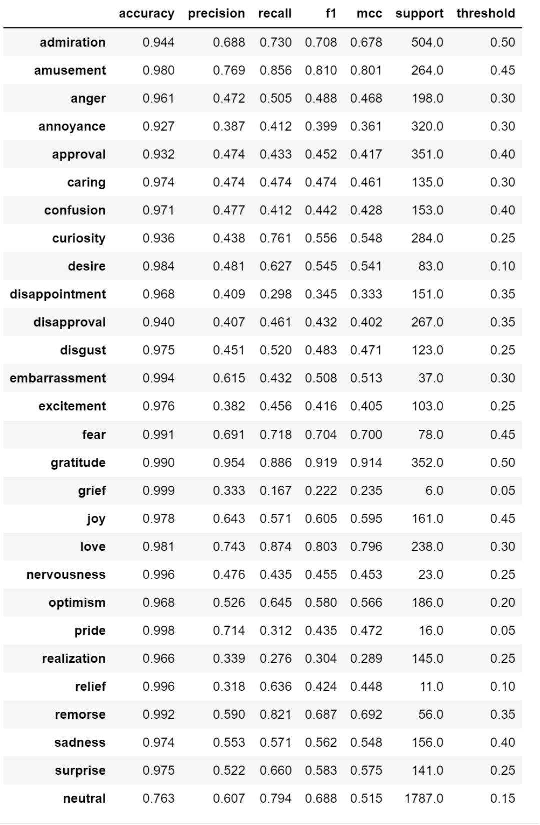
Text	Label
I miss them being alive.	Sadness, Grief.
Ok, then what the actual fuck is your plan?	Anger, curiosity
I'm not saying your analysis is wrong or not based on evidence, just that the data it generates is not useful.	Disapproval, Neutral
Thank you so much! I love Germany! I was in Berlin 2 years ago for research work.	Gratitude, Love

Table 1: Text and emotion labels

specific emotion. Across 27 iterations for each model, the optimal hyperparameter configurations were identified as a maximum sequence length of 32, a learning rate of 2e-05, and a batch size of 16. Among these models, the XLNet-based pre-trained model exhibited the most impressive performance, achieving an F1 score of 0.61.

The F1 score improves by 14.7% when comparing the simple mean of the F1 score between the fixed and variable thresholds. When considering the weighted F1 score, which accounts for the number of instances of each emotion, there's an improvement of 10.5%. These results suggest that optimizing thresholds for each emotion category significantly enhances the model's ability to classify emotions accurately, especially in a dataset with an uneven distribution of emotions. This approach seems particularly effective for improving recall, which indicates a better detection rate of true positive emotions, while also maintaining a balance with precision.

The classifier's performance across various emotions demonstrates a tailored approach, with decision thresholds specifically optimized for each category as shown in Figure 1. For instance, 'curiosity' and 'love' have lower thresholds (0.25 and 0.30 respectively), which corresponds with their higher recall rates (0.761 and 0.874 respectively), indicating that the classifier is set to be more inclusive in identifying potential instances of these emotions. Conversely, 'gratitude', with a higher threshold of 0.50, shows both high precision (0.990) and recall (0.954), indicating that even with a more stringent threshold, the classifier can reliably identify this emotion with fewer false positives.



