RoBERTa-Based Emotion Classification of Essays: Improving Performance on Imbalanced Data

Anonymous ACL submission

Abstract

This paper presents a study on using the RoBERTa language model for emotion classification of essays as part of the 'Shared Task on Empathy Detection, Emotion Classification and Personality Detection in Interactions' (Barriere et al., 2023), organized as part of 'WASSA 2023' at 'ACL 2023'. Emotion classification is a challenging task in natural language processing, and imbalanced datasets further exacerbate this challenge. In this study, we explore the use of various data balancing techniques in combination with RoBERTa (Liu et al., 2019) to improve the classification performance. We evaluate the performance of our approach (denoted by adityapatkar on Codalab (Pavao et al., 2022)) on a benchmark multi-label dataset of essays annotated with eight emotion categories, provided by the Shared Task organizers. Our results show that the proposed approach achieved the best macro F1 score in the competition's training and evaluation phase. Our study provides insights into the potential of RoBERTa for handling imbalanced data in emotion classification. The results can have implications for the natural language processing tasks related to emotion classification.

1 Introduction

Emotion detection and classification in natural language processing (NLP) is a crucial task with various applications such as sentiment analysis, recommendation systems, and chat-bots. In recent years, deep learning-based models, particularly those based on transformer architectures, have shown remarkable performance in a range of NLP tasks. Among them, the RoBERTa model has gained significant attention for its superior performance on various benchmarks.

However, emotion classification is a particularly challenging task, as emotions are subjective and context-dependent, and often manifest in subtle and nuanced ways. Additionally, imbalanced datasets, where certain emotion categories have fewer instances than others, are common in emotion classification tasks, further complicating the task.

Dataset	Rows
Training	792
Development	208
Test	100
Total	1100

Table 1: Number of rows in each split of the dataset.

We had to work with a highly imbalanced dataset. The task being multi-label added further complexity. Previous works on emotion classification mainly focus on single label classification (Barriere et al., 2022), use somewhat balanced dataset (Demszky et al., 2020) or work with texts shorter than essays (Mohammad, 2012). The last point is crucial as when working with essays, we have to consider the perceived emotions of the complete essay.

Our proposed system contains a RoBERTa-large model. We will discuss various techniques that we tried to overcome the challenges faced by an imbalanced dataset, biggest of which was over-fitting to the majority label. Some approaches include using paraphrasing to increase the size of the dataset, adding class weights as a feature, weight decay etc.

2 Dataset

The dataset provided for the emotion classification task contains essays written in response to news articles where harm to individuals or groups is present. The dataset was divided into training, development, and test sets. Table 1 shows the data split. We focused our analysis on the 'essay' and 'emotion' columns, discarding other columns like gender and age, which were not relevant to the task.

During the exploratory data analysis, we observed that the training and development datasets

Emotion	Occurrences
Sadness	383
Neutral	240
Joy	10
Anger	124
Surprise	19
Disgust	100
Fear	33
Hope	32

Table 2: Number of samples for each emotion in the training dataset.

were comparable in terms of essay length (averaging between 75-80 words per essay) and the split between single-label and multi-label rows. Furthermore, we analyzed the top five most frequent words after removing stop-words in the training dataset, which included ’people,’ ’like,’ ’feel,’ ’think,’ and ’sad.’ This finding suggests that the essays in the dataset often expressed personal opinions and feelings related to human experiences.

To gain further insights, we also examined the top direct objects of the verb in the sentences for each label. For example, the words ’journey’ and ’hardship’ were amongst the top 5 most frequent direct objects for the label ’Joy,’ while for ’Sadness,’ the words ’life’ and ’child’ were present amongst the top 5. This observation highlights the differences in language use across different emotions and provides clues to the underlying emotional experiences. Our exploratory data analysis sheds light on the characteristics of the dataset and provides valuable insights into the language use associated with different emotions.

One of the major challenges encountered in the emotion classification task was the presence of ’data imbalance’. Also, the emotion column in the dataset allowed for a single essay to have multiple emotions, making the classification task more complex. For instance, an essay could express both ’Disgust’ and ’Anger.’

The training dataset exhibited a highly skewed distribution towards ’Sadness’ and ’Neutral’ emotions, as shown in Table 2. The issue of data imbalance is evident from the fact that our baseline model failed to predict the under-represented emotions, such as ’Joy’ and ’Surprise.’

In natural language processing tasks, data imbalance is often addressed through under-sampling or oversampling techniques. We implemented a few techniques to address this problem. We will discuss those in the forthcoming sections.

3 Baseline

To establish a baseline for our multi-label emotion classification task, we opted to fine-tune the BERT model (Devlin et al., 2019) using the HuggingFace Transformers library (Wolf et al., 2020). We detached the head of the model to customize it for our task. For data pre-processing, we retained only the ’essay’ and ’emotion’ columns as they were relevant for our task. We utilized the ’bert-base-uncased’ tokenizer to obtain the final embedding of the essays, which we one-hot encoded for multi-label classification.

