Multi-Emotion Classification: Back-Translation Augmentation and Model Contrast

1st Honghao Wang
Dept. of Statistics
University of Michigan, Ann Arbor
Ann Arbor, United States
whonghao@umich.edu

***Abstract*—Sentiment analysis is a key task in Natural Language Processing (NLP) with wide applications, but binary classification benchmarks cannot meet industrial multi-class needs. This study focuses on multi-class sentiment classification using the dair-ai/emotion dataset (6 emotion categories, notable class imbalance), exploring three core objectives: the impact of multilingual back-translation augmentation, performance differences between fine-tuned BERT models and traditional machine learning (TF-IDF+SVM/XGBoost/Random Forest), and effects of BERT parameter scales. Experimental results show that back-translation effectively balances class distribution and improves minority category recognition; fine-tuned BERT outperforms traditional methods via contextual semantics capture; larger-parameter BERT yields limited performance gains with higher training costs. This study validates data augmentation and pre-trained model advantages for imbalanced multi-class sentiment tasks.**

***Index Terms*—Sentiment Analysis, Data Augmentation, Large Language Models, Class Imbalance, Natural Language Processing.**

I. Introduction

A. Overview

Sentiment analysis, a key task in Natural Language Processing (NLP), is widely applied in practical scenarios—such as monitoring social media public opinion, e-commerce customer feedback analysis, and early warning of financial user sentiment risk. Its results directly provide data-driven support for strategic enterprise decisions and product iteration optimization.

Among popular GLUE (General Language Understanding Evaluation) benchmarks in NLP, only the SST-2 task (Stanford Sentiment Treebank-2) focuses on sentence-level sentiment analysis. However, SST-2 is limited to binary positive-negative classification and cannot meet industrial multi-class sentiment classification needs. Thus, this project adopts the public dair-ai/emotion dataset from Hugging Face. It covers 6 emotion categories (joy, sadness, anger, fear, surprise, love) and serves as a classic multi-class sentiment analysis benchmark. Nevertheless, the data set has an unbalanced sample distribution in some categories. Additionally, research gaps exist in validating specific data augmentation technologies, comparing traditional machine learning with pre-trained large models, and analyzing performance differences among pre-trained models of varying parameter scales—all rendering the dataset practically and academically valuable.

The core research directions of the project are threefold: (1) Exploring how data augmentation based on back-translation affects model training for multi-class sentiment tasks; (2) Verifying performance differences between pre-trained large model fine-tuning and traditional machine learning methods; (3) Analyzing performance improvements of pre-trained large models with different parameter scales on the task.

B. Related Work

In text sentiment classification, data sparsity and class imbalance have long constrained model performance. Text augmentation—by generating semantically consistent yet diverse samples—has become a key solution to these issues. In basic text operation augmentation: Wei proposed the EDA method, which combines four core operations (e.g. synonym replacement, random insertion) to provide a simple and efficient augmentation approach for small-sample text classification [1]. Guo applied the Mix-up method to sentence classification, expanding its scope through linear interpolation of random sample features and synchronous label distribution modeling [2]. In translation-driven augmentation: Body adopted multilingual back-translation, leveraging cross-linguistic expression differences to enrich text diversity [3]. Kumar's research verified that Transformer-based translation models excel in preserving sentiment semantics [4]. In generative augmentation: Whitfield used the GPT-2 model to generate new samples with consistent sentiment tendencies, effectively improving data set diversity and coverage [5]. After addressing class imbalance, two main approaches dominate text sentiment classification: the traditional "representation-then-classification" framework and end-to-end classification with large language models. For traditional methods: Salton introduced the TF-IDF algorithm to quantify the importance of words in text [6]; Mikolov proposed Word2Vec, mapping words to dense low-dimensional vectors to allow automatic clustering of semantically similar terms [7]. For classifiers: Pang adopted SVM as a classic traditional sentiment classification model, whose suitability for high-dimensional text features is widely recognized [8]; P Dhanalakshmi combined logistic regression with the VADER tool, achieving superior performance in analyzing the sentiment of airline tweets [9]; Samih used the IWVS method and improved the F1-score using the XGboost algorithm [10]. For large language model-based classification: Devlin adapted

the BERT model to text classification through fine-tuning, establishing a practical framework for applying pre-trained models to this field [11]; Liu proposed an optimized BERT pre-training strategy (RoBERTa), which improves the stability and precision of text classification by adjusting the pre-training data scale and re-inspection of training details [12].

II. Method

A. Problem Formulation

This study focuses on multi-class sentiment classification. Its core goal is to accurately map text inputs to six predefined emotion categories (joy, sadness, anger, fear, surprise, love). Formally, let X denote the text input space (i.e., the set of all text samples), and a single text sample is denoted by $x \in X$. The set of emotion categories is $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$, which corresponds to the six aforementioned emotions. The core of this task is to learn a classification function $f : X \rightarrow Y$, which minimizes the model's prediction error in unseen test samples and thus ensures its generalizability.

