SemEval: Emotion Detection

Hamza Manzoor
Matriculation number: 5201765
Module: Models of Computation
hamza.manzoor@mailbox.tu-dresden.de

Abstract

This project delves into text-based emotion detection, emphasizing perceived emotions—the feelings that a general audience would attribute to a speaker based on textual input. The primary objective was to identify the emotion most commonly inferred from a given text snippet. To optimize model performance, a range of techniques was employed, including transformer-based models, hyperparameter tuning, ensemble learning, and data augmentation methods such as backtranslation and the combination of backtranslation with paraphrasing. Notably, the study highlights the paramount importance of data quality in achieving accurate results, affirming the adage "garbage in, garbage out." By enhancing the dataset through strategic augmentation, the final approach achieved a remarkable F1 score of 0.95. All the relevant files are available at GitHub.

1 Introduction

Emotions are integral to human communication and cognition, shaping thoughts and interactions (Mohammad et al., 2018). While emotions are universally familiar, they are inherently complex and nuanced. Certain emotions—such as joy, sadness, and fear—are considered fundamental according to the categorical model of emotions (Seyeditabari et al., 2018), as they are deeply rooted in physiological and cognitive mechanisms. Text serves as a valuable medium for emotion detection due to the growing availability of textual data from sources such as microblogs, emails, and SMS messages. The rapid expansion of emotion-rich textual data necessitates automated methods for identifying and analyzing emotions expressed in text (Automatic emotion detection in text streams by analyzing Twitter data). Social media platforms and microblogging tools (e.g., Twitter, Facebook) have become major channels for individuals to share their emotions and opinions on daily life, current events, and social issues. These messages often contain explicit and implicit indicators of emotions, making them a useful resource for emotion detection research (Hasan et al., 2019). Emotion detection in computational linguistics involves identifying discrete emotions conveyed through text. It extends beyond traditional sentiment analysis by offering a more detailed understanding of sentiment (Seyeditabari et al., 2018). Moving past binary sentiment classification (positive vs. negative) allows for more insightful applications, such as targeted marketing strategies and socio-political sentiment monitoring. For example, while both fear and anger convey negative sentiment, but latter is particularly relevant in political discourse and marketing campaigns (Seyeditabari et al., 2018). Emotion recognition in text presents unique challenges due to the subjective nature of emotions and linguistic subtleties (Wilson et al., 2005; Kiritchenko and Mohammad, 2018; Mohammad, 2022, 2023). This project addresses perceived emotion detection, aiming to predict the emotions that a general audience would infer from a given text snippet. To achieve this, various deep learning approaches—especially transformer-based models—were explored, alongside data augmentation techniques in an attempt to maximize F1-score.

1.1 Data and Scope

This task involves predicting the perceived emotion(s) of a speaker based on a given text snippet. Specifically, the model determines whether any of the following emotions apply: joy, sadness, fear, anger, and surprise. The multilabel dataset used in this project is sourced from SemEval 2025 Task 11: Bridging the Gap in Text-Based Emotion Detection (Track A) (Abdulmumin, 2025) and consists of English language text snippets annotated for five primary emotion labels: joy, sadness, fear, anger, and surprise. This task centers on perceived emotion detection, aiming to determine the emotion that most

people would attribute to the speaker based on a given sentence or short text snippet (Abdulmumin, 2025). What This Task Does Not Cover:

- The emotion experienced by the reader upon interpreting the text.
- The emotion of another individual mentioned within the text.
- The speaker's actual emotion, as it cannot be definitively inferred from a short text snippet alone.

It is crucial to recognize this distinction, as perceived emotions may differ significantly from actual emotions. This discrepancy arises due to several factors, including cultural context, individual differences in emotional expression, and the inherent limitations of text-based communication (Woensel and Nevil, 2019; Wakefield, 2021). A significant challenge in this dataset is class imbalance, where certain emotions are significantly underrepresented compared to others. Through extensive experimentation, it became evident that data imbalance acted as a bottleneck in maximizing the F1 score. Addressing this limitation was a central focus of this project, leading to the application of various data augmentation strategies. It is important to emphasize that the scope of this study was strictly limited to the original dataset provided by SemEval 2025 (Abdulmumin, 2025). No external data sources were incorporated at any stage of the research. All data enhancement techniques, including augmentation and rebalancing, were exclusively applied to the provided dataset, ensuring that the scope and integrity of the task remained consistent with the original data constraints.

