Title: Model Implementation & Hyperparameter optimization using One Versus Rest Classifier with the estimator of Logistic Regression on Multi Label Dataset Problems, such as GoEmotion and DairAI

Author & The Course: Emre Erdoğan - COMP4602 Natural Language Processing

Abstract:

Emotion classification in natural language processing often suffers from severe class imbalance, where frequent emotions dominate the training process and rare emotions are underrepresented. This issue becomes more critical in multi label settings, as a single text can express multiple emotions simultaneously. This study investigates the impact of class imbalance on multi label emotion recognition and evaluates strategies to improve rare emotion detection. Two publicly available datasets were used: GoEmotions and the dair-ai/emotion Twitter dataset, which were reformatted into a unified multi label representation. A baseline machine learning pipeline using TFIDF vectorization and a OnevsRest Logistic Regression classifier was implemented, followed by hyperparameter optimization using RandomizedSearchCV. Model performance was evaluated using label wise metrics, primarily Hamming loss, to avoid the limitations of subset accuracy. The results show that hyperparameter optimization significantly improved label level accuracy, increasing from 0.960 to 0.990. Interestingly, the best performing configuration did not rely on class weight balancing, suggesting that weighting may increase false positives for rare emotions in this setting. These findings highlight the importance of imbalance aware evaluation and careful metric selection in multi label emotion classification tasks.

Introduction:

Emotion classification in natural language processing (NLP) is often challenged by the fact that real world datasets are rarely clean and balanced. In many emotion datasets, a small number of emotions appear very frequently, while several important emotions occur only in a limited number of samples. When standard machine learning models are trained under this imbalance, they tend to optimize performance on frequent labels and struggle to detect rare ones, which can introduce systematic bias into predictions. This issue becomes even more critical in multi label emotion recognition, where a single text may express multiple emotions simultaneously and rare emotions can carry key information about the deeper context and meaning of the text.

Motivated by the challenges posed by class imbalance in multi label emotion recognition, this paper examines how imbalance affects the detection of rare emotions and evaluates whether a baseline machine learning pipeline can be improved through hyperparameter optimisation and imbalance aware design choices. To this end, two publicly available datasets are used: GoEmotions (Reddit) and dair-ai/emotion (Twitter). GoEmotions contains approximately 58,000 Reddit text samples annotated with 27 emotion labels plus a neutral label (28 labels in total), allowing multiple emotion labels per instance. In contrast, the dair-ai/emotion dataset contains approximately 20,000 Twitter texts labeled with six basic emotions and is originally formatted for multi class classification using numerical labels from 0 to 5. Although the datasets differ in both domain and labeling scheme, their emotion categories overlap conceptually, enabling joint analysis after

appropriate preprocessing and label alignment.

To enable consistent modelling, the dair-ai/emotion labels were reformatted into a binary representation aligned with the GoEmotions label structure (0/1 columns per emotion). Both datasets were then preprocessed and used within a unified modelling framework. The baseline approach was implemented as a pipeline consisting of TFIDF vectorization followed by a OnevsRest Logistic Regression classifier, where each emotion label is treated as a separate binary classification problem. To improve performance, the same pipeline structure was tuned using RandomizedSearchCV, searching over key vectorizer parameters and classifier settings.

The key findings indicate that hyperparameter optimization improved label level performance substantially. Using label wise accuracy (1 − Hamming loss) as a primary metric, the baseline model achieved 0.960, while the tuned model reached 0.990. Interestingly, the best-performing configuration did not require [class_weight="balanced"]. A possible explanation is that applying class weights made the model more likely to predict rare emotions, which can increase false positives and slightly reduce overall label level correctness. These results reinforce the importance of imbalance aware evaluation and careful metric selection in multi label emotion classification, especially when the objective is to improve detection quality for rare emotions rather than maximise overall accuracy alone.

Related Work:

Prior research in multi label learning and emotion recognition consistently highlights two recurring challenges: severe label imbalance, where frequent emotions dominate learning, and evaluation ambiguity, where strict instance level metrics can understate meaningful partial correctness. Within this literature, common approaches range from simple problem transformation baselines to more complex modelling and imbalance handling strategies, yet reported gains can depend strongly on dataset specific properties and the chosen metric. Building on these observations, this project contributes a focused empirical study that frames rare emotion detection as an imbalance sensitive multi label problem, aligns datasets with different annotation schemes into a consistent multi label formulation for comparable analysis, and emphasizes label level evaluation to separate “all or nothing” correctness from per label behavior. In doing so, the work complements existing research by demonstrating how methodological choices that are often treated as implementation details such as the evaluation metric and imbalance handling can materially change conclusions about model quality, especially for rare emotions.