To objectively evaluate the classification performance of the model, three key metrics are adopted:

- Accuracy: The proportion of correctly classified samples among all. intuitively reflects the overall classification effect of the model;
- Weighted F1-score: The F1-score weighted by the sample size of each category. It effectively mitigates the impact of class imbalance on evaluation results and improves the fairness of performance comparison;
- Confusion Matrix (CofMtx): It clearly shows the model's ability to recognize majority and minority classes, and intuitively reflects misclassification between categories, providing clear directions for future model optimization.

B. Dataset Description

This experiment uses the public dair-ai/emotion dataset from the Hugging Face Hub, a classic benchmark for multi-class sentiment analysis tasks. The data set contains text samples annotated into six emotion categories, with notable class imbalance: joy (5,362 samples) and sadness (4,666 samples) are majority classes, while love (1,304 samples) and surprise (572 samples) are minority classes. The lengths of the text in the data set range from 3 to 70 words, and there is no significant correlation between the length of the text and the emotion category.

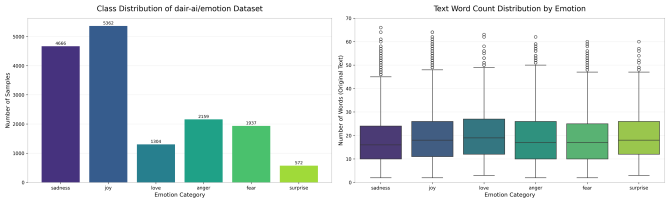

Fig. 1. Data Distribution of the dair-ai/emotion Dataset

C. Methodological Framework

a) Multilingual Back-Translation Strategy: To alleviate the class imbalance of the dataset, this study adopts a multilingual back-translation strategy to generate augmented samples with consistent semantics and diverse lexicon. Eight pairs of English-foreign languages are selected: English-German, English-Spanish, English-French, English-Chinese, English-Russian, English-Arabic, English-Italian and English-Dutch, to ensure the diversity of expression transformations. The translation models uniformly use the Helsinki-NLP/opus-mt series from the Hugging Face Hub, which perform excellently in preserving sentiment semantics. The specific method is as follows: for each emotion category, randomly sample one original sample and generate a new sample through bidirectional translation ("source language $\rightarrow$ target language $\rightarrow$ source language") using a randomly selected language pair; repeat the process after deduplication until the sample size of each category is not less than 80% of the largest category (joy) in the original dataset.

b) TF-IDF Feature Representation: All traditional machine learning models in this study adopt the "TF-IDF representation + classifier" framework. Text is converted into structured feature vectors via TF-IDF to capture key semantic information in the text.Term Frequency (TF) refers to the frequency of a word in a single sample. Inverse Document Frequency (IDF) reflects the importance of a word throughout the entire data set. The TF-IDF value of a word is defined as

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \text{IDF}(w) \tag{1}$$

where w denotes a word and d denotes a sample.

To balance feature dimensionality and information retention, the maximum number of features for TF-IDF is set to 5000. This parameter is consistent across all traditional classifiers to ensure the fairness and reproducibility of the experiment.

c) Traditional Machine Learning Classifiers: Following the suggestions of the teaching assistant, three different machine learning methods are adopted for comparative experiments.

- Support Vector Machine: SVM is a discriminative classifier that uses the kernel trick to map text features into a high-dimensional space, thereby finding the optimal separating hyperplane to distinguish different emotion categories. The linear kernel function is selected and the model parameter is set as $C = 1.0$.
- XGBoost [13]: it is an optimized ensemble learning algorithm based on Gradient-Boosted Decision Trees (GBDT). It iteratively constructs weak classifiers and optimizes the model using gradient information from the loss function, gradually reducing prediction errors. The key parameters of the model are configured as follows: $n\_estimators = 100$ and $learning\_rate = 0.1$.
- Random Forest [14]: it is an ensemble classifier composed of multiple independent decision trees. The final prediction result is generated by majority voting of the outputs from individual decision trees. The model parameter is set as $n\_estimators = 100$.

d) BERT Series Fine Tuning: Two BERT-series pre-trained models with different parameter scales are selected for end-to-end classification:

- bert-base-uncased: A lightweight model with 110 million parameters, balancing training efficiency and basic semantic capture capability;
- bert-large-uncased: A large-scale model with 340 million parameters, designed to capture more refined and complex semantic features.

In the fine-tuning process, we freeze the bottom feature extraction layers to retain the general language knowledge learned from pre-training; only the top classification layer is fine-tuned to adapt to the specific multi-class sentiment classification task. Consistent training parameters are used for both models: $max\_length = 128$, $batch\_size = 16$, $learning\_rate = 2e\text{-}5$, $training\_epochs = 4$, the loss function adopts the cross-entropy loss function.

