Fine-grained Emotion Detection Modelling Improvements

Pedro Forli

UC Berkeley Data SCI

pedroforli@berkeley.edu

Abstract

Emotion detection is a challenging task aimed at identifying emotions from text. This paper focuses on enhancing GoEmotions fine-grained classification by selecting a new baseline model and employing a novel approach to the problem reaching for a higher macro F1-Score. We assess public transformers and published papers to establish a new baseline to improve of 0.53. Our approach aims at refining this target by evaluating transfer learning from colloquial language transformers and applying various balancing techniques, including oversampling, augmenting, and minority shuffling. Text cleaning techniques, such as Wordnet lemmatization and Porter stemming, were also tested. Our best model reached a F1-Score of 0.54. Notably, minority shuffling significantly contributes to the model's improved performance. However, further advancements can be explored through larger networks, hyperparameter tuning, and fine-tuning individual class combination models.

1. Introduction

Emotion detection in NLP is about identifying a person's emotional state from their text [1]. Techniques like sentiment analysis and deep learning have been steadily improving the ability of machines to provide accurate classification of these states. For example, [2] achieved 92% accuracy in identifying emotions from social media.

To run emotion detection models, proper emotion naming and classification are crucial. The traditional approach has been to use the Ekman taxonomy [3], which identifies six basic emotions: happiness, sadness, anger, fear, disgust, and surprise. Improving on that, Google's GoEmotions dataset [4] was designed to offer a more fine-grained and comprehensive range of emotions, expanding to 28 emotional states.

However, this expansion comes at a cost. The baseline BERT model achieved by the paper [4] has a macro F1-score of 0.46, far from the reported performances using data with the Ekman taxonomy (for example, 0.81 F1-score for [5]).

Limited work has been done to improve this on the full range of emotions, with few pre-trained models displaying higher F1-score on Hugging Face [6].

This paper aims to select a new baseline for GoEmotions classification, evaluate novel approaches, such as balancing and text cleaning, to reach a higher macro F1-Score.

2. Background

Most of the work done using GoEmotions dataset has been developed aiming at further evaluation of algorithms performance within the context of Ekman's classification. For instance, [7] conducted a comparative analysis of various algorithms, assessing their effectiveness on a subset of datasets, including GoEmotions.

Certain research papers have excluded specific emotions from the dataset for focused emotion detection or loss evaluation purposes. For example, [8] explored contrastive fine-tuning of pre-trained language models, achieving an impressive F1-score of 0.63. However, it is important to consider that their evaluation on the GoEmotions dataset omitted

the "Neutral" class, potentially affecting the classifier's real-world accuracy.

Nevertheless, some studies have successfully advanced fine-grained detection models using GoEmotions full class range. [9] employed pooled DNN and Bi-LSTM built on top of BERT, which surpassed previous baseline with a 0.48 F1-score. [10] proposed a sequence-to-emotion model with a bi-directional decoder, leading to a performance improvement with a 0.47 F1-score. [11] integrated external knowledge into a pre-trained self-attention model using Knowledge-Embedded Attention (KEA), which effectively incorporates information from emotion lexicons to enhance the contextual representations derived from pre-trained ELECTRA and BERT models. This framework achieved notable improvements with a 0.50 F1-score.

Furthermore, [12] presented a multi-task learning framework for fine-grained emotion prediction, combining Class Definition Predictions (CLP) and Masked Language Modelling (MLM) reaching a 0.52 F1-score. [13] aimed to improve the BERT model for emotion detection by inducing sentiment/ emotion-specific biases into the model through eMLM (Emotion Masked Modelling), achieving a performance of 0.48 F1-score.

Recent advancements in few and zero-shot learning led to evaluations of their effectiveness in this classification task. [14] explored few-shot emotion recognition by transferring knowledge from GoEmotions to SemEval corpus, and in this process achieved 0.49 F1-score using SBERT, while [15] explored zero-shot learning with a variety of model, showing low accuracy performance for GoEmotions.

Additionally, [16] proposed the CUE framework to interpret uncertainties in Pre-trained Language Models, reaching 0.49 F1-score. Lastly, [17] evaluated ChatGPT's performance on GoEmotion, revealing its limited accuracy compared to specialized classification approaches.

3. Methods

3.1. Data Analysis

To initiate our project, we will start by analyzing the data. Upon examining the sentences, we ascertained that they predominantly made of concise lengths (< 200 characters) and limited word counts (< 35 tokens), as evidenced in Figure 1.