Methodology:

This project uses two emotion datasets from different social media platforms: GoEmotions (Reddit) and dair-ai/emotion (Twitter). Because the two datasets use different labeling formats, the first step is to make them compatible under a single multi label learning setup. GoEmotions already supports multi label annotation and is represented as separate binary columns per emotion (1 if the emotion is present, 0 otherwise) across 28 labels (27 emotions plus Neutral). In contrast, the Twitter dataset is originally multi class and stores its label as a single integer from 0 to 5, representing six emotions (sadness, joy, love, anger, fear, surprise). To align this dataset with the GoEmotions

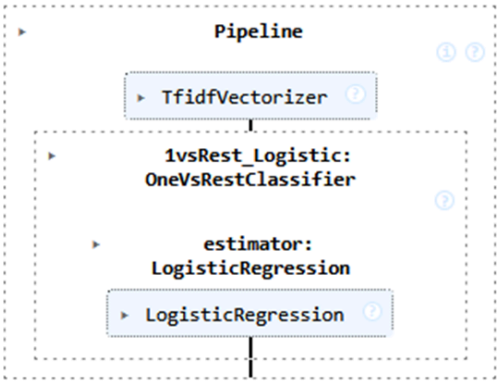
representation, each Twitter label is mapped to its corresponding emotion name and then converted into a multi label binary format by creating the same emotion columns and assigning a 1 only to the corresponding label for each text. After this alignment, both datasets follow the same schema: one text column and multiple binary label columns.

After preprocessing, the data is split into training and testing. GoEmotions is split using an 80/20 ratio with a fixed random seed for reproducibility. The training set is then extended by concatenating it with the Twitter training set, allowing the model to learn from both domains under the same multi label format.

The baseline model is implemented as a simple but strong classical machine learning pipeline. The text is first converted into numerical features using TFIDF vectorization, which represents each document as a sparse vector based on how informative each word (or word sequence) is across the dataset. These TFIDF features are then passed into a OnevsRest classification framework using Logistic Regression as the base estimator. OnevsRest is a standard approach for multi label problems because it trains one independent binary classifier per label, meaning each emotion is predicted separately as present (1) or absent (0). Logistic Regression is appropriate here because the target for each label is binary and the method performs well with sparse TFIDF features. The model is configured with a high iteration limit to ensure stable convergence.

To improve performance beyond the baseline, hyperparameter optimization is performed using RandomizedSearchCV. Rather than trying every possible parameter combination, this method samples a fixed number of random configurations from a defined search space and evaluates each configuration using cross validation. In this project, 20 random configurations are evaluated with 3 fold cross validation, leading to 60 total model fits. The hyperparameter search includes TFIDF settings such as n-gram range (whether to use single words only or include word pairs), minimum document frequency (filtering out extremely rare tokens), and maximum document frequency (filtering out overly common tokens). On the classifier side, the search includes the regularization strength of Logistic Regression and whether to apply class weighting. Class weighting is included as a potential imbalance handling mechanism because it can increase the importance of rare labels during training; however, it can also increase false positives depending on the dataset and feature setup.

Evaluation is performed using metrics suitable for multi label classification. Standard “accuracy” in multi label settings is often subset accuracy, which is very strict because a prediction is only counted correctly when the entire set of labels for a sample matches exactly. This can produce very low scores even when a model is doing reasonably well on most labels. Therefore, this project emphasizes label level evaluation using Hamming loss, which measures the fraction of individual label decisions that are wrong across all labels and samples. Label level accuracy is then interpreted as one minus the Hamming loss. This metric is more informative for multi label emotion recognition because it credits partially correct predictions and allows performance comparison at the label decision level rather than an all or nothing sample level.



Results and Analysis:

Model performance was evaluated using complementary metrics to reflect the nature of multi label emotion classification under class imbalance. First, subset accuracy (exact match accuracy) was reported, where a prediction is counted as correct only if the entire 28 label set matches the ground truth; this is a strict form of metric because predicting 27 labels correctly but missing a single label marks the whole sample as incorrect. In addition, label wise evaluation was performed using Hamming loss, which measures the fraction of incorrect label decisions across all labels and samples; label level accuracy is then interpreted as (1 - Hamming loss).

