COMP9444 - Emotion Classification using Tweets

(Joe) Jiaheng Kang z5419190, (Dylan) Yuyang Xiao z5529467, Silu Yang z5526686, Tong Zhou z5418247, Xiao Dong z5571806

***Abstract* — Emotion classification in social media (or any informal context) remains challenging due to linguistic noise. The work proposes an enhanced framework developed based on other research papers. The model would leverage preprocessed emotion data from CARER, then combine class-imbalance mitigation techniques with fine-tuned DeBERTa to resolve the emotion multiclass classification problem.**

Keywords — Emotion Recognition, Class imbalance, DeBERTa, Twitter data

I. INTRODUCTION

Emotion recognition holds great promise across domains such as mental health monitoring and human-computer interaction. However, real-world deployment scenarios would inevitably face the three obstacles.

- Linguistic complexity: Informal expressions like "ughhh why meeeee?" requires nuanced interpretations.
- Complicated contextual information: Sarcasm and verbal irony are hard to identify. Saying "What a beautiful day!" when it's raining heavily could lead to misunderstanding by the model.
- Data sparsity / class imbalance: When one category makes up most of the data, it can overshadow the smaller categories. As such, accuracy alone doesn't give a fair picture of performance, so other measures like the F1-score were also considered.

While other studies have shown that CARER handles linguistic complexity and ReBERTa achieves strong baseline performance, both neglect class imbalance. Thus, our proposed solution would be to bridge this gap through three enhancements: first, leveraging the preprocessed data described in the CARER study; second, applying techniques to mitigate imbalance; and finally, fine-tuning the model with a DBeERTa framework.

II. RELATED WORK

Saravia et al. introduced CARER, a semi-supervised graph-based framework that enriches contextual patterns with word embeddings; This design captures non-standard forms such as "waaaaking" (as discussed earlier) and delivers state-of-the-art F1 scores across eight emotions. CARER, however, is trained on naturally skewed label distributions; the authors acknowledge the need to use F1 rather than accuracy precisely because the dataset is imbalanced, yet they leave unexplored the mitigation of this imbalance.

To streamline reproducibility, the same team later released a Colab notebook that fine-tunes RoBERTa on a six-emotion variant of the data set. Such a dataset is preprocessed with the first several steps described in the paper. Thus, it yields a strong deployable baseline. Despite its empirical gains and linguistic ReBERTa inherits CARER's limitation: majority classes dominate training, so minority emotions remain under-represented.

The FEEL-IT model (Basile et al., 2021), developed for the Italian language, was trained with high-quality Italian tweets sourced from the COMMOWNCRAW ITA corpus, achieving high performance results with the UmBERTo-FT model for emotion classification. Thus, we adapted this approach for English tweets; the dair-ai/emotion dataset was reformatted by mapping *love* and *surprise* into the *joy* category, resulting in a four-class schema (*anger*, *fear*, *joy*, *sadness*). A model was also retrained using the pre-trained FEEL-IT weights and the English dataset. But, due to the fact that original emotion annotations were produced by Italian speakers, the label definitions and examples inherently reflect Italian cultural and linguistic norms. Even after we fine-tuned the model with English data, these cultural and syntactic biases persisted, limiting the model's performance in F1 score and accuracy as seen in Figure 1 and 2 below.