2 Related Work

Text-based emotion detection has garnered significant attention in recent times, leading to the development of various approaches aimed at accurately identifying emotions from textual data. Early methods primarily relied on lexicon-based techniques, utilizing predefined dictionaries of emotion-related words to assess the emotional content of text. While straightforward, these approaches often struggled with context sensitivity and the subtleties of language, as they typically consider single words in isolation and may not account for qualifiers or negations preceding a word (Wakefield, 2021).

The advent of machine learning introduced supervised learning methods, where models were trained on labeled datasets to recognize emotional expressions. These models demonstrated improved performance over lexicon-based approaches by learning patterns directly from data. Common algorithms employed include k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Artificial Neural Networks (ANNs). For instance, a study explored the application of k-NN, SVM, and ANNs for emotion detection in text, highlighting their potential in improving human-machine interaction (Machová et al., 2023).

In recent years, deep learning techniques, particularly transformer-based models like BERT and RoBERTa, have achieved state-of-the-art results in emotion detection tasks. For instance, a study proposed a hybrid deep learning network combining convolutional and recurrent layers to detect emotions in text, highlighting the effectiveness of such architectures in capturing complex linguistic features (Kusal et al., 2022).

The effectiveness of machine learning models, including neural networks, is intrinsically linked to the quality and quantity of their training data. This relationship is succinctly captured by the adage "Garbage In, Garbage Out," which underscores that poor-quality input data inevitably leads to suboptimal outputs. In the context of artificial intelligence, data serves as the foundational element, the lifeblood of AI (Kuismin-Raerinne and Nieminen, 2022), upon which models are built. An article from Open Data Science (Ron Ozminkowski, 2021) elaborates on this concept, stating that if the data fed into a model is erroneous or irrelevant, the resulting predictions or analyses will be compromised.

Despite these advancements, challenges persist, notably the generalizability of models across different domains and languages. Research has indicated that models trained on specific datasets may not perform well when applied to data from different sources, emphasizing the need for approaches that enhance cross-domain robustness (Zanwar et al., 2022).

Furthermore, the issue of class imbalance in emotion datasets remains a significant hurdle. Certain emotions are often underrepresented, leading to models that are biased towards more prevalent emotions. Addressing this imbalance through data augmentation techniques, such as backtranslation and paraphrasing, has been explored to create more

balanced datasets and improve model performance.

For example, a study by Koufakou et al. (Koufakou et al., 2023) investigated the impact of data augmentation on small, imbalanced emotion datasets. They applied techniques including Easy Data Augmentation (EDA), embedding-based methods, and ProtAugment, finding that these methods significantly improved classifier performance. Additionally, case studies utilizing ChatGPT for paraphrasing demonstrated promising potential in enhancing emotion detection models.

Similarly, In their 2020 study, "Semi-Supervised Models via Data Augmentation for Classifying Interactive Affective Responses," Chen et al. (Chen et al., 2020) introduced a semi-supervised text classification system that employs data augmentation techniques, including backtranslation, to enhance the classification of interactive affective responses. Their approach demonstrated notable improvements in F1-score and accuracy.

Overall, the field has seen a progression from simple lexicon-based methods to sophisticated deep learning models, with ongoing research focusing on improving generalizability, handling class imbalance, and developing multilingual capabilities.

3 Methodology

This study focuses on predicting perceived emotions—specifically joy, sadness, fear, anger, and surprise—from English text snippets. The methodology involves data preparation, model development, and evaluation, structured as follows.