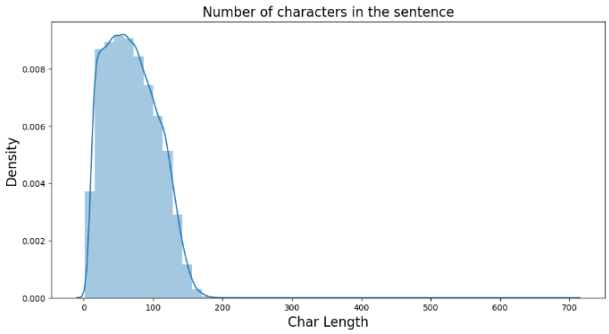

Figure 1: Distribution of sentences by the number of characters and tokens in it

Additionally, it is noteworthy that these sentences encompass a blend of emojis, abbreviations, and colloquial expressions, which aligns with our expectations, given their extraction from Reddit, as exemplified below.

"I hate when these crooks say it ain't so cuz it is FR 😳 ***"***

On Figure 2 we can see that there are some imbalances in terms of emotion representation within the dataset. This will be important, given that on [4] it was shown that these lower represented classes had close to 0 F1-score, and might require some specialized approach.

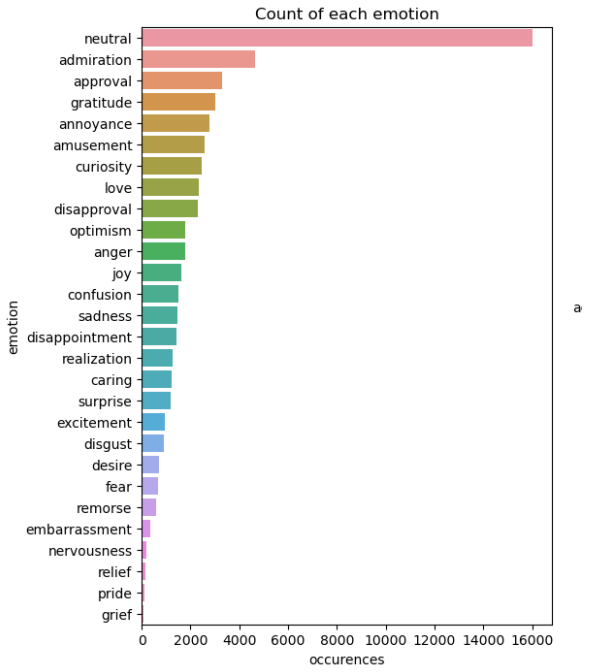

Figure 2: Distribution of sentences by emotion classification

Moreover, these classifications are not mutually exclusive, as seem in Figure 3. Around 16% of all sentences have more than one classification.

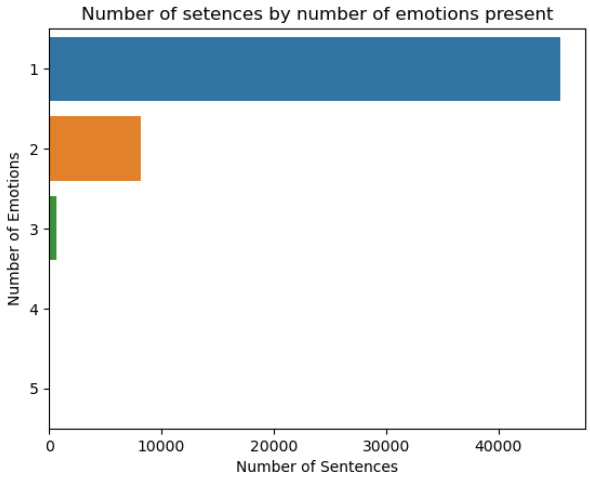

Figure 3: Number of sentences by number of emotions detected

3.2. Selection of a Baseline

In this study, we initially established a baseline model using BERT base cased with a dense output layer and a hidden layer with 128. The model's performance, evaluated with a two-epoch execution and a batch size of 16 at a learning rate of 2e-5, yielded a macro F1-score of 0.46, consistent with the results reported in the GoEmotions initial paper [4].

Our analysis also examined the impact of data imbalance, revealing certain classes like "grief" and "pride" with significantly lower F1-scores due to limited representation within the dataset (refer to Table 4 in the appendix).