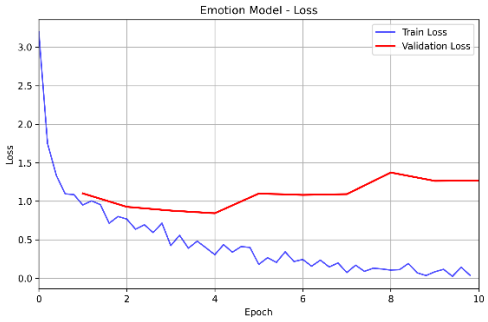

Figure 1: FEEL-IT Loss across Epochs

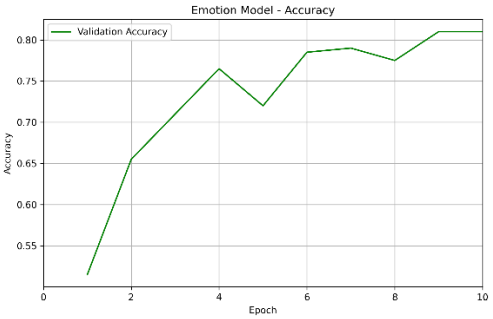

Figure 2 FEEL-IT Accuracy across Epochs

Rumali et al. (2021) compared CNN and BiLSTM models for multi-class emotion classification on Twitter data, establishing both as strong baselines as seen in Figure 3-6 below.

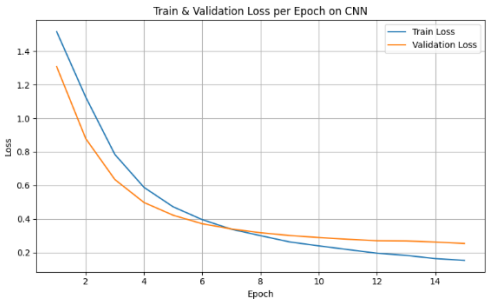

Figure 3: CNN Loss across Epochs

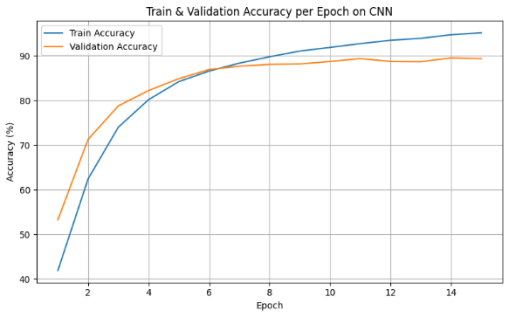

Figure 4: CNN Accuracy across Epochs

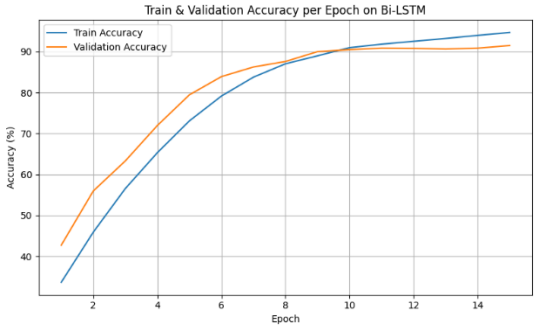

Figure 5: Bi-LSTM Accuracy across Epochs

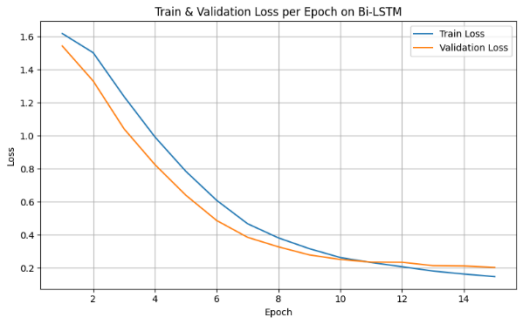

Figure 6: Bi-LSTM Loss across Epochs

CNNs, especially when paired with pre-trained static embeddings such as GloVe, are computationally efficient and effective at capturing local patterns (Kim, 2014), but struggle with long-range dependencies and contextual nuance. BiLSTMs improve global context modelling and consistently outperform CNNs in accuracy and F1 score, yet incur higher computational cost (Graves & Schmidhuber, 2005) and share the same limitations of static embeddings and lack of imbalance handling. Attempts to improve these baselines via deeper architectures or attention pooling add complexity with marginal gains. These constraints motivate our focus on transformer-based models (RoBERTa, DeBERTa), which deliver stronger contextual understanding and allow more flexible data imbalance mitigation through sampling and fine-tuning.