3.1 Data Preparation

The dataset went through basic preprocessing and preparation for model training. Initially, the data was examined for any missing values, and none were found. Subsequently, the dataset was divided into training and testing sets using an 80/20 split, reserving a sufficient portion for evaluation. The text snippets were then tokenized using BERT's uncased tokenizer, converting the text into numerical representations suitable for model input.

3.2 Baseline Model Development

The initial approach involved training a BERT-base-uncased model, chosen for its proven effectiveness in similar tasks (as discussed in the Related Work section). The problem was defined as a multi-label classification task, enabling the model to assign multiple emotions to a single text snippet. Training commenced with the default BERT parameters, using the F1-score as the evaluation metric. The baseline model achieved moderate performance, highlighting room for optimization.

3.3 Hyperparameter Optimization

Given the limitations of the baseline model, several hyperparameter optimizations were performed. The number of epochs was increased to 25, allowing the model to converge fully and ensuring that performance plateaued due to model saturation rather than insufficient training. The 'load_best_model_at_end' parameter was set to True to avoid retaining an overfitted model. The learning rate was set to 2e-5, a standard value for fine-tuning BERT, balancing convergence speed and stability. Weight decay was increased to 0.1 to reduce overfitting by penalizing large weight values. Batch size was increased to 32, ensuring a more stable gradient estimate during training. Warmup steps were set to 500, allowing for a gradual increase in learning rate, stabilizing the early stages of training. Despite these optimizations, the improvement in F1-score was minimal.

3.4 Custom BERT Model Development

Given the limited impact of hyperparameter tuning, a customized BERT model was developed, leveraging transfer learning to refine the model's adaptability. To retain foundational language representations while allowing task-specific adaptation, the first encoder layers were frozen, with only the last six layers unfrozen for fine-tuning (bui, 2024).Additional modifications included the introduction of dropout layers with a 0.5 probability after fully connected layers and activation functions to mitigate overfitting (Yadav, 2022), the addition of a fully connected layer to reduce the dimensionality of BERT outputs to 512 neurons, application of layer normalization to stabilize training and accelerate convergence, utilization of the GELU activation function to enhance non-linearity and complex pattern learning, and an output layer to map the processed features to emotion labels. raining adjustments involved increasing warmup steps to 1000 for a smoother optimization process and raising weight decay to 0.2 to strengthen regularization. Despite these modifications, the F1-score remained stagnant, indicating that additional architectural improvements or data-centric approaches were necessary. Despite these modifications, the F1-score

remained stagnant, indicating that additional architectural improvements or data-centric approaches were necessary.

3.5 Ensemble Models with Fine-Tuning

To leverage the strengths of multiple transformer architectures, an ensemble approach was explored, combining BERT-base-uncased, RoBERTa-base, and DeBERTa-v3-base (Jiang, 2024).The Ensemble model was first evaluated without fine-tuning to establish reference scores. Following this, each model underwent individual fine-tuning using a learning rate of 2e-5, a weight decay of 0.01, a batch size of 16, and 500 warmup steps over 10 epochs. Mixed precision training was enabled for computational efficiency, and the best-performing models were saved after each epoch. After fine-tuning, each model was assigned a weight based on its validation F1-score, with BERT-base-uncased receiving a weight of 0.5, RoBERTa-base assigned 0.2, and DeBERTa-v3-base assigned 0.3. For ensemble prediction, the outputs from each model were aggregated using the assigned weights, and a threshold of 0.4 was applied to determine the final binary emotion labels. Despite fine-tuning and ensembling, the F1-score remained comparable to previous approaches. While ensemble learning helped balance predictions across models, it did not significantly outperform any individual fine-tuned model, suggesting that data-related issues were the primary limiting factor rather than model selection.

3.6 Data Analysis and Observations

After observing the stagnant F1 scores across multiple model architectures and fine-tuning strategies, I conducted a detailed analysis of the dataset to identify the underlying bottlenecks that hindered performance improvements. Despite employing transfer learning, fine-tuning different transformer models, and ensemble techniques, the model's overall performance remained largely unchanged. This analysis focused on key aspects such as class imbalance, label co-occurrence patterns, and misclassification trends.