To account for recent advancements in pre-trained models on Hugging Face, we evaluated their efficacy using the macro average F1-score on the same test set. Encouragingly, the publicly available models outperformed the previous baseline, providing a new and improved starting point for model refinement (refer to Table 1)

Table 1: Macro average F1-score by model (threshold of 0.2)

Model	F1-Score
BERT Fine-tuned[1]	0.532
DistilBERT Fine-tuned	0.442
RoBERTa Base	0.499
RoBERTa Large	0.534
RoBERTa BNE	0.348
EmoRoBERTa	0.446
EmoGPT	0.502

Upon examining of the individual models' performance, is observed that all exhibit misclassification errors predominantly in the context of longer sentences (higher average token count). However, it is noteworthy that the RoBERTa Large and EmoRoBERTa models stand out as the sole performers at effectively addressing the lower-represented class, thereby attaining a more balanced confusion matrix overall.

[1] We could not replicate the model execution, given compatibility issues

3.3. Experiments

Given our initial performance baseline and background, we conducted an evaluation of potential strategies aimed at enhancing the model's overall performance. The following approaches were considered:

1. **Leveraging Transfer Learning**: Inspired by the success demonstrated in [18], we adopted a transfer learning approach using prominent models like RedditBERT and BERTweet [19]. Building upon the knowledge acquired during pre-training on large corpora, these models were fine-tuned on our specific task.
2. **Implementation of Balancing Techniques**: Addressing the challenge posed by underrepresented classes, we explored methodologies proposed in [20]. These techniques included merging minority classes or applying round-trip translation to augment the data and enhance model performance for such classes.
3. **Text Cleaning Techniques**: To further enhance the quality of the input data, we applied various text cleaning techniques. Specifically, we tested character normalization, lemmatization, stemming, and emoji tagging to ensure a consistent and refined dataset for training and evaluation.

Importantly, these approaches were not mutually exclusive, and we explored their combination to identify potential synergistic effects. This was achieved through an iterative process that assessed which changes to make to the model based on its performance. In the appendix we layed out how we went about running the experiments and computing the results.

4. Results and Discussion

From our baseline (BERT with 2 epochs) we analyzed the model results and found that training loss was increasing over time. To address this issue we removed the hidden layer and added a learning rate scheduler that allowed us to reach better performance with healthier loss behaviour.

Subsequently, we conducted an analysis of misclassification metrics and discovered a notable prevalence of emojis in the misclassified data, with an incidence rate over 60 times higher than in correctly classified instances. Despite the relatively low occurrence of emojis in the overall dataset (only 0.3% of tokens), we investigated the impact of introducing emoji tokens on the performance of BERT. While this modification did not yield significant improvements, it did lead to a minor enhancement in model performance.

Additionally, we examined the influence of uppercase letters on model performance. Given that there were no significant differences in upper case representation between correctly and incorrectly classified data, we made the decision to transition from using the cased BERT model to the uncased variant, which led to no discernable impacts on performance.

Given the informal nature of the conversational data, we further explored alternative pre-trained models: BERTweet [19], given that researchers have found that these variants outperformed BERT-base in Tweet NLP classes; and RedditBERT, which has been pre-trained on reddit posts. However, these models failed to surpass the baseline performance. Moreover, both models demonstrated similar performance in terms of class unbalancing and misclassification metrics

compared to the BERT-uncased model, with the latter slightly outperforming in predicting minority classes.

Once we identified our new best-performing model we focused on exploring balancing techniques for improving the minority classes performance. We initially attempted a simple oversampling method, randomly selecting samples to match the majority class representation. This, however, resulted in massive decreases, given deteriorated accuracy for the majority classes.

To address this issue, we attempted to balance minority classes to achieve the average number of data points per class (~2k). While this approach did not increase the overall model performance, by analyzing the confusion matrix (Figure 4) we see that we enhanced the model's ability to predict minority classes, particularly for grief and pride. However, further exploring loss curves reveled that we had a more severe overfitting problem compared against the unbalanced model variant.

To further improve balancing, we decided to test augmentation. The core idea is that we could reduce overfitting by applying technics that would slightly change the text while keeping its overall meaning [21]. For our model we applied round-trip translation using a selection of >10 languages. This had our highest performance yet, while decreasing performance for some majority classes.

As an alteranative to oversampling, we also tested minority shuffling. For this goal, we tested combinations of some minority classes (grief, relief, pride, annoyance and nervousness). These approaches yielded good results, having increased accuracy minority all minority classes selected substantially, leding to a new benchmark model of 0.539 F1-score.