III. Results

A. Validation of Data Augmentation Effect

The original dair-ai/emotion dataset exhibits significant class imbalance (fig 1), after augmentation via the multilingual back-translation strategy, the sample sizes of "love", "anger", "fear" and "surprise" are all increased to 4289, effectively balancing the class distribution; meanwhile, the text word count distribution of the augmented dataset (fig 2) is consistent with the original data, with no distortion introduced.

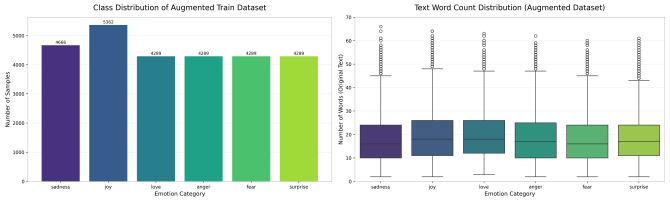

Fig. 2. Data Distribution of the augmented Dataset

Table I presents the corresponding quantitative metrics, while Figure 3 displays the confusion matrices of TF-IDF+SVM in the original and augmented data: model precision (0.8670) and weighted F1 score (0.8703) in the augmented data slightly decrease compared to the original data, but the recognition performance of minority categories improves significantly—specifically, the accuracy of the "surprise" category increases from 0.58 to 0.77, and that of the "love" category rises from 0.74 to 0.86. This result indicates that the multilingual back-translation data augmentation strategy effectively alleviates the class imbalance issue, enabling the model to better learn features of scarce classes and enhancing its robustness and practical value in real-world imbalanced data scenarios.

B. Comparison Between Traditional ML and Fine-Tuned LLM

Quantitative results in Table II show that the Bert_base_ft_emotion model achieves significantly better overall performance than traditional machine learning models

TABLE I
THE TF-IDF+SVM ON ORIGINAL AND AUGMENTED DATA

Method	Dataset	Accuracy	Weighted F1 Score
TF-IDF+SVM	original_data	**0.8935**	**0.8921**
	augmented_data	0.8670	0.8703

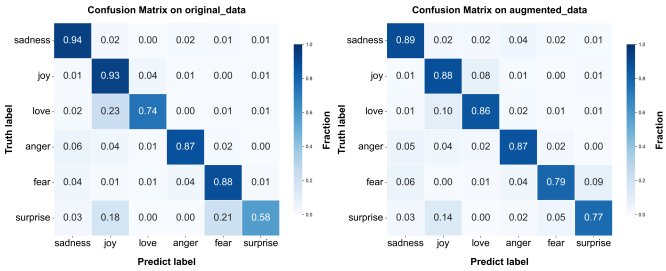

Fig. 3. CofMtx of the TF-IDF+SVM on Original Data and Augmented Data

(including TF-IDF+SVM, TF-IDF+XGBoost, and TF-IDF+Random_forest)—in terms of overall accuracy (0.9300), weighted F1 score (0.9288), and recognition rate of most emotion categories, there are notable improvements, with TF-IDF+SVM being the best-performing traditional model; confusion matrix results indicate that 18% of "surprise" samples are misclassified as "joy" in TF-IDF+SVM model, while Bert_base_ft_emotion completely avoids this misclassification. It significantly improves the classification accuracy of most categories (e.g., the recognition rate of "surprise" itself increases to 76%), with only the recognition rate of the fear category 0.72 (slightly lower than the 0.74 of the TF-IDF+SVM method), and the misclassification between other similar emotions is significantly reduced—this difference comes from the fact that bert_base_uncased can capture contextual semantics of text, while traditional models rely only on the lexical features of TF-IDF.

TABLE II
COMPARISON BETWEEN TRADITIONAL ML AND FINE-TUNED LLM

Method	Accuracy	Weighted F1 Score
TF-IDF+SVM	0.8935	0.8921
TF-IDF+XGboost	0.8520	0.8530
TF-IDF+Random_forest	0.8820	0.8819
Bert_base_ft_emotion	**0.9300**	**0.9288**

C. Comparison of Fine-Tuned LLM with Different Parameters

Table III presents the performance comparison between Bert_base_ft_emotion (110M parameters) and Bert_large_ft_emotion (340M parameters): both models achieve the same overall accuracy of 0.9300, while the weighted F1 score of Bert_large_ft_emotion slightly increases to 0.9295. Confusion matrix results further show that Bert_large_ft_emotion improves the recognition of most categories (especially minority classes): the accuracy of "love" rises from 0.72 to 0.77, "anger" from 0.91 to 0.94, and "surprise" from 0.76 to 0.82; however, the accuracy