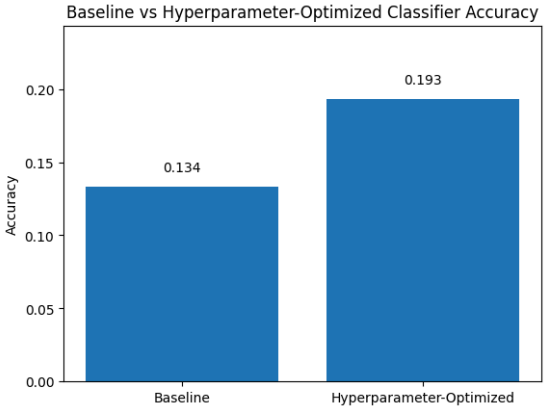
To reveal systematic misunderstandings, a confusion analysis was also conducted. Since a standard single “confusion matrix” is not directly defined for multi label tasks, confusion was visualized in a multi label appropriate way by examining label level error patterns: for each label, binary confusion counts (true positives, false positives, false negatives, true negatives) were computed, and the most frequent confusion pairs were summarized (e.g., cases where one emotion is predicted while a semantically similar emotion is present in the ground truth). This helps identify consistent error trends such as confusion between closely related emotions and over prediction of frequent labels.

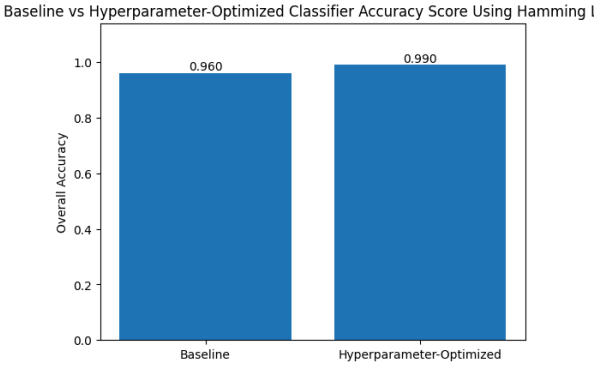
Finally, to properly evaluate performance under imbalance, per class precision, recall, and F1 score were reported for each label. These label wise metrics are crucial because aggregate scores can hide failures on rare emotions: precision indicates how often predicted positives are correct, recall indicates how often true positives are detected, and F1 summarizes the trade off between them. Reporting per class metrics makes it possible to explicitly observe which rare emotions remain difficult and whether improvements come from better recall, better precision, or both.

```
Baseline accuracy: 0.1335

Baseline classification report:
```

	precision	recall	f1-score	support
admiration	0.68	0.26	0.37	3443
amusement	0.63	0.29	0.39	1921
anger	0.54	0.08	0.14	1619
annoyance	0.38	0.02	0.04	2701
approval	0.53	0.03	0.05	3465
caring	0.55	0.05	0.08	1195
confusion	0.55	0.04	0.07	1522
curiosity	0.52	0.05	0.10	1995
desire	0.45	0.05	0.10	771
disappointment	0.58	0.02	0.04	1674
disapproval	0.34	0.02	0.03	2338
disgust	0.63	0.07	0.13	1055
embarrassment	0.50	0.03	0.05	481
excitement	0.48	0.05	0.08	1122
fear	0.67	0.16	0.26	620
gratitude	0.92	0.70	0.80	2303
grief	0.33	0.01	0.01	139
joy	0.49	0.11	0.18	1594
love	0.70	0.34	0.45	1652
nervousness	0.42	0.01	0.03	356
optimism	0.60	0.15	0.24	1719
pride	0.57	0.01	0.03	274
realization	0.55	0.01	0.03	1794
relief	0.50	0.01	0.02	252
remorse	0.48	0.13	0.21	499
sadness	0.58	0.12	0.19	1336
surprise	0.52	0.10	0.16	1154
neutral	0.57	0.19	0.29	10919
micro avg	0.64	0.15	0.24	49913
macro avg	0.54	0.11	0.16	49913
weighted avg	0.56	0.15	0.21	49913
samples avg	0.17	0.16	0.16	49913