III. METHODS

This gap motivates our three-tiered approach: (1) reuse CARER's cleaned data pipeline / preprocessed data, (2) utilise a DeBERTa model to do the final prediction, (3) introduce under sampling and oversampling (repeated sampling).

We selected Microsoft/DeBERTa-v3-small over RoBERTa because of its capabilities to understand and analyse both content and relative position, which enhanced the detection of subtle cues like sarcasm, emojis, and indirect emotions. In tweets, these cues are often spread across the sentence in non-standard ways — for example, sarcasm may rely on a positive phrase followed much later by a contradicting phrase ("Oh great… another Monday 🙃"). Unlike absolute positional embeddings in RoBERTa, which tie meaning to fixed token positions, DeBERTa's relative positional encoding models the distance between tokens directly, allowing it to link related words even when separated by multiple tokens. This is particularly effective for informal, irregular tweets containing fragmented grammar, unusual word order, and mixed modalities such as hashtags, emojis, and abbreviations, enabling the model to better capture context and subtle emotional contrasts

Benchmarks such as the GoEmotions dataset also demonstrated the DeBERTa model achieving a higher F1 score and faster convergence than RoBERTa.

The DeBERTa-v3-small model architecture can be found in Figure 7 below. The DeBERTa model consists of 6 transformer encoder layers, each with 768 hidden units and 12 attention heads, using GELU activation and a 128k vocabulary.

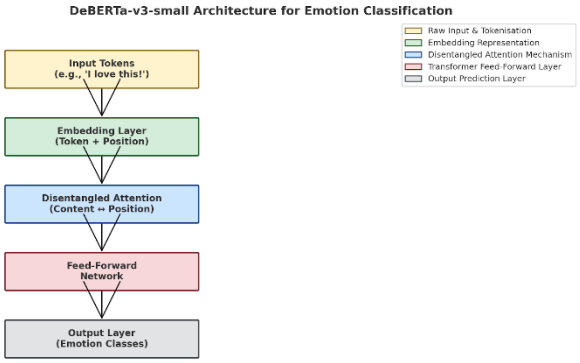

Figure 7: DeBERTa-v3-small model architecture

For the ultimate classification stage, we used this pre-trained model and fine-tuned it on our emotion dataset. Fine-tuning allowed the model to adapt its general language understanding to the specific characteristics of Twitter data — short, noisy, and often context-dependent — while retaining the semantic and syntactic knowledge gained during large-scale pre-training. This transfer learning approach is particularly effective for emotion detection, as it enables the model to recognise subtle affective cues without requiring vast amounts of task-specific data. Tokenization was handled by the DeBERTa-v3-small's native tokenizer, which is optimised to work with its sub word vocabulary and relative positional encoding. Using the native tokenizer ensured consistent treatment of domain-specific patterns such as hashtags (mondayblues), emojis (😂, 😡), and multi-character abbreviations (LMAO, LOL, smh) that carry strong emotional or tonal meaning. Preserving these elements intact — rather than splitting them into unrelated sub-tokens — helped the model maintain semantic coherence, allowing it to

better leverage these high-signal tokens when inferring sentiment and emotion.

We explored oversampling methods such as SMOTE to address class imbalance but found them less suited for the natural language processing task. In text classification tasks, synthesised samples can often result in non-continuous token sequences or simply invalid sentences, which can confuse the model. To mitigate these issues, we used repeated random oversampling of real examples instead of traditional synthetic generation, combined with under sampling of over-represented classes to achieve a more balanced training distribution.
For example:
Suppose that we have 2 original valid sentences:
"I hate you!" (anger)
"I want to kill!" (anger)
Possible invalid outcomes:
Want I you to! (Nonsense, invalid sentence)
I hate to kill! (Loses the anger context, ambiguous tone)

IV. EXPERIMENTS

We used the emotion dataset introduced by Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen∗ from National Tsing Hua University, Hsinchu, Taiwan, as described in their work CARER: Contextualised Affect Representations for Emotion Recognition.