	accuracy	precision	recall	f1	mcc	support	threshold
admiration	0.944	0.688	0.730	0.708	0.678	504.0	0.50
amusement	0.980	0.769	0.856	0.810	0.801	264.0	0.45
anger	0.961	0.472	0.505	0.488	0.468	198.0	0.30
annoyance	0.927	0.387	0.412	0.399	0.361	320.0	0.30
approval	0.932	0.474	0.433	0.452	0.417	351.0	0.40
caring	0.974	0.474	0.474	0.474	0.461	135.0	0.30
confusion	0.971	0.477	0.412	0.442	0.428	153.0	0.40
curiosity	0.936	0.438	0.761	0.556	0.548	284.0	0.25
desire	0.984	0.481	0.627	0.545	0.541	83.0	0.10
disappointment	0.968	0.409	0.298	0.345	0.333	151.0	0.35
disapproval	0.940	0.407	0.461	0.432	0.402	267.0	0.35
disgust	0.975	0.451	0.520	0.483	0.471	123.0	0.25
embarrassment	0.994	0.615	0.432	0.508	0.513	37.0	0.30
excitement	0.976	0.382	0.456	0.416	0.405	103.0	0.25
fear	0.991	0.691	0.718	0.704	0.700	78.0	0.45
gratitude	0.990	0.954	0.886	0.919	0.914	352.0	0.50
grief	0.999	0.333	0.167	0.222	0.235	6.0	0.05
joy	0.978	0.643	0.571	0.605	0.595	161.0	0.45
love	0.981	0.743	0.874	0.803	0.796	238.0	0.30
nervousness	0.996	0.476	0.435	0.455	0.453	23.0	0.25
optimism	0.968	0.526	0.645	0.580	0.566	186.0	0.20
pride	0.998	0.714	0.312	0.435	0.472	16.0	0.05
realization	0.966	0.339	0.276	0.304	0.289	145.0	0.25
relief	0.996	0.318	0.636	0.424	0.448	11.0	0.10
remorse	0.992	0.590	0.821	0.687	0.692	56.0	0.35
sadness	0.974	0.553	0.571	0.562	0.548	156.0	0.40
surprise	0.975	0.522	0.660	0.583	0.575	141.0	0.25
neutral	0.763	0.607	0.794	0.688	0.515	1787.0	0.15

Figure 1: Model evaluation for each emotion category with different threshold values.

The optimized thresholds highlight the model's adaptability to the intricacies of each emotional expression. The threshold acts as a fine-tuning parameter, balancing the trade-off between precision and recall to enhance overall performance. This adaptive approach is beneficial, especially in an imbalanced dataset where some emotions are less represented than others. However, the stark difference in performance metrics for emotions with low support, such as 'grief' (support of 6), despite an optimized threshold of 0.05, suggests that there are still challenges in recognizing rare emotions that go beyond threshold tuning.

5.2 Discussion

The variable threshold strategy's impact on the emotion classification model's performance is evident from the results. For commonly occurring emotions, the model demonstrates high precision and recall, indicating that the chosen thresholds are highly effective for these categories. This reflects a model that is sensitive to the distinct characteristics of each emotion and can adjust its predictions accordingly.

The poor performance metrics for 'grief' raise

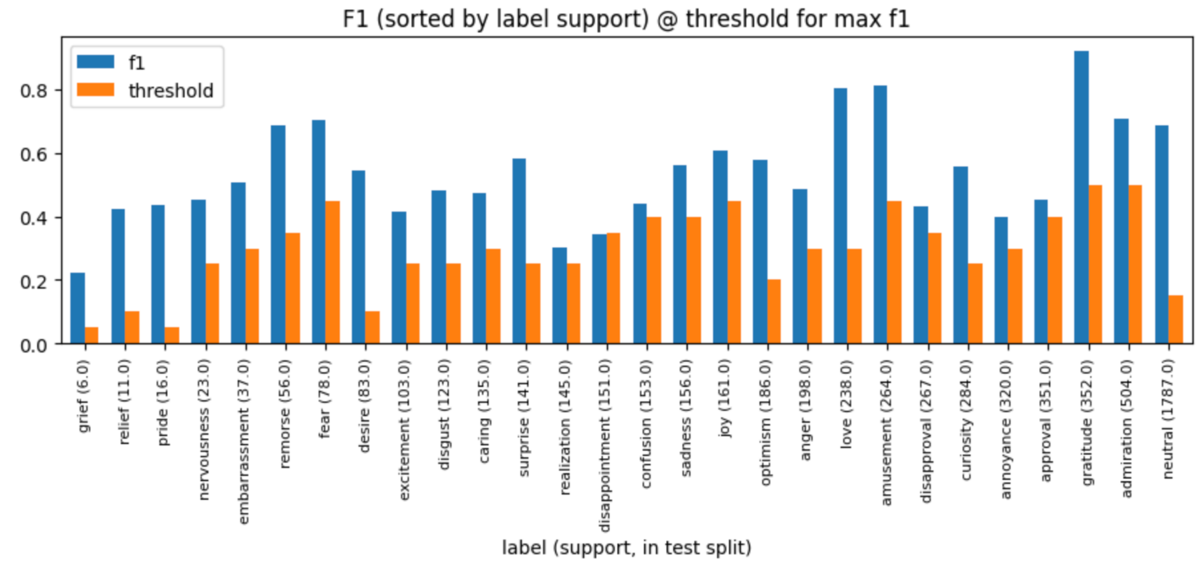


Figure 2: Overall model evaluation report.

critical questions. Despite a tailored threshold intended to maximize detection (set at 0.05), the model's low recall and F1 score indicate a potential overfitting to more common emotions and underrepresentation of this particular emotion in the training data. This could suggest that the model may be inadequately generalizing to less frequent, more complex emotions, which are likely underrepresented in the dataset.

