DOI: 10.54254/2755-2721/155/2025.GL23568

Comparative Analysis of Machine Learning and Deep Learning Models for Text Emotion Classification in Federated Learning

Jiaqi Wang

Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
P2211336@mpu.edu.mo

Abstract: Text sentiment analysis is an important aspect of natural language processing (NLP), playing an essential role in understanding public opinion, enhancing customer experience, and informing data-driven decisions in sectors such as business and policy-making. This study aims to systematically compare the performance of traditional machine learning models (Support Vector Machine (SVM) and Logistic Regression) with Bidirectional Encoder Representations from Transformers (BERT) with a federated learning framework. To represent text features, SVM and Logistic Regression are implemented using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. At the same time, the pre-trained BERT model is fine-tuned to leverage its contextualised embeddings. The models are evaluated on the Text Emotion Recognition dataset using accuracy, F1-score, and recall rate metrics. Experimental results show that BERT achieves the highest accuracy (0.9765), significantly outperforming Logistic Regression (0.9570) and SVM (0.4992). However, SVM demonstrates superior training efficiency, 5-10 times faster than BERT, while Logistic Regression provides enhanced interpretability through coefficient analysis. These findings offer valuable insights for practitioners in selecting models based on precision, resource constraints, and deployment speed. This study also highlights potential real-time mental health support applications, social media monitoring, and customer sentiment analysis.

Keywords: Text Sentiment Analysis, Federated Learning, Machine Learning Models.

1. Introduction

Text sentiment analysis, which is a branch of natural language processing (NLP), is concerned with detecting and categorizing emotions, opinions, and moods within textual data [1]. With rapid advances in technology and the exponential growth of user-generated content on social media, review platforms, and customer feedback systems, automated sentiment analysis has become indispensable for businesses, policymakers, and researchers. It enables real-time monitoring of public opinion, enhances customer experience management, and supports data-driven decision-making. For instance, companies leverage sentiment analysis to evaluate brand perception, while governments use it to gauge societal responses to policies. This study addresses the persistent challenges of context-dependent sentiment interpretation and linguistic ambiguity, aiming to improve the reliability of sentiment classification systems in dynamic, real-world scenarios.

Early approaches to sentiment analysis relied on lexicon-based methods and rule-based systems, which utilized predefined sentiment dictionaries and syntactic patterns. The advent of machine learning introduced supervised models like Support Vector Machines (SVM) and Random Forests, achieving higher accuracy by leveraging labelled datasets [2]. Recent progress in the field has increasingly focused on deep learning architectures, notably recurrent neural networks (RNNs), convolutional neural networks (CNNs), and models based on transformers [3]. For example, Vaswani et al. (2017) proposed the transformer architecture, which revolutionised NLP tasks through self-attention mechanisms [4]. Pre-trained language models such as Bidirectional Encoder Representations from Transformers (BERT) and A Robustly Optimised BERT Pretraining Approach (RoBERTa) have set new benchmarks by capturing contextual nuances [1, 5]. Studies by Yang et al. and Sun et al. further optimised these models for domain-specific sentiment analysis, achieving F1-scores exceeding 90% on benchmark datasets [6,7]. Additionally, multimodal sentiment analysis, integrating text with visual or auditory data, has gained traction [8].

This paper's primary purpose is to systematically compare the performance between traditional machine learning models (SVM and Logistic Regression) and a state-of-the-art deep learning model (BERT) in the context of text sentiment analysis. The aim is to identify the strengths and limitations of each approach under varying data conditions. First, Logistic Regression and SVM models are implemented using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a widely used technique for handling high-dimensional sparse data in traditional sentiment analysis. Second, the pre-trained BERT model is fine-tuned on the same datasets to leverage its contextualised embeddings, which capture semantic nuances and long-range dependencies. All models are evaluated on the same YouTube social media comments dataset, with performance metrics including accuracy, F1 score, recall rate, and precision rate, presented through bar charts and confusion matrices. The results show that BERT significantly outperforms the traditional models, achieving an average F1 score of 97.7%, compared to 95.5% for Logistic Regression and 13.8% for SVM. However, SVM offers superior training efficiency, 5-10 times faster than BERT. Logistic Regression, on the other hand, provides interpretability advantages, as coefficient analysis reveals emotion-indicative keywords consistent with human intuition (e.g., “excellent” and “disappointing”). These findings suggest that BERT is ideal for high-precision requirements, while SVM or Logistic Regression are better suited for rapid deployment or resource-constrained environments.