Data Source URL: https://huggingface.co/datasets/dair-ai/emotion

In this project, we reused their preprocessed data. The dataset contains 6 basic emotions (sadness, joy, love, anger, fear and surprise) with approximately 16k training samples, 2k validation samples and 2k test samples.

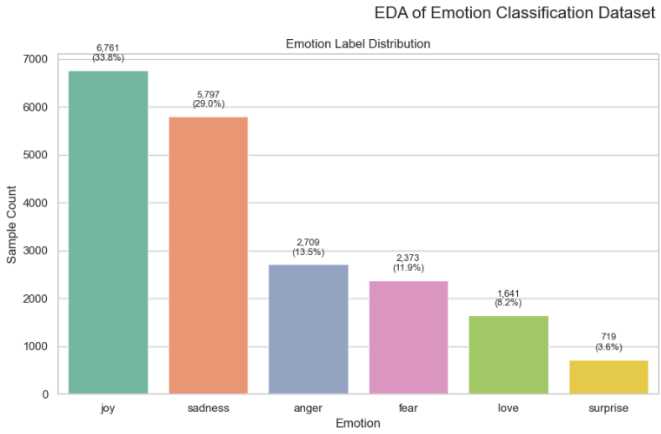

Figure 8: Dataset Class Distribution

Figure 8 shows severe class imbalance with a max/min ratio of 9.40x, where joy (33.8%) and sadness (29%) dominate, while surprise accounts for only 3.6% of the data. We also noticed that Tweet lengths can vary widely, running from 7 to 300 characters. No significant data leakage was detected. Text length alone yields only 33.8% accuracy in predicting labels, confirming that content rather than length is decisive for classification.

We also added an attention pooling layer before the classifier to assign greater weight to emotionally salient tokens, allowing the model to focus more on the words, emojis, or symbols most indicative of sentiment while reducing the influence of irrelevant or neutral content. This approach is particularly valuable in tweets, where emotionally charged words may be sparse but highly informative.

Key training settings include:
Max sequence length: 512 tokens, to ensure complete coverage of longer tweets or threads while maintaining computational feasibility.
Batch size: 32 to balance gradient stability with available GPU memory.
Learning rate: 2e-5, optimised for fine-tuning large language models, paired with the AdamW optimiser for better weight regularization.
Dropout: 0.3 to mitigate overfitting by introducing controlled noise during training.
Loss function: Cross-entropy loss which is commonly used for multi class classification problems
Early stopping based validation F1 score to avoid overfitting and prioritise balanced performance across classes rather than raw accuracy.

To address class imbalance, we implemented a two-way resampling strategy. A constant max_samples_per_class threshold was set: any class exceeding this limit was undersampled to reduce its dominance, while any class below this limit was oversampled to increase its representation. Unlike purely synthetic methods such as SMOTE, we relied on repeated sampling of authentic examples to preserve natural language coherence. The value of max_samples_per_class was empirically tuned, with multiple trials conducted to determine the setting that yielded the highest overall F1 score. This ensured that minority emotions received adequate representation without excessively discarding data from majority classes, maintaining both dataset balance and linguistic diversity.

V. RESULTS

Our proposed model achieved a test accuracy of 93.37% and a macro-average F1 score of 0.9260 on the dataset. Per-class evaluation also demonstrated consistently strong performance, with F1 scores exceeding 0.85 for all categories as seen in Table 1 below. Notably, performance gains are most significant for the two minority classes (love and surprise) compared to the prior works.

	precision	**recall**	**f1-score**	**support**
sadness	0.961872	0.955250	0.958549	581
joy	0.958580	0.932374	0.945295	695
love	0.806818	0.893082	0.847761	159
anger	0.936567	0.912727	0.924494	275
fear	0.886364	0.870536	0.878378	224
surprise	0.674699	0.848485	0.751678	66

accuracy	0.923500	2000		
macro avg	0.870817	0.902076	0.884359	2000
weighted avg	0.926988	0.923500	0.924647	2000

Table 1 – Improved Model Performance Across Classes

When compared with the RoBERTa baseline provided by the Saravia team (accuracy 92.35% macro F1 0.884), our approach improves F1 by about 4.2% as a result of the result of DeBERTa's stronger contextual modelling and our integration with oversampling/undersampling methods to counteract label imbalance.