Table 2: Summary of experiments results

Model	Epochs	Scheduler	Emoji Tag	Balancing	Cleaning	F1-Score
BERT cased	2	×	×	×	×	0.459
BERT cased	10	✓	×	×	×	0.482
BERT cased	10	✓	✓	×	×	0.487
BERT uncased	10	✓	✓	×	×	0.490
BERTweet	10	✓	✓	×	×	0.479
RedditBERT	10	✓	✓	×	×	0.489
BERT uncased	10	✓	✓	R.O.M.	×	0.112
BERT uncased	10	✓	✓	R.O.A.	×	0.486
BERT uncased	10	✓	✓	A.O.A.	×	0.502
BERT uncased	10	✓	✓	M.S.	×	0.539
BERT uncased	10	✓	✓	M.S.	Contractions	0.529
BERT uncased	10	✓	✓	M.S.	Lemmatization	0.529
BERT uncased	10	✓	✓	M.S.	Stemming	0.515

***Legends**: R.O.M. = Random Oversampling for Majority class | R.O.A. = Random Over Sampling for Average | A.O.A. = Augmented Oversampling for Average | M.S. = Minority Shuffling*

In order to enhance the efficacy of our model, we conducted a series of experiments involving text cleaning techniques. An analysis of the model revealed that misclassified examples exhibited approximately 10% more contractions, comparable frequencies of stop words, capitalized letters, and abbreviations, as well as longer token lengths.

With this insight, our initial focus was on investigating the impact of contraction cleaning on the model's performance. Surprisingly, this approach resulted in a deterioration of the model's predictive capabilities. Subsequent investigation unveiled that the majority of correctly predicted instances belonged to classes with a higher incidence of contractions.

Drawing inspiration from a prior study (reference [22] that demonstrated improved model accuracy in emotion classification through stemming and lemmatization, we proceeded to experiment with WordNet lemmatization and Porter stemming as preprocessing techniques for our text. Unfortunately, both of these approaches yielded decreased performance.

Table 2 presents a concise summary of the results obtained from our experiments. Ultimately, the most successful model achieved an F1-score of 0.54 on both the validation and test sets. The best model results and weights have been made publicly available on a google drive folder.

In this comparative study, we rigorously evaluate our model against the baseline Large RoBERTa architecture. The results, as displayed in Table 3 demonstrates notable performance improvements across multiple metrics, confirming our approach's efficacy. However, recall shows a slight decrement, highlighting the importance of tailored threshold optimization for enhanced performance.

Table 3: Summary comparison of model performance on the test set

	Large RoBERTa	**BERT w/ M.S.**
Macro F1-Score	0.53	0.54
Micro F1-Score	0.59	0.61
Macro Precision	0.49	0.54
Macro Recall	0.60	0.54
AUC	0.93	0.93

Furthermore, details about the model performance, such as its confusion matrix, classification report, and an assessment of its weaknesses can be found on Appendix 3.

5. Conclusion

In conclusion, our research successfully surpassed the baseline performance, achieving an impressive F1-Score benchmark of 0.54. The application of minority shuffling proved to be pivotal in elevating the model's performance substantially. However, further enhancements are still possible by engaging in hyperparameter tuning, employing more complex and larger models, and fine-tuning the individual models comprising the combined classes, especially given the overfitting patterns drawn from an evaluation of their loss behavior. These avenues for improvement pave the way for even more robust and accurate models, opening up exciting possibilities for advancing the state-of-the-art in this field.

References

[1] G. U. C. M. S. R. G. M. A. P. Chatterjee A, "Understanding Emotions in Text Using Deep Learning and Big Data," *Computers in Human Behavior,* vol. 93, no. April, pp. 309-317, 2019.

[2] B. Gaind, V. Syal and S. Padgalwar, "Emotion Detection and Analysis on Social Media," arXiv, 2019.

[3] P. Ekman, "Basic Emotions," in *Handbook of cognition and emotion*, San Francisco, CA, USA, John Wiley & Sons Ltd., 1999, pp. 45-60.

[4] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," ArXviz, 2020.

[5] S. Bharti, S. Varadhaganapathy, R. Gupta, P. Shukla, M. Bouye, S. Hingaa and A. Mahmoud, "Text-Based Emotion Recognition Using Deep Learning Approach," Comput Intell Neurosci, 2022.

[6] "Papers With Code," [Online]. Available: https://paperswithcode.com/sota/text-classification-on-go-emotions. [Accessed 16 July 2023].

[7] S. Zanwar, D. Wiechmann, Y. Qiao and E. Kerz, "Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features," arxiv, 2022.

[8] V. Suresh and O. C. Desmond, "Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification," arXiv, 2021.

[9] N. Alvarez-Gonzalez, A. Kaltenbrunner and V. Gómez, "Uncovering the Limits of Text-based Emotion Detection," arXiv, 2021.