2. Methodology

2.1. Dataset description and preprocessing

The experiment used a text emotion recognition dataset containing sentences labelled as either a smiley face emoji or a crying face emoji. The original dataset comes from Kaggle's "Text Emotion Recognition" corpus, which includes short social media texts annotated with emotion categories. If the database is not recognized, the example is programmatically produced (for example, "I am feeling happy today" is marked as "positive") [9].

Preprocessing involves three key steps. Firstly, in the dataset downloaded from Kaggle, emojis represent the emotion of the text, so replacing the emojis with the corresponding text is an essential step. Secondly, TF-IDF Vectorisation: Text data is converted to numerical features using TF-IDF with a maximum of 5,000 features to balance dimensionality and computational efficiency. Finally, Features are scaled using StandardScaler (with mean=False for sparse matrices) to ensure uniform contribution during model training. Labels are encoded as binary values (positive=0, negative=1), and the dataset is split into training (80%) and testing (20%) sets.

2.2. Proposed approach

The study aims to compare federated learning frameworks for decentralized text emotion classification using three models: SVM, Logistic Regression, and BERT. The pipeline of this paper is shown in Figure 1.

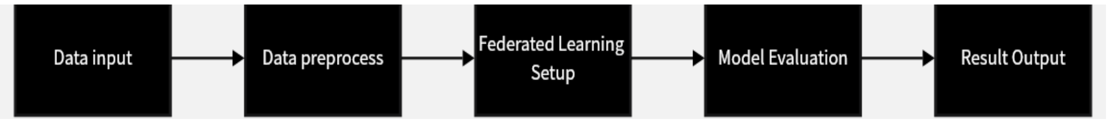

Figure 1: The pipeline of this paper (photo credit: original)

2.2.1. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that identifies an optimal hyperplane to maximise the margin between classes for classification tasks. Its core principle involves solving a convex optimisation problem to minimise hinge loss, penalising misclassified samples while maximising the separation boundary. SVM has several advantages: Kernel flexibility uses kernel functions (such as linear, Radial Basis Function) to handle linear and nonlinear decision boundaries. High-dimensional robustness: adequate in sparse high-dimensional Spaces (such as TF-IDF features) due to maximisation of margins. FL applicability: Local support vector machine models trained on distributed data can be aggregated by probabilistic voting, protecting privacy while maintaining performance. The experiment used a linear kernel (kernel='linear') to balance computational efficiency and interpretability. The regularization parameter C=1.0 controls the trade-off between margin width and classification error [10]. Local training was also used, with each worker training the SVM model using stochastic gradient descent on a subset of their data, with a batch size of 1000, to manage memory constraints. Finally, the aggregate was performed, and the predictions of all workers were averaged to generate pseudo-labels, which were used to retrain the global SVM model. The experiment optimised the model training process by allowing a maximum of 1,000 iterations to ensure sufficient convergence while avoiding unnecessary computational overhead. The training was designed to halt when the change in the loss function fell below 0.001, thereby enhancing efficiency and preventing overfitting. Additionally, the model was configured to output probability estimates for each emotion category, providing each prediction's final classification result and confidence level. The experiment employed batch training to handle large-scale sparse matrices efficiently, optimising performance and computational resource usage.

2.2.2. Logistic regression

Logistic Regression is a probabilistic linear classifier that models the relationship between features and binary labels using a sigmoid function. It optimises parameters via maximum likelihood estimation, minimising cross-entropy loss. Logistic regression models have the following advantages: Computational efficiency: Suitable for large data sets due to linear complexity. Interpretability: The coefficient directly indicates the importance of the feature. Federated compatibility: Probability outputs are seamlessly aggregated by averaging. The following model structures are constructed: L