3.6.1 Class Imbalance in the Dataset

One of the most striking observations from the Label Distribution plot was the severe class imbalance:

- Fear was the most dominant emotion, with over 1600 instances, making it disproportionately overrepresented.
- Anger, on the other hand, had significantly fewer examples ($\sim$ 300), making it challenging for the model to learn distinguishing patterns for this emotion.

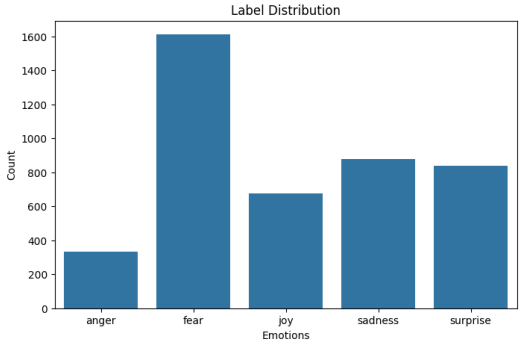


Figure 1: Label Distribution

3.6.2 Label Co-occurrence Patterns

The Label Co-occurrence Matrix further highlighted the complexity of the dataset:

- Fear and sadness frequently co-occurred, indicating a strong semantic overlap between these emotions.
- Surprise and joy also had considerable overlap, making it difficult for the model to differentiate between them.
- Anger, due to its low occurrence and fewer co-occurrences, was particularly difficult for the model to classify accurately.

This highlights the intricate interplay between emotions, which is also playing a part in complicating the classification task and reducing the model's ability to make accurate predictions.

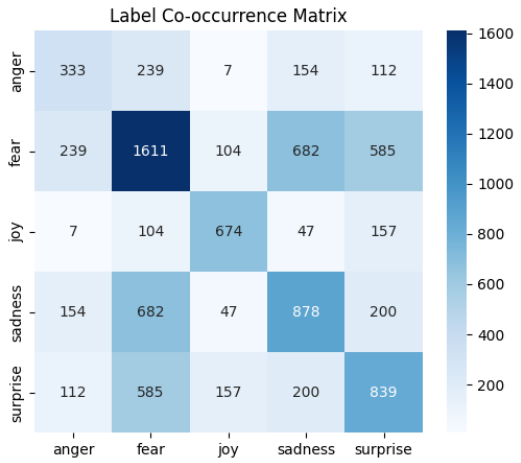
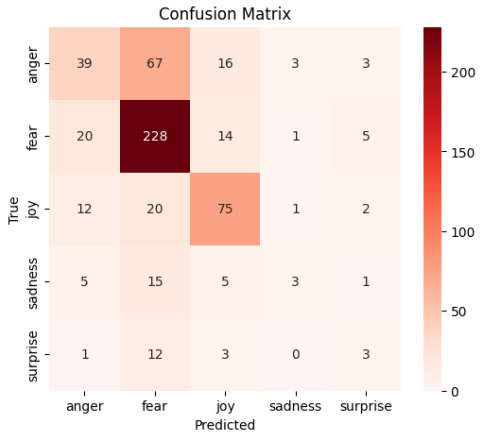


Figure 2: Co-Occurrence of labels

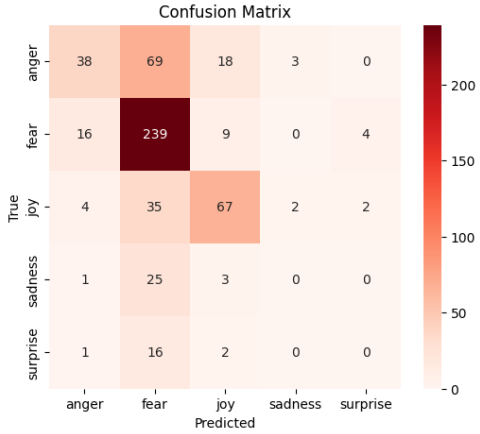
3.7 Confusion Matrix Comparison and Observations

To further analyze model performance, I compared the confusion matrices for the two primary models used:

1. Optimized BERT Model 3a
2. Fine-Tuned Ensemble Model 3b



(a) Confusion Matrix of the Fine-Tuned BERT Model



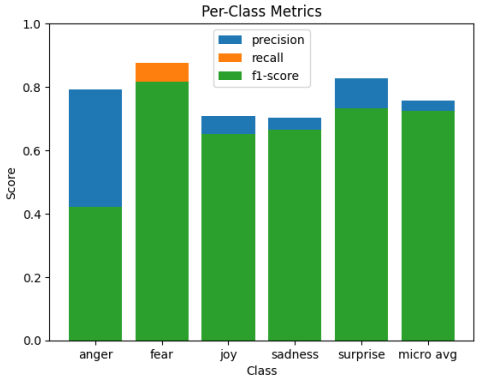
(b) Confusion Matrix of the Fine-Tuned Ensemble Model

Figure 3: Comparison of Misclassification Pattern

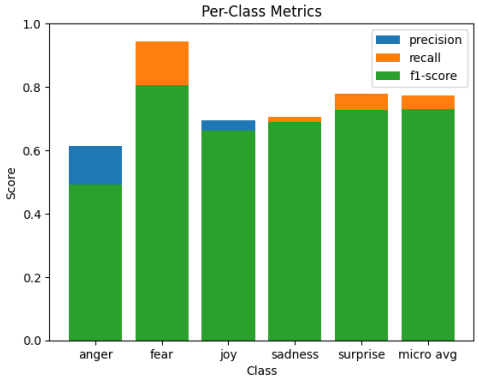
Despite architectural differences and fine-tuning, both models exhibited nearly identical misclassification patterns, indicating that modifying the model alone was insufficient to address performance limitations. Key insights from the confusion matrices include: 'Fear' is the most confidently classified emotion in both models. 'anger' is consistently misclassified as 'fear', reinforcing the observation that these emotions share significant linguistic similarities. 'Joy' saw a slight improvement in recall in the ensemble model (75 correct predictions in Fine-Tuned Ensemble vs. 67 in the BERT model), but its misclassification rates remained high. 'sadness' and 'surprise' remain highly misclassified, likely due to their lower representation in the dataset.

3.8 Per-Class Performance Analysis

Upon further assessing model performance across different emotions, per-class precision, recall, and F1-score, several observations were made.The optimized BERT model demonstrated higher precision in identifying 'anger'; however, it exhibited lower recall, indicating a conservative approach that missed several instances of this emotion. On the other hand, the fine-tuned ensemble model achieved a slight improvement in recall for both 'anger' and 'surprise', but this gain was offset by a decrease in precision, resulting in no significant overall enhancement in the F1-score. Emotions such as 'fear' and 'sadness' were consistently classified accurately, likely due to their higher prevalence in the dataset, which provided the models with more examples to learn from.



(a) Per Class Metrics of the Fine-Tuned BERT Model



(b) Per Class Metrics of the Fine-Tuned Ensemble Model

Figure 4: Comparison of Misclassification Pattern

3.9 Data Augmentation through Backtranslation

Following the analysis of dataset limitations, it was evident that class imbalance and label ambiguity were major bottlenecks affecting model performance. To address these issues, data augmentation techniques were employed to artificially expand the dataset and improve class representation. The primary augmentation method utilized was backtranslation, a widely used technique in natural language processing (NLP) to generate diverse paraphrases of text while preserving its semantic meaning(Keita, 2022).

3.9.1 Addressing Dataset Limitations Through Augmentation

Backtranslation was applied using MarianMT models from the Hugging Face Transformers library, where English text snippets were first translated into French and then translated back into English. This method was specifically applied to augment underrepresented classes in the dataset, ensuring a more balanced class distribution and providing the model with more diverse training examples to improve learning.