The strong performance for 'gratitude' with an optimized threshold of 0.5 showcases the model's capability to maintain high standards of precision and recall, even with a stringent criterion for classification. This indicates that 'gratitude' is expressed with clear, discernible patterns in the data, which the model can learn and identify effectively.

Overall, the variable threshold approach indicates a sophisticated model calibration technique that can significantly improve performance on individual emotions. However, it also underscores the importance of a balanced dataset, especially for emotions with fewer instances, to ensure the model can generalize well across all categories. The results emphasize the need for continued research and development, especially in the area of rare emotion detection, where current models may still fall short.

6 Conclusions

Our project has embarked on an ambitious journey to explore the depths of emotional expression in text through advanced sentiment analysis techniques. Utilizing the comprehensive GoEmotions dataset and deploying sophisticated language models like XLNet and ALBERT, we have made significant strides in emotion recognition from textual data. Our models' ability to discern and categorize up to 27 distinct emotional states stands as a testament to the efficacy of our approach and the potential of NLP in understanding complex human emotions.

The utilization of the GoEmotions dataset, with its rich and diverse array of annotated Reddit comments, has been instrumental in the development and refinement of our models. Our approach of experimenting with various hyperparameters and selectively fine-tuning model layers has proven effective, as evidenced by our experiments' high precision and recall rates.

One of the key findings is the importance of tailored decision thresholds for each emotion category. This customization allows our models to adapt to the unique characteristics of each emotion, balancing precision and recall to optimize performance. This approach is particularly effective in dealing with an imbalanced dataset, where some emotions are less represented than others.

However, challenges remain, particularly in recognizing rare emotions. The low recall and F1 scores for emotions such as 'grief', despite optimized thresholds, highlight the need for further fine-tuning and innovative feature engineering techniques. This indicates an area for future research and development, emphasizing the need to enhance our dataset and explore new methodologies to better capture these less-represented emotions.

Looking forward, the potential applications of our work are vast. The implications are profound, from improving customer experience through more responsive chatbots to enabling businesses to gain deeper insights into consumer sentiments. Furthermore, our research paves the way for more nuanced and empathetic AI, capable of understanding and interacting with human emotions in a more meaningful manner. Our journey in understanding and interpreting the subtle nuances of human emotion through text continues, promising exciting advancements in the field of NLP.

7 Contributions

Yingjie Xu and Zitao He were tasked with the construction and training of the ALBERT-based and XLNet-based models. Henghui Bao, Yanli Zhang, and Peiyao Zhu were responsible for fine-tuning and optimizing the models. Introduction, Results, Discussion, and Abstract: Yanli Zhang. Method: Yingjie Xu. Experiments: Henghui Bao. Conclusions: Zitao He.

References

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

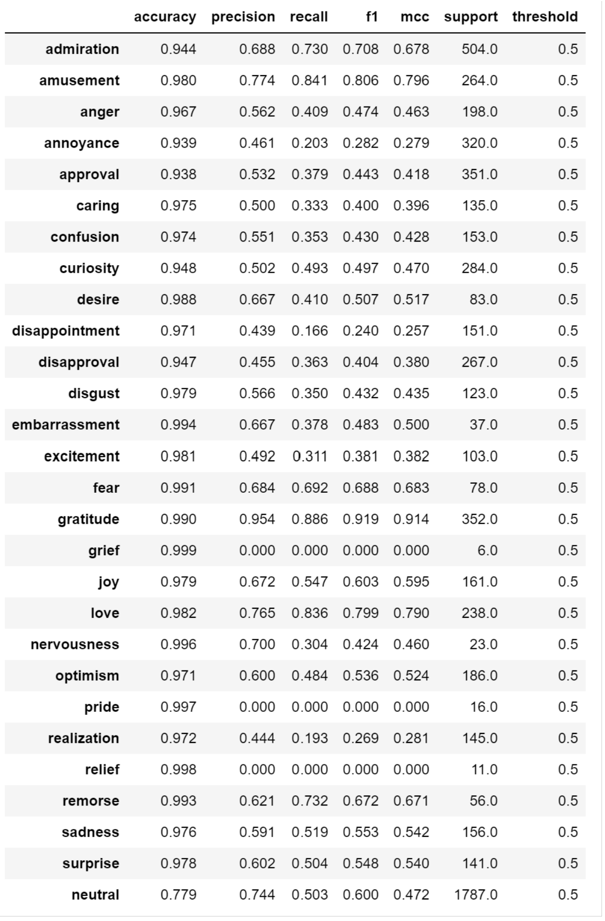
Sam Lowe. 2023. Roberta base model for go emotions. *Hugging Face Model Hub*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