[10] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou and O. Zaïane, "Seq2Emo: A Sequence to Multi-Label Emotion Classification Model," ACL Anthology, 2021.

[11] V. Suresh and C. D. Ong, "Using Knowledge-Embedded Attention to Augment Pre-trained Language Models for Fine-Grained Emotion Recognition," arXiv, 2021.

[12] G. Singh, D. Brahma, P. Rai and A. Modi, "Fine-Grained Emotion Prediction by Modeling Emotion Definitions," arXiv, 2021.

[13] T. Sosea and C. Caragea, "eMLM: A New Pre-training Objective for Emotion Related Tasks," ACL Anthology, 2021.

[14] J. Olah, S. Baruah, D. Bose and S. Narayanan, "Cross domain emotion recognition usign few shot knowledge transfer," arXiv, 2021.

[15] A. Gera, A. Halfon, E. Shnarch, Y. Perlitz, L. Ein-Dor and N. Slonim, "Zero-Shot Text Classification with Self-Training," arXiv, 2022.

[16] J. Li, Z. Sun, B. Liang, L. Gui and Y. He, "CUE: An Uncertainty Interpretation Framework for Text Classifiers Built on Pre-Trained Language Models," arxiv, 2023.

[17] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Ł. Radliński, K. Wojtasik, S. Woźniak and P. Kazienko, "ChatGPT: Jack of all trades, master of none," arXiv, 2023.

[18] M. Mozhdehi and A. Moghadam, "Textual emotion detection utilizing a transfer learning approach," *The Journal of Supercomputing,* vol. 79, p. 13075–13089, 2023.

[19] D. Q. Nguyen, T. Vu and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," arXiv, 2020.

[20] S. Vora, R. Mehta and S. Patel, "Impact of Balancing Techniques for Imbalanced Class Distribution on Twitter Data for Emotion Analysis: A Case Study," in *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance*, 2021, pp. 211-231.

[21] M. Bayer, M.-A. Kaufhold and C. Reuter, "A Survey on Data Augmentation for Text Classification," arXiv, 2021.

[22] H. Mulki, C. B. Ali, H. Haddad and I. Babaoglu, "Tw-StAR at SemEval-2018 Task 1: Preprocessing Impact on Multi-label Emotion Classification," ACL Anthology, 2018.

Appendix

Appendix 1 – Baseline Model

The table below shows our baseline model performance. It is noted that grief and pride classes are at where our model is performing worst, with no correct grief predictions

Table 4: Baseline model performance by class

	Precision	Recall	F1
admiration	0.63	0.69	0.66
amusement	0.76	0.87	0.81
anger	0.42	0.53	0.46
annoyance	0.31	0.40	0.35
approval	0.37	0.40	0.39
caring	0.43	0.36	0.39
confusion	0.41	0.37	0.38
curiosity	0.46	0.68	0.55
desire	0.55	0.45	0.49
disappointment	0.26	0.34	0.30
disapproval	0.34	0.52	0.41
disgust	0.48	0.45	0.46
embarrassment	0.46	0.30	0.36
excitement	0.41	0.40	0.41
fear	0.49	0.79	0.61
gratitude	0.94	0.89	0.91
grief	0.00	0.00	0.00
joy	0.60	0.55	0.58
love	0.72	0.86	0.79
nervousness	0.23	0.22	0.22
neutral	0.61	0.77	0.68
optimism	0.55	0.55	0.55
pride	1.00	0.06	0.12
realization	0.29	0.14	0.19
relief	1.00	0.09	0.17
remorse	0.56	0.75	0.64
sadness	0.55	0.49	0.52
surprise	0.48	0.50	0.49
macro avg	0.51	0.47	0.46

Appendix 2 – Experiment Approach

In this study, we executed a systematic approach for emotion prediction using machine learning models. The experimentation process began with the creation of a parameterized notebook, enabling us to explore multiple aspects of the model's configuration. We experimented with different transformer architectures, selecting various heads for emotion prediction, tuning learning rates, and implementing diverse balancing techniques to address class imbalances in the dataset. By systematically varying these parameters, we aimed to investigate their impact on the model's performance and identify the most effective combinations.

Upon completing the model training, we logged a comprehensive set of performance metrics to gain insights into the model's effectiveness. We meticulously tracked the model's training history, monitoring its performance during successive epochs. Additionally, we calculated and recorded essential metrics, including the area under the ROC curve (AUC) and F1-score, which served as key evaluation measures for our emotion prediction task. Moreover, we saved the model's predictions on the validation/ test set and recorded instances of misclassification, which provided valuable information for our subsequent model analysis.