Backtranslation Process

Step 1: Translation to French

Original English: *"I snapped my phone shut, clenching my fist around it tightly."*
French Translation: *"J'ai fermé mon téléphone, serrant fermement mon poing autour."*

Step 2: Backtranslation to English

Backtranslated English: *"I closed my phone, clenching my fist tightly around it.".*

This slight variation in structure serves as a valuable augmented snippet to be included in the dataset. The subtle transformation in wording, such as "snapped" being changed to "closed", retains the underlying emotion of anger while introducing diversity in expression. Such modifications help the model learn different ways in which emotions can be conveyed, improving its generalization and robustness in recognizing emotional nuances across varying sentence structures.

Targeted Class Augmentation Strategy

To prevent model bias toward dominant classes and address data scarcity, augmentation was selectively applied:

- Fear (Most Frequent Class): No additional augmentation.
- Anger, Joy, Sadness, and Surprise (Minority Classes):
  - Joy, Sadness, and Surprise: Doubled (2× augmentation)
  - Anger: Tripled (3× augmentation)

This balancing approach ensured that all emotions were adequately represented in the dataset while avoiding excessive distortions.

3.9.2 Training and Impact of Backtranslation Augmentation

After generating the backtranslated dataset, it was merged with the original data to balance class distribution and improve diversity. The BERT-based multi-label classification model was fine-tuned using the same optimized hyperparameters, with 15 training epochs and micro F1-score as the evaluation metric.

This augmentation led to a massive improvement, with the F1-score increasing by approximately 35% compared to the best-performing model on non-augmented data. The model showed better generalization across diverse linguistic expressions and reduced bias toward dominant emotions. These findings confirm that the key bottleneck was data quality, not model architecture. A detailed evaluation of this impact is provided in the Evaluation section.

3.10 Backtranslation with Paraphrasing

In an another attempt to further enhance data augmentation, backtranslation was combined with paraphrasing to introduce greater linguistic diversity while preserving the original meaning of the text. Unlike the previous backtranslation-only approach, which utilized a single language pair (English ↔ French), this method employed randomized selection among multiple language pairs (English ↔ French, German, Spanish, Italian) to increase variability in sentence structure.

Implementation

- **Backtranslation**: Each text snippet was translated from English to a randomly selected language and then back to English, introducing structural variations.
- **Paraphrasing**: The backtranslated text was further rephrased using Pegasus paraphrasing

model (git), generating alternative phrasings while maintaining the original intent.

- **Augmentation Strategy**: Minority emotion classes were selectively expanded using controlled augmentation factors to ensure better class balance in the dataset.

By incorporating randomized language selection and paraphrasing, this approach increased linguistic diversity and robustness, reducing the risk of overfitting to specific phrasing patterns and improving generalization.

3.10.1 Further Augmentation Experiments

Two variations of the above mentioned augmentation strategy were tested:

Moderate Augmentation Volume: Applied a combination of backtranslation and paraphrasing with augmentation ratios of 3× for anger and 2× for other emotions, except fear (0×). This approach resulted in performance improvements over non-augmented models but did not surpass the backtranslation-only model. The decline in performance is likely due to semantic drift and increased label noise.

Higher Augmentation Volume: Increased augmentation to 5× for anger and maintained 2× for other emotions, except fear (0×). This led to slight improvements compared to previous augmentation but still fell short of the backtranslation-only model's performance.

While increasing data volume led to some improvements, excessive augmentation was not pursued due to concerns over data quality loss. Given the multi-label nature of the dataset, increasing one class also impacts co-occurring labels (see figure 2). For instance, anger and fear, or fear and sadness frequently co-occur, meaning augmenting anger or sadness indirectly increases fear instances for many of the cases, only marginally reducing class imbalance. Therefore, this approach was not pursued further.

4 Evaluation

This section evaluates the trained models, highlighting the impact of different architectures, fine-tuning strategies, and augmentation techniques.

4.1 Baseline Model Performance

The fine-tuned BERT-base model trained on the original dataset served as the baseline. While effective in capturing dominant emotions, its F1-score remained moderate due to class imbalance and label ambiguity, especially affecting underrepresented emotions.