A Appendix: Best Performance

When the threshold is set to 0.5, simple mean of labels: 'precision': 0.544, 'recall': 0.426, 'f1': 0.468, 'mcc': 0.46; Weighted average (using support): 'precision': 0.647, 'recall': 0.509, 'f1': 0.561, 'mcc': 0.518

Threshold from per label search (for best F1 score), simple mean of labels: 'precision': 0.533, 'recall': 0.562, 'f1': 0.537, 'mcc': 0.522; Weighted average (using support): 'precision': 0.576, 'recall': 0.658, 'f1': 0.61, 'mcc': 0.548



	accuracy	precision	recall	f1	mcc	support	threshold
admiration	0.944	0.688	0.730	0.708	0.678	504.0	0.5
amusement	0.980	0.774	0.841	0.806	0.796	264.0	0.5
anger	0.967	0.562	0.409	0.474	0.463	198.0	0.5
annoyance	0.939	0.461	0.203	0.282	0.279	320.0	0.5
approval	0.938	0.532	0.379	0.443	0.418	351.0	0.5
caring	0.975	0.500	0.333	0.400	0.396	135.0	0.5
confusion	0.974	0.551	0.353	0.430	0.428	153.0	0.5
curiosity	0.948	0.502	0.493	0.497	0.470	284.0	0.5
desire	0.988	0.667	0.410	0.507	0.517	83.0	0.5
disappointment	0.971	0.439	0.166	0.240	0.257	151.0	0.5
disapproval	0.947	0.455	0.363	0.404	0.380	267.0	0.5
disgust	0.979	0.566	0.350	0.432	0.435	123.0	0.5
embarrassment	0.994	0.667	0.378	0.483	0.500	37.0	0.5
excitement	0.981	0.492	0.311	0.381	0.382	103.0	0.5
fear	0.991	0.684	0.692	0.688	0.683	78.0	0.5
gratitude	0.990	0.954	0.886	0.919	0.914	352.0	0.5
grief	0.999	0.000	0.000	0.000	0.000	6.0	0.5
joy	0.979	0.672	0.547	0.603	0.595	161.0	0.5
love	0.982	0.765	0.836	0.799	0.790	238.0	0.5
nervousness	0.996	0.700	0.304	0.424	0.460	23.0	0.5
optimism	0.971	0.600	0.484	0.536	0.524	186.0	0.5
pride	0.997	0.000	0.000	0.000	0.000	16.0	0.5
realization	0.972	0.444	0.193	0.269	0.281	145.0	0.5
relief	0.998	0.000	0.000	0.000	0.000	11.0	0.5
remorse	0.993	0.621	0.732	0.672	0.671	56.0	0.5
sadness	0.976	0.591	0.519	0.553	0.542	156.0	0.5
surprise	0.978	0.602	0.504	0.548	0.540	141.0	0.5
neutral	0.779	0.744	0.503	0.600	0.472	1787.0	0.5