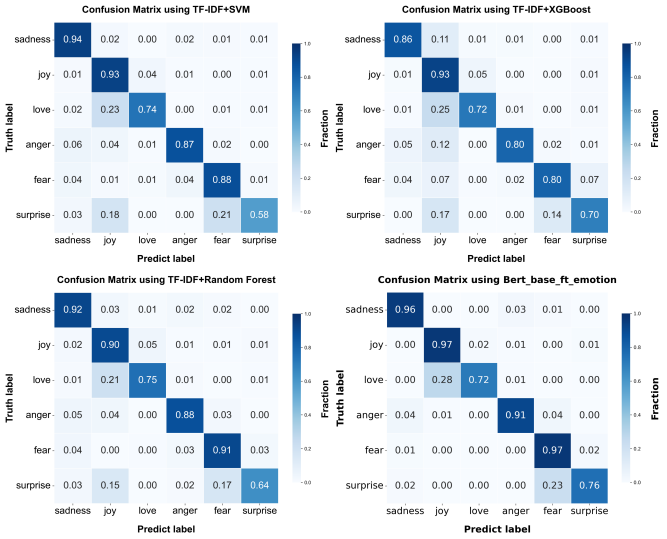

Fig. 4. CofMtx of the Traditional ML and Fine-Tuned LLM

of "fear" drops from 0.97 to 0.85. This indicates that the larger-parameter model can capture partial emotional features more precisely, but its fitting effect fluctuates for specific categories, leading to relatively limited overall performance gain.

TABLE III
COMPARISON OF FINE-TUNED LLM WITH DIFFERENT PARAMETERS

Method	Accuracy	Weighted F1 Score
Bert_base_ft_emotion	**0.9300**	0.9288
Bert_large_ft_emotion	**0.9300**	**0.9295**

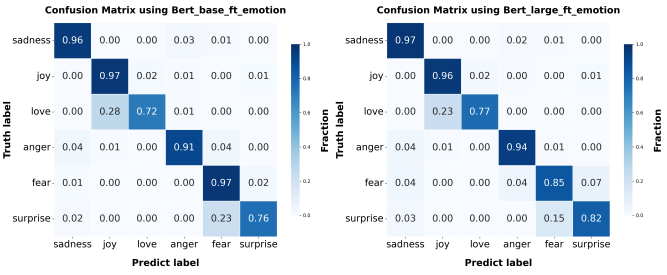

Fig. 5. CofMtx of Fine-Tuned LLM with Different Parameters

IV. CONCLUSION

In response to the three main objectives proposed in the introduction, this study draws the following conclusions:

(1) The multilingual back-translation data augmentation strategy can effectively alleviate the class imbalance problem of the data set. It also retains the core features of the original text without introducing data distortion. Although the overall performance of the model fluctuates slightly after augmentation, it significantly improves the recognition effect of minority categories, which helps to improve the applicability and robustness of the model in real-world imbalanced data scenarios.

(2) The fine-tuned pre-trained large model has better comprehensive performance than traditional machine learning methods in multi-class sentiment tasks. Traditional machine learning methods are based on lexical-level feature representation and are difficult to capture the contextual semantic information of text. In contrast, pre-trained large models have strong contextual semantic encoding capabilities, which can better distinguish similar emotions and reduce classification confusion.

(3) Pre-trained large models with larger parameter scales have certain improvements in the recognition of some emotional categories and can capture complex emotional features more accurately. However, as the parameter scale increases, the training cost of the model increases significantly, and the performance of some categories may fluctuate, with limited overall performance improvement.

REFERENCES

[1] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," arXiv preprint arXiv:1901.11196, 2019.
[2] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," arXiv preprint arXiv:1905.08941, 2019.
[3] T. Body, X. Tao, Y. Li, L. Li, and N. Zhong, "Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models," Expert Systems with Applications, vol. 178, p. 115033, 2021.
[4] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," arXiv preprint arXiv:2003.02245, 2020.
[5] D. Whitfield, "Using GPT-2 to create synthetic data to improve the prediction performance of NLP machine learning classification models," arXiv preprint arXiv:2104.10658, 2021.
[6] K. Sparck Jones, "Index term weighting," Information Storage and Retrieval, vol. 9, pp. 619–633, 1973.
[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint arXiv:cs/0205070, 2002.
[9] P. Dhanalakshmi, G. A. Kumar, B. S. Satwik, K. Sreeranga, A. T. Sai, and G. Jashwanth, "Sentiment analysis using VADER and logistic regression techniques," in 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 139–144.
[10] A. Samih, A. Ghadi, and A. Fennan, "Enhanced sentiment analysis based on improved word embeddings and XGboost," International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no. 2, pp. 1827–1836, 2023.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[12] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.
[14] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.