4.2 Impact of Backtranslation Augmentation

The first augmentation approach involved backtranslating text exclusively to French and then back to English. As detailed in Section 3.9.1, this process introduced subtle sentence variations while preserving meaning, addressing class imbalance without introducing additional noise.The BERT model trained on data from this approach outperformed all prior models, demonstrating that data quality was the key bottleneck, not model architecture. The improved dataset allowed the model to generalize better, reducing misclassification rates across all emotion classes.

4.3 Backtranslation with Randomized Language Selection & Paraphrasing

A second augmentation strategy (see Section 3.10) introduced randomized backtranslation to one of four languages (French, German, Spanish, or Italian) followed by paraphrasing. This approach aimed to increase linguistic diversity and prevent the model from overfitting to specific translation patterns.

While this hybrid augmentation method improved over non-augmented models, it did not surpass the French-only backtranslation approach. The most likely reason was semantic drift and label noise, as paraphrasing sometimes altered sentence meaning, making classification more challenging (see Section 3.10.1).

4.4 Expanded Augmentation Volume

To examine the impact of scaling augmentation, a higher augmentation ratio was tested (5× for anger, 2× for others except fear (0×)). While this approach showed minor improvements over lower-volume augmentation, it still did not outperform the French-only backtranslation model.

Additionally, since emotions co-occur in multi-label classification (see Section 3.6.2), this was not further pursued for the reason(s) mentioned earlier(see Section 3.10.1).

4.5 Performance Comparison

The following table summarizes the F1-scores for all trained models, with percentage changes calculated relative to the best-performing non-

augmented model. The highest F1-score and percentage increase are highlighted in green.

Table 1: Performance Comparison of Trained Models

Model	**F1 Score**	**% Change**
Baseline Fine-Tuned BERT	0.7033	—
Optimized BERT	0.7245	+3.01%
Custom BERT	0.6960	-1.04%
Fine-Tuned Ensemble	0.7284	+3.57%
Backtranslation Augmented Model	**0.9507**	**+35.18%**
Backtranslation + Paraphrasing (3× Anger, 2× Others, 0× Fear)	0.8918	+26.80%
Expanded Backtranslation + Paraphrasing (5× Anger, 2× Others, 0× Fear)	0.9182	+30.56%

4.6 Per-Class Performance Analysis

The best-performing model was analyzed across individual emotion classes. Figure 5 shows per-class precision, recall, and F1-score, demonstrating significant improvements in recognizing minority emotions like anger and surprise.

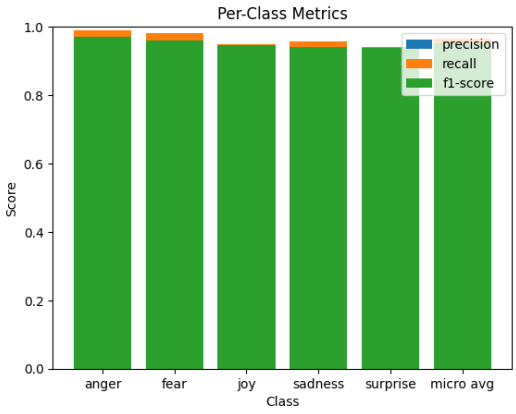


Figure 5: Per-Class Metrics of the Best-Performing Model (Backtranslation only)

4.7 Confusion Matrix Insights

The confusion matrix in Figure 6 further illustrates the improvements achieved by the French-Only Backtranslation model. Misclassifications significantly decreased, especially for underrepresented emotions.

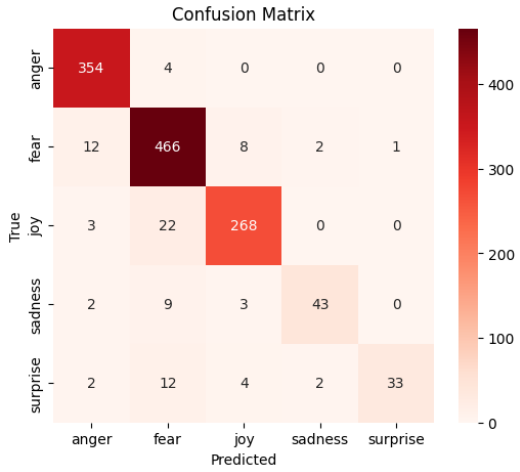


Figure 6: Confusion Matrix of the Best-Performing Model (Backtranslation only)

Key improvements compared to earlier models:

- Anger misclassification dropped significantly(see Figure 3a), increasing recall.
- Fear remained the most confidently classified emotion, but with fewer false positives.
- Overall accuracy signifcantly improved.