We trained the BERT model using TensorFlow, with a learning rate of 2e-05, binary cross-entropy as the loss, and the Adam optimizer. We trained the model for 25 epochs, with a batch size of 16. To prevent over-fitting, we implemented early stopping by monitoring the validation loss. To optimize the threshold on the logits for the labels, we performed a random search. As we trained the model, our training loss kept going down, but after a point, the validation loss did not go down. Instead it increased. This is a sign of overfitting. Figure 1 shows us how the baseline model performed for each individual label.

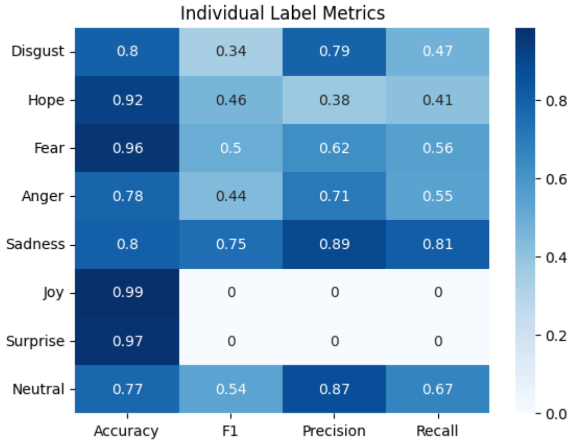

Figure 1: Individual label accuracy, precision, recall and F-1 score of the baseline model on the development dataset.

4 Proposed Approach

We introduce a system[1] that takes into account the limitations put forward by the data imbalance. We

[1] Source code available at https://github.com/adityapatkar/WASSA2023_EMO

use the ’roberta-large’ model with a few tweaks which helped us beat the results of the baseline in every metric.

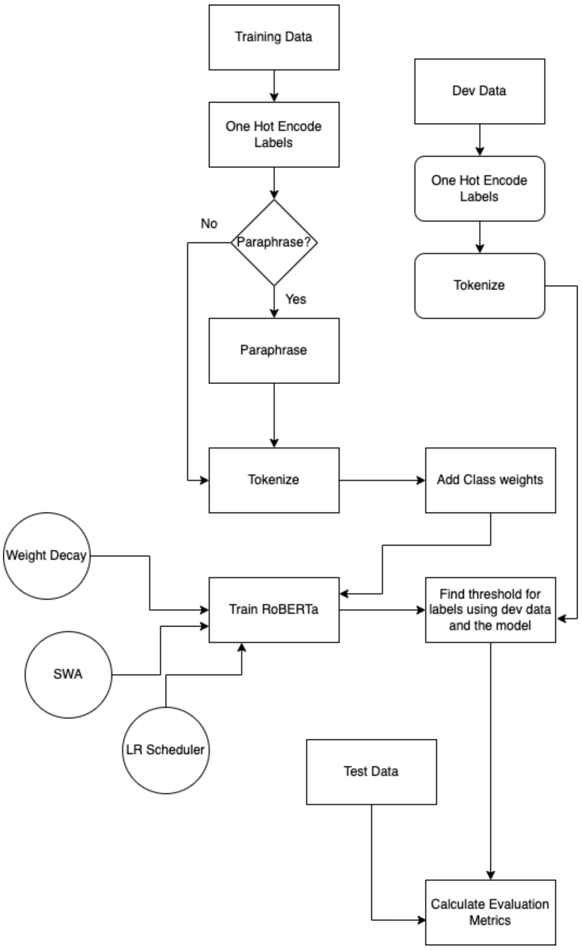

Figure 2: High-level flowchart of our proposed system.

4.1 Our System

Our proposed system deviates from the baseline in several important ways, which we believe have contributed to improved performance on the task at hand.

Firstly, we utilize the RoBERTa model, which has a deeper architecture than the BERT model used in the baseline. This deeper architecture allows for more effective learning of context, which is particularly important for the purpose of this multi-label classification task.

In addition to using a different model architecture, we also implement weight decay to address the issue of over-fitting. This involves setting a defined percentage of weights to exponentially decay to zero, helping to prevent the model from becoming too specialized to the training data. Furthermore, we also implement Stochastic Weight Averaging (SWA) as a means of improving generalization (Izmailov et al., 2019). Our implementation of SWA involves modifying the learning rate schedule and averaging weights after each epoch.

One major innovation we introduced was the use of class weights as a feature in the training dataset. By calculating class weights for each essay based on their inverse frequency, we were able to improve the macro F-1 score by 11%. Specifically, for each class i, we calculated the weight w_i as $w_i = \frac{n}{k \times n_i}$, where n is the total number of samples in the training data, k is the number of classes, and n_i is the number of samples in class i. These weights were then added to the data as a feature.

Finally, we attempted to augment the dataset size by paraphrasing essays associated with under-represented emotions. However, we found that this approach was not effective, as the model started to over-fit to the training data.