Compared to the baseline, DeBERTa-v3-small's higher capacity (143M vs. 82M parameters) and significantly faster training speed (6× faster per epoch) allowed for both improved efficiency and better generalisation. (As seen in Figure 9 vs Figure 10 below)

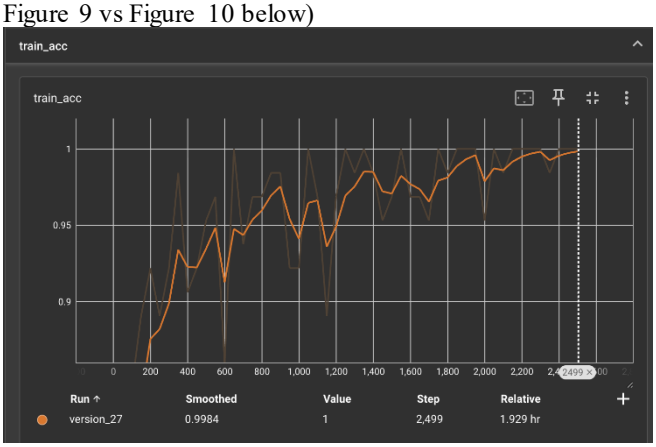

Figure 9: Baseline Model Training Accuracy across Epoch

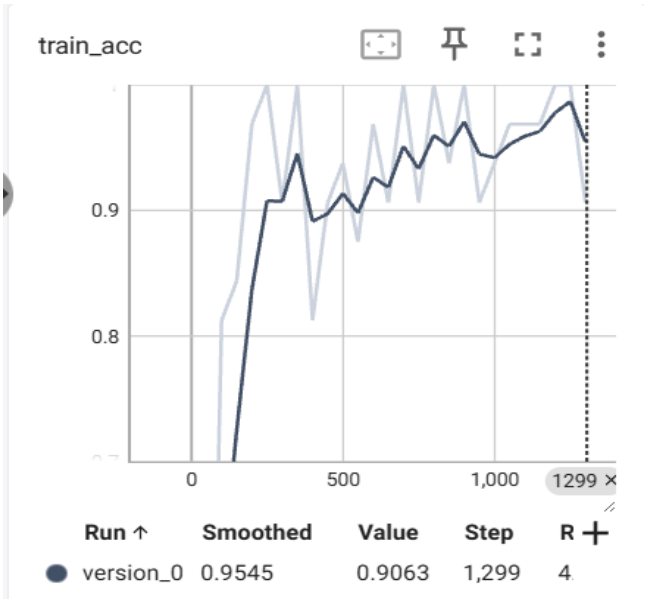

Figure 10: Improved Model Training Accuracy across Epoch

This efficiency is likely due to factors such as DeBERTa's improved attention mechanism reducing computational redundancy, the relative positional encoding requires fewer parameters than absolute embeddings for capturing positional relationships, and the model's enhanced representational capacity may enable faster convergence to optimal solutions and Microsoft's optimisation towards pytorch.

Additionally, early stopping and faster convergence further reduced the risk of overfitting.

However, some limitations remain. While oversampling helped address class imbalance, it may have inadvertently introduced redundancy in the training set, leading to slight overfitting on certain minority classes. This could result in the model relying heavily on repetitive patterns rather than generalisable features.

Furthermore, the classifier continued to struggle with differentiating between semantically similar emotional states—particularly pairs such as Joy vs. Love and Fear vs. Surprise—where the linguistic and contextual cues often overlap. For instance, affectionate language can easily be interpreted as both joy and love, while sudden or intense descriptions may convey elements of both fear and surprise. Such confusion is exacerbated in short, informal texts, where contextual markers are sparse and emotion boundaries are inherently fuzzy.