Figure 3: Per-label metrics with constant 0.5 thresholds

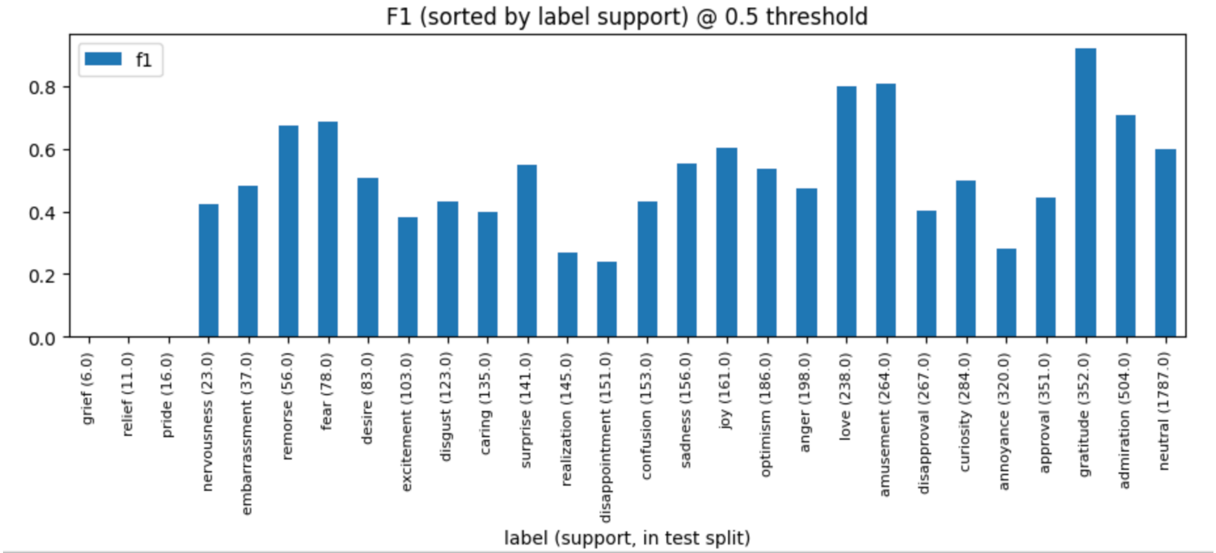


Figure 4: Result for model evaluation. Per-label metrics with constant 0.5 thresholds

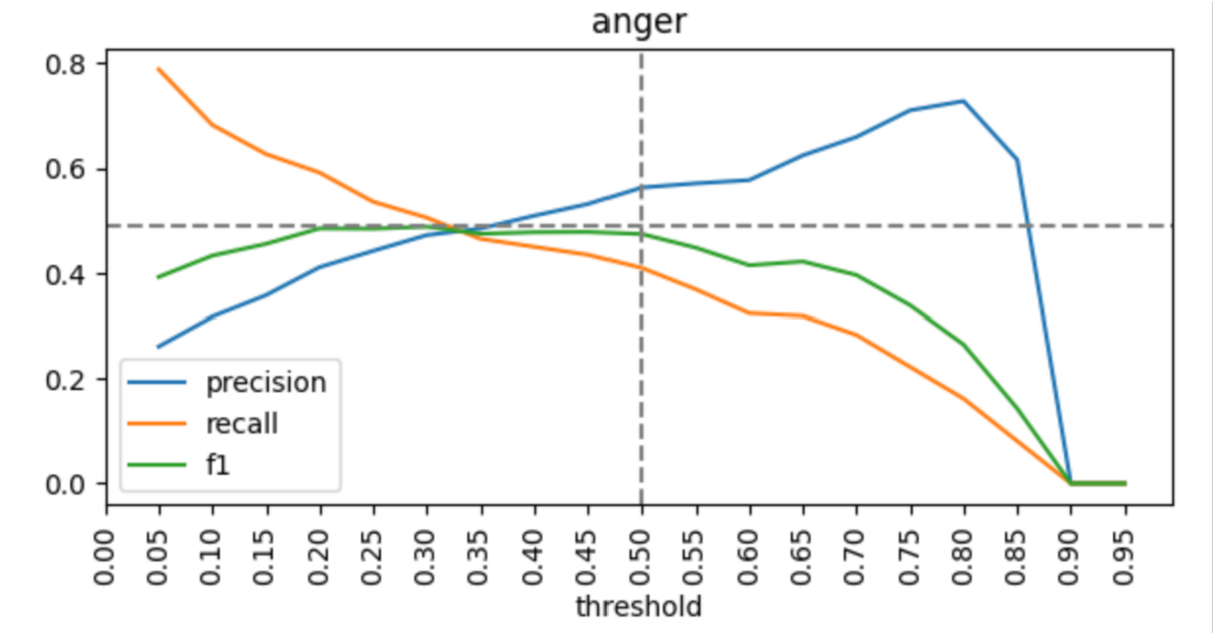


Figure 5: Example threshold and f1 relationship.