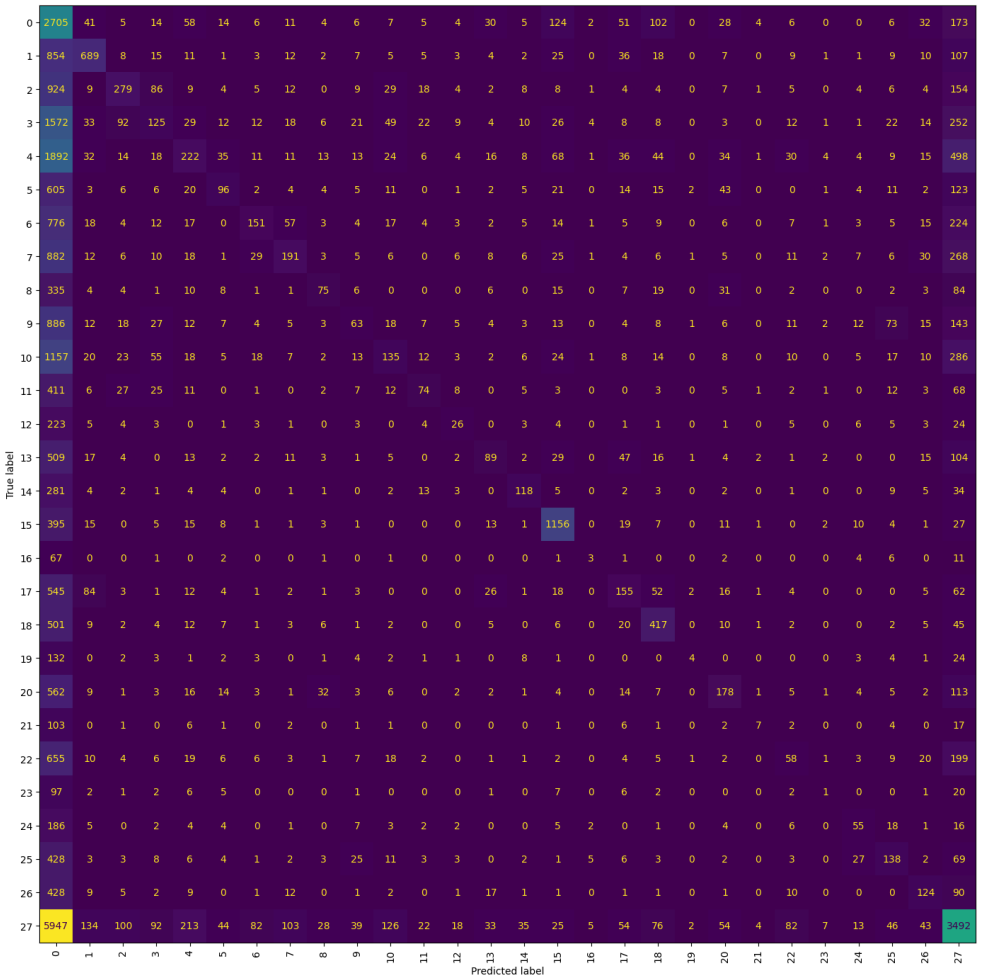




Test accuracy (best model): 0.1935

Optimized model classification report:

	precision	recall	f1-score	support
admiration	0.62	0.35	0.45	3443
amusement	0.60	0.38	0.46	1921
anger	0.45	0.17	0.25	1619
annoyance	0.31	0.07	0.11	2701
approval	0.30	0.07	0.12	3465
caring	0.37	0.10	0.15	1195
confusion	0.46	0.11	0.18	1522
curiosity	0.45	0.12	0.19	1995
desire	0.41	0.11	0.17	771
disappointment	0.27	0.04	0.08	1674
disapproval	0.31	0.07	0.12	2338
disgust	0.46	0.10	0.16	1055
embarrassment	0.41	0.10	0.16	481
excitement	0.36	0.10	0.15	1122
fear	0.54	0.22	0.31	620
gratitude	0.89	0.72	0.79	2303
grief	0.15	0.03	0.05	139
joy	0.43	0.19	0.26	1594
love	0.64	0.41	0.50	1652
nervousness	0.39	0.03	0.06	356
optimism	0.53	0.19	0.28	1719
pride	0.53	0.07	0.12	274
realization	0.29	0.06	0.09	1794
relief	0.12	0.02	0.03	252
remorse	0.47	0.19	0.27	499
sadness	0.48	0.20	0.28	1336
surprise	0.44	0.18	0.25	1154
neutral	0.52	0.33	0.40	10919
micro avg	0.52	0.22	0.31	49913
macro avg	0.44	0.17	0.23	49913
weighted avg	0.47	0.22	0.29	49913
samples avg	0.25	0.24	0.24	49913