Future research could involve scaling to larger and more context-sensitive transformer architectures (e.g., DeBERTa-large or GPT-based encoders), which may capture subtler semantic distinctions through deeper contextual embeddings.

Another area of research is emotion-specific pretraining, where models are exposed to large, emotion-rich corpora to better learn affective representations, or contrastive learning strategies that explicitly push apart similar but distinct emotion categories in embedding space.

Additionally, integrating external knowledge sources—such as commonsense reasoning datasets or psychological emotion taxonomies—could provide the model with richer grounding to disambiguate nuanced expressions. Finally, shifting from a single-label framework to a multi-label classification setup may better reflect the multifaceted nature of emotional expression in real-world language, where multiple emotions frequently co-occur within the same utterance.

VI. CONCLUSION

We combined DeBERTa with attention pooling and balanced sampling, achieving higher macro F1 over the CARER RoBERTa baseline. The approach avoids synthetic oversampling, maintains linguistic integrity, and is more efficient to train. But we still identified some limitations: The model still struggles to distinguish joy vs. love and fear vs. surprise whenever the linguistic cues overlap. Thus, for future improvements, we plan to explore larger DeBERTa variants and multi-label classification and try to capture co-existing emotions more effectively.

REFERENCES

[1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Emotion recognition and its applications," in Advances in intelligent systems and computing, 2014, pp. 51–62. doi: 10.1007/978-3-319-08491-6_5.

[2] S. Kiritchenko, X. Zhu, and S. M. Mohammad, “Sentiment analysis of short informal texts,” Journal of Artificial Intelligence Research, vol. 50, pp. 723–762, Aug. 2014, doi: 10.1613/jair.4272.

[3] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, “A multimodal corpus for emotion recognition in sarcasm,” ACL Anthology, Jun. 01, 2022. https://aclanthology.org/2022.lrec-1.756/

[4] T. Meng, Y. Shou, W. Ai, N. Yin, and K. Li, “Deep imbalanced learning for multimodal emotion recognition in conversations,” IEEE Transactions on Artificial Intelligence, vol. 5, no. 12, pp. 6472–6487, Aug. 2024, doi: 10.1109/tai.2024.3445325.

[5] A. Papenmeier, D. Kern, D. Hienert, Y. Kammerer, and C. Seifert, “How accurate does it feel? – Human perception of different types of classification mistakes,” CHI Conference on Human Factors in Computing Systems, Apr. 2022, doi: 10.1145/3491102.3501915.

[6] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, “CARER: Contextualized Affect Representations for Emotion recognition,” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Jan. 2018, doi: 10.18653/v1/d18-1404.

[7] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, “Multi-label emotion classification in texts using transfer learning,” Expert Systems With Applications, vol. 213, p. 118534, Sep. 2022, doi: 10.1016/j.eswa.2022.118534.

[8] F. Bianchi, D. Nozza, and D. Hovy, “FEEL-IT: Emotion and Sentiment classification for the Italian language,” ACL Anthology, Apr. 01, 2021. https://aclanthology.org/2021.wassa-1.8/

[9] Y. Takenaka, “Performance Evaluation of Emotion Classification in Japanese Using RoBERTa and DeBERTa,” arXiv preprint arXiv:2505.00013, Apr. 2025, doi: 10.48550/arXiv.2505.00013.

[10] N. Rumali, R. E. R. Hridoy, M. H. Sami, and S. N. Jyoti, “Emotion Extraction and Classification from Twitter Text,” GitHub repository, Available: https://github.com/rejonehridoy/Emotion_Extraction_and_Classification_from_Twitter_Text_using_Deep_Learning. [Accessed: Oct. 13, 2021].

[11] Y. Kim, Convolutional Neural Networks for Sentence Classification. Doha, Qatar: Association for Computational Linguistics, 2014,pp. 1746–1751. doi: 10.3115/v1/d14-1181.

[12] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” Neural Networks, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/j.neunet.2005.06.042.