To gain deeper insights into the model's behavior, we conducted an in-depth model analysis using a separate notebook. Through this analysis, we examined loss training curves to detect signs of overfitting or underfitting, ensuring the model's appropriate complexity. Furthermore, we evaluated class-specific F1-scores to assess the model's performance for each emotion category, enabling us to identify emotions that might be particularly challenging for the model to predict accurately.

An essential component of our model analysis involved scrutinizing the confusion matrix to understand the model's confusion patterns. This analysis allowed us to identify emotions that were frequently confused with each other, shedding light on the model's strengths and weaknesses. Additionally, we conducted a thorough comparison

of statistics between correctly classified and misclassified examples to detect potential biases or patterns that might influence the model's predictions.

To facilitate comparison across different experiments, we exported the analysis notebook as an HTML file. This enabled us to visually and quantitatively compare the performance of various model configurations side by side, aiding in identifying trends and making informed decisions regarding the model's hyperparameters and architecture.

Throughout the experimentation process, we engaged in an iterative refinement approach, using only the validation results as guidence. Drawing insights from the model analysis, we made informed decisions to modify the model's configurations and parameter settings. This iterative process allowed us to continuously fine-tune the modeling approach, ultimately leading to improved emotion prediction performance.

Appendix 3 – Final Model Results

The optimal performance is achieved through a combination of four pairs of classes: Grief and Sadness, Pride and Admiration, Anger and Annoyance, and Relief and Joy. These specific combinations were derived from the analysis of the best-performing BERT model without any oversampling, as illustrated in Figure 5, presented in Appendix 4. Additionally, an attempt was made to combine "nervousness" and "fear" without implementing any modeling improvements. Moreover, it is evident from the confusion matrix that the combination of "relief" and "joy" does not appear immediately viable. However, insights gleaned from the word cloud analysis of different classes, illustrated in Figure 7 of Appendix 5, reveal certain similarities in the prominent words like "good," "glad," and "thank" between these classes. Hence, the decision to explore this combination was prompted by "relief" being one of the lowest performing classes, and its lack of improvement upon combination with other classes exhibiting higher confusion in the matrix.

Each class combination was subjected to training via a BERT model, incorporating a dropout layer with a rate of 0.3 and a dense output layer with the corresponding number of classes. During the training process, we conducted 10 epochs, and in instances where the categorical loss did not decrease within three epochs, early stopping was implemented. The training procedure utilized a batch size of 16, employing the Adam optimizer with an initial learning rate of 5e-5, and a scheduler to decrease the learning rate by a factor of 5 at the commencement of each epoch.

To create our final predictions, we multiply the probabilities of the combined class with their corresponding individual model to reach a final probability for the individual classes in the original classification.

Table 5 provides a detailed breakdown of the model's performance by class. Notably, substantial enhancements are observed when comparing the model's performance with the initial baseline model, particularly for the minority classes, with a pronounced improvement noted for the class "grief." Additionally, noteworthy improvements are evident for classes such as "nervousness" and "embarrassment." Conversely, certain classes, including "amusement" and "fear," exhibit a decrease in performance. These observations collectively highlight the model's capability to

effectively address specific class imbalances while also revealing potential areas that require further investigation and optimization.

Table 5: Model performance by class

	Precision	Recall	F1
admiration	0.68	0.81	0.74
amusement	0.73	0.83	0.78
anger	0.43	0.63	0.51
annoyance	0.35	0.47	0.40
approval	0.38	0.43	0.40
caring	0.49	0.54	0.51
confusion	0.47	0.51	0.49
curiosity	0.47	0.72	0.57
desire	0.62	0.51	0.56
disappointment	0.36	0.31	0.33
disapproval	0.43	0.43	0.43
disgust	0.47	0.48	0.48
embarrassment	0.68	0.49	0.57
excitement	0.39	0.34	0.37
fear	0.63	0.63	0.63
gratitude	0.89	0.90	0.90
grief	0.50	0.31	0.38
joy	0.59	0.57	0.58
love	0.70	0.85	0.77
nervousness	0.55	0.29	0.37
neutral	0.61	0.76	0.68
optimism	0.59	0.58	0.59
pride	0.90	0.60	0.72
realization	0.39	0.22	0.28
relief	0.23	0.17	0.19
remorse	0.72	0.69	0.71
sadness	0.46	0.63	0.53
surprise	0.53	0.57	0.54
macro avg	0.54	0.54	0.54

A more in-depth examination of the model's misclassification patterns reveals persistent challenges in accurately classifying texts characterized by lengthier content and a higher frequency of emojis, as demonstrated in Table 6. These findings underscore the model's susceptibility to such complex textual structures, warranting future research efforts aimed at enhancing its robustness and addressing the limitations encountered in these specific scenarios.