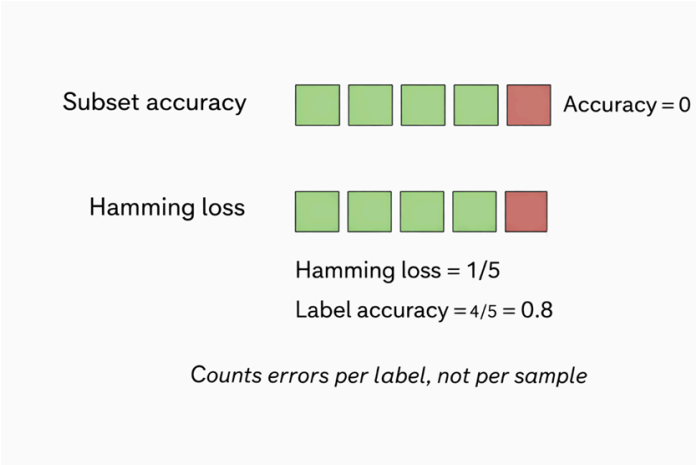
Discussion and Future Work:

The experiments show that hyperparameter tuning substantially improved performance in multi label emotion classification under class imbalance. Using label wise accuracy (defined as 1 - Hamming loss), the baseline pipeline achieved 0.960, while the optimized configuration reached 0.990. This metric choice is important because subset accuracy (standard accuracy_score) marks an entire sample as incorrect if any label is wrong, which can be overly strict for multi label settings. In contrast, Hamming loss evaluates errors at the individual label level, making it more informative for assessing overall label correctness in datasets where each instance has many possible labels.

An interesting outcome is that the best performing setup did not use [class_weight="balanced"]. Class weighted training changes the loss to penalize minority class mistakes more heavily (typically using inverse frequency weighting), which often encourages the model to predict rare labels more frequently. While this can improve recall for underrepresented emotions, it can also increase false positives, reflecting the well known precision recall trade off (pushing toward higher recall can reduce precision by introducing more false positives). In this project's feature space (TFIDF) and decision rule, the additional false positives appear to slightly reduce overall label-level correctness, explaining why [class_weight="none"] performed better.

However, several limitations remain. First, TFIDF with linear classifiers does not capture deeper context, sarcasm, or long range dependencies, which are common in emotion heavy text. Second, relying primarily on Hamming loss based accuracy can hide label specific weaknesses; rare emotions may still perform poorly even when overall label level correctness is high. Finally, the two datasets come from different platforms (Reddit vs. Twitter), so domain differences may affect generalization.

Future work should prioritize deep learning models that better encode context, such as fine tuned transformer based architectures, and compare them against the current TF IDF baseline. Additionally, imbalance aware approaches can be explored beyond class weighting, including threshold calibration per label and loss functions designed for multi-label imbalance (e.g., asymmetric loss variants). A deeper error analysis should also be included, reporting per class precision/recall/F1 and inspecting systematic confusions to identify which emotions remain difficult and why.



Conclusion:

This project investigated class imbalance in multi label emotion recognition using GoEmotions and the dair-ai/emotion Twitter dataset under a unified multi label format. A TFIDF + OnevsRest Logistic Regression baseline was implemented and then improved through hyperparameter optimization, which increased label wise performance measured by (1 − Hamming loss) from 0.960 to 0.990. The best configuration did not require [class_weight="balanced"], suggesting that weighting can increase false positives for rare emotions and reduce overall label level correctness in this feature setup. Overall, the findings emphasize that both imbalance handling and metric selection strongly affect conclusions in multi label emotion classification, and that label wise evaluation provides a more informative view than strict exact match accuracy for real world, imbalanced emotion datasets.

References:

*A Review on Multi-Label Learning Algorithms by Min-Ling Zhang - https://www.researchgate.net/publication/263813673_A_Review_On_Multi-Label_Learning_Algorithms

*GoEmotions: A Dataset of Fine-Grained Emotions - https://arxiv.org/pdf/2005.00547

*dair-ai/emotion - https://huggingface.co/datasets/dair-ai/emotion

*Asymmetric Loss For Multi-Label Classification - https://arxiv.org/pdf/2009.14119