5 Discussion

This study underscores the critical importance of data quality and augmentation in text-based emotion detection. While advanced model architectures, such as custom BERT models, offer potential, my findings indicate that enhancing the dataset through backtranslation significantly improves model performance, particularly in addressing class imbalances. This aligns with existing challenges faced and the research emphasizing the necessity of high-quality, well-documented datasets for reliable emotion detection models.

6 Contribution statement

This project was conducted independently, encompassing all facets from conceptualization and experimentation to manuscript preparation. The research leveraged High-Performance Computing (HPC) resources provided by the Center for Information Services and High Performance Computing (ZIH) at TU Dresden. Additionally, Large Language Models (LLMs) were helpful in various aspects of the project.

I would like to express my gratitude to Professors Michael Färber and Simon Razniewski

for their insightful course, "Behind the Secrets of Large Language Models." Their teachings provided valuable perspectives that informed this research. A special thanks to Alisamar Husain, the course tutor, for his invaluable guidance and support.

I also acknowledge the contributions of platforms such as Stack Overflow, Codabench and Hugging Face, just to name a few, which provided invaluable resources and community support throughout this project.

References

GitHub - google-research/pegasus — github.com. https://github.com/google-research/pegasus. [Accessed 06-02-2025].

2024. What Is Transfer Learning? A Guide for Deep Learning | Built In — builtin.com. https://builtin.com/data-science/transfer-learning. [Accessed 04-02-2025].

Idris Abdulmumin. 2025. Codabench — codabench.org. https://www.codabench.org/competitions/3863/. [Accessed 03-02-2025].

Jiaao Chen, Yuwei Wu, and Diyi Yang. 2020. Semi-supervised models via data augmentationfor classifying interactive affective responses. *CoRR*, abs/2004.10972.

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2019. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7:35–51.

Xintong Jiang. 2024. A novel ensemble model for emotion classification in social media text. *Applied and Computational Engineering*, 86:30–41.

Zoumana Keita. 2022. Data Augmentation in NLP Using Back Translation With MarianMT — medium.com. https://medium.com/towards-data-science/data-augmentation-in-nlp-using-back-translation-with-mariannmt-a8939dfea50a [Accessed 05-02-2025].

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *Preprint*, arXiv:1805.04508.

Anna Koufakou, Diego Grisales, Oscar Fox, et al. 2023. Data augmentation for emotion detection in small imbalanced text data. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1508–1513. IEEE.

Atte Kuismin-Raerinne and Liisa Nieminen. 2022. Garbage in, garbage out.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection–techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

Kristína Machová, Martina Szabóova, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Front. Psychol.*, 14:1190326.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

PhD Ron Ozminkowski. 2021. Garbage In, Garbage Out — medium.com. https://medium.com/towards-data-science/garbage-in-garbage-out-721b5b299bc1. [Accessed 03-02-2025].

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Jane Wakefield. 2021. Ai emotion-detection software tested on uyghurs — bbc.com. https://www.bbc.com/news/technology-57101248. [Accessed 03-02-2025].

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Lieve Van Woensel and Nissy Nevil. 2019. What if your emotions were tracked to spy on you? *URL https://www. europarl. europa. eu/RegData/etudes/ATAG/2019/634415/EPRS ATA (2019)*, 634415.

Harsh Yadav. 2022. Dropout in Neural Networks | Towards Data Science — towardsdatascience.com. https://towardsdatascience.com/dropout-in-neural-networks-47a162d621d9/. [Accessed 04-02-2025].

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 1–13, Abu Dhabi, UAE. Association for Computational Linguistics.