Figure 2 provides a high-level flowchart of our system, highlighting the key differences from the baseline. Overall, our changes to the system architecture and training approach have led to improved performance on the task at hand.

4.2 Training

Our system employs the ’roberta-large’ model, which was trained for 25 epochs using a learning rate of 2e-5. To avoid over-fitting, we implemented a weight decay of 0.8%. We also employed the Stochastic Weight Averaging (SWA) technique, which averages the model weights after each epoch, thereby enhancing generalization. We used the Adam optimizer for optimization and binary cross-entropy as the loss function.

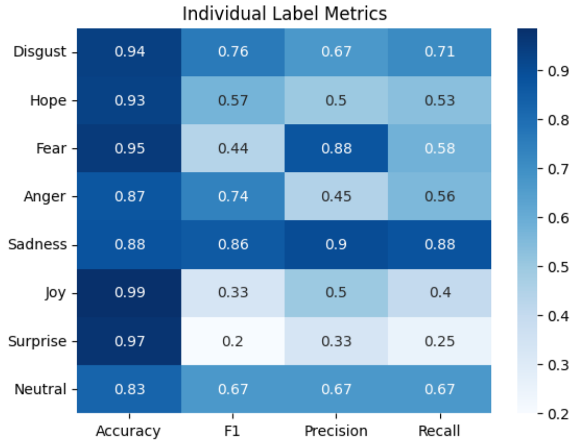

Figure 3: Individual label accuracy, precision, recall and F-1 score of our system on the development dataset.

Development Metrics

Model	Macro F-1	Macro Recall	Macro Precision
`DistilBERT`	0.4021	0.3797	0.5560
`Baseline BERT`	0.4409	0.4458	0.4835
`BERT + Paraphrasing`	0.4207	0.4062	0.4614
`BERT + RoBERTa Ensemble`	0.4251	0.3703	0.5250
`Proposed System with RoBERTa`	**0.5790**	**0.6251**	**0.5711**

Test Metrics

Model	Macro F-1	Macro Recall	Macro Precision
`Baseline BERT`	0.5464	**0.7257**	0.5039
`Proposed System with RoBERTa`	**0.7012**	0.6773	**0.8105**

Table 3: Evaluation metrics on the development and test set.

4.3 Evaluation

Similar to the baseline model, we performed a random search on the logits to determine the best threshold. Our search yielded a value of -0.075 as it maximized the F-1 score on the development set. Figure 3 displays the performance of our proposed system for each individual label. It is evident that there is an improvement in performance compared to the baseline. Notably, emotions with a low number of samples are now being predicted, which the baseline failed to predict. Table 3 compares the results of the baseline and the proposed system on the development set and the test set. Our system shows a considerable improvement in all metrics except the recall over the baseline.

5 Alternate Approaches

We employed alternative approaches in our study, in addition to the baseline model. The first approach involved utilizing a DistilBERT model (Sanh et al., 2020), for a strictly single-label classification. The model was trained for 100 epochs with a learning rate of 1e-5. However, the results were unsatisfactory.

In the second approach, we utilized BERT as the underlying model, similar to the baseline approach, but expanded the dataset by paraphrasing essays related to labels with less than 40 samples. However, this approach did not perform better than the baseline, due to overfitting of the training data.

Finally, we had observed that ensemble models had a good performance on emotion classification tasks (Maheshwari and Varma (2022), Ganaie et al. (2022)). We used the PyTorch implementation of RoBERTa and BERT and fine-tuned both models using the binary cross-entropy loss and optimized them using the Adam optimizer with a learning rate of 1e-05. We used the RoBERTa tokenizer for RoBERTa and the BERT tokenizer for BERT to tokenize the input text. After adding the sigmoid head for 8 classes, we trained both models for 40 epochs, with early stopping based on validation loss criteria. For the final prediction, we took the average output probability of the two models and used a threshold of 0.066 to predict the labels.

Details of the performance of these approaches on the development set can be found in Table 3.

6 Conclusion

In conclusion, this paper presents a study on using the RoBERTa language model for emotion classification of essays, focusing on addressing the challenges posed by imbalanced datasets. The proposed approach combines various data balancing techniques with RoBERTa to improve classification performance, and the results show that the proposed approach achieved the best macro F1 score in the competition's training and evaluation phase. The study provides valuable insights into the potential of RoBERTa for handling imbalanced data in emotion classification, which can have implications for natural language processing tasks related to emotion classification. Overall, the proposed approach offers a promising direction for future research in this field.

References

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Work-*

shop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 214–227.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Salvatore Giorgi. 2023. Wassa 2023 shared task: Predicting empathy, emotion and personality in interactions and reaction to news stories. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. Averaging weights leads to wider optima and better generalization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Himanshu Maheshwari and Vasudeva Varma. 2022. An ensemble approach to detect emotions at an essay level. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 276–279, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.