Table 6: Model misclassification analysis

Metric	✓	×	rate
Char Length	65	74	+14%
Token Length	12	14	+14%
Upper Case rate	27%	27%	-1%
Ponctuations Rate	4.1%	3.9%	-5%
Emoji Rate	0.04%	0.08%	+83%
Stop Words Rate	44%	44%	+1%
Contractions Rate	3%	3%	+5%

Analysis of the model's confusion matrix, on Figure 6 reveals several areas with potential for further improvements. Despite the performance gains achieved through class combinations (Grief and Sadness, Pride and Admiration, Anger and Annoyance), the persistently high level of confusion among these classes indicates that individual models for each of them should be refined to enhance overall performance. Moreover, the model exhibits exacerbated confusion between "nervousness" and "fear," necessitating deeper investigations into the underlying reasons for this challenge and exploring strategies to potentially disentangle these classes. Additionally, the neutral class exhibits substantial confusion with most other classes, warranting the exploration of a dedicated model for distinguishing neutral from non-neutral instances, aiming to effectively capture their inherent differences. These avenues of investigation hold promise for achieving further advancements in the model's accuracy and overall performance.

Appendix 4 – Confusion Matrixes

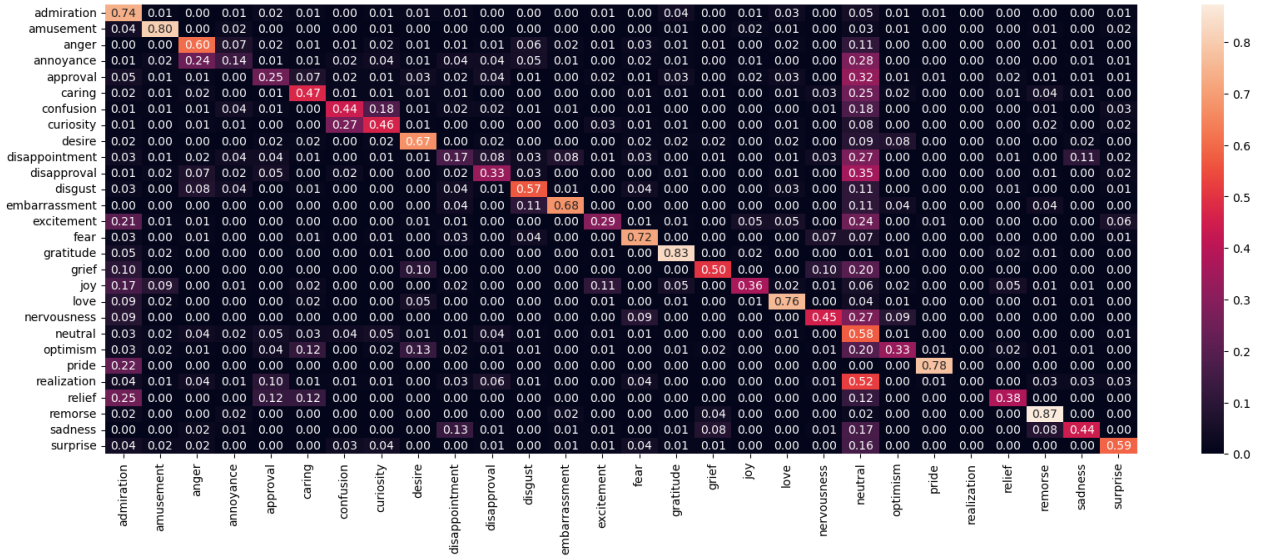

Figure 4: Confusion matrix for average balanced model without augmentation

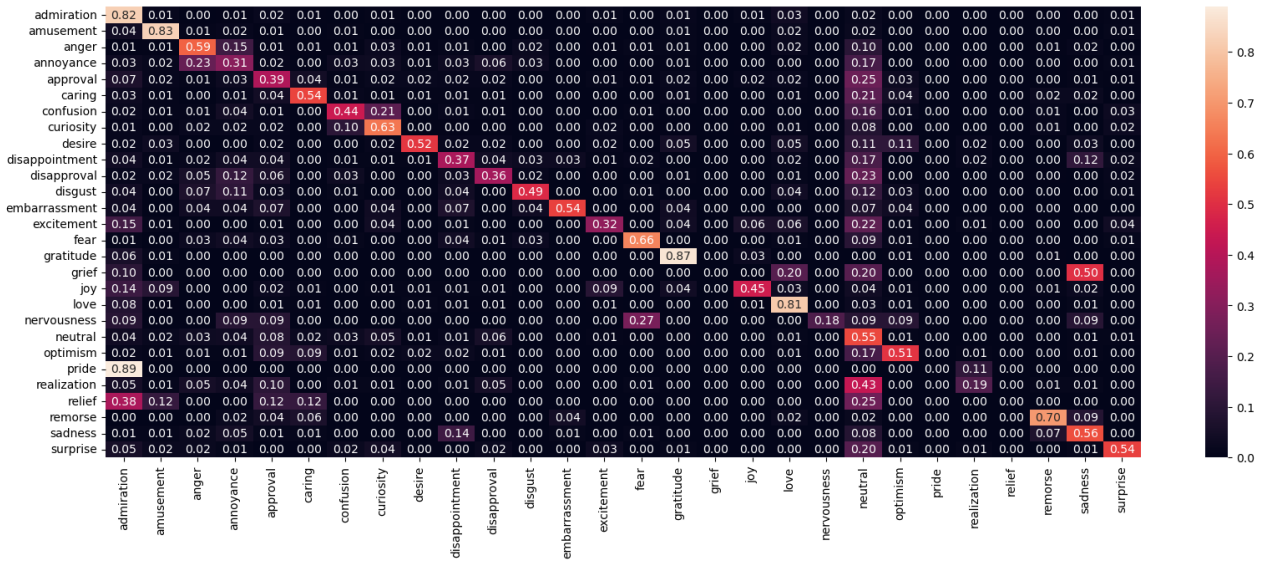

Figure 5: Confusion matrix for average balanced model with augmentation

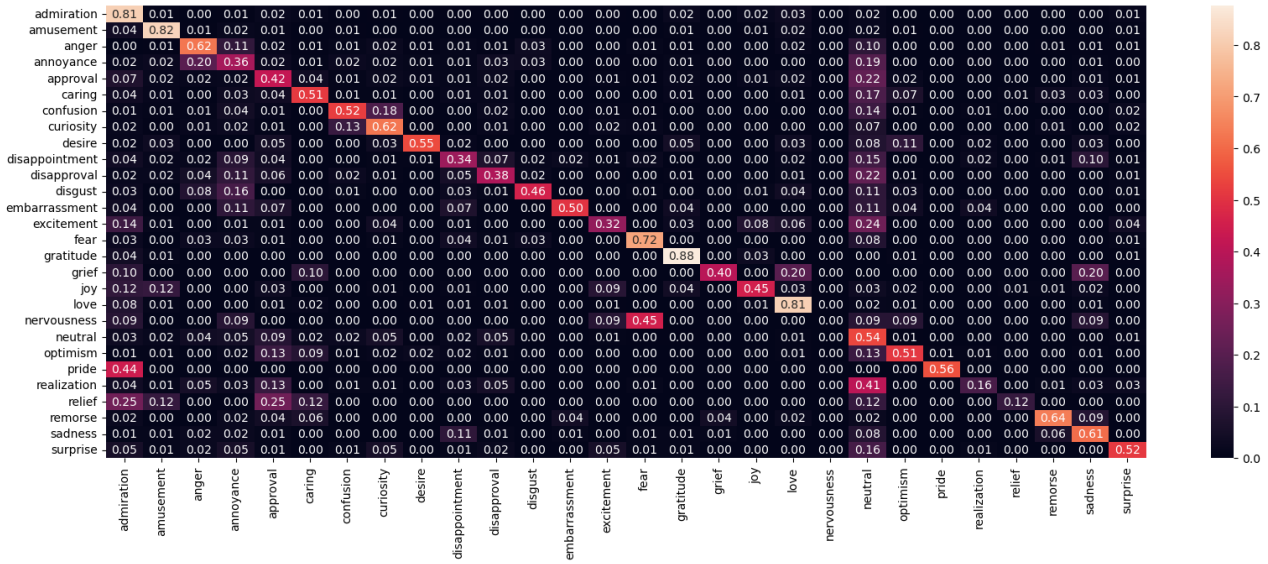

Figure 6: Confusion Matrix for best performing model

Appendix 5 – Word Clouds

Figure 7: Word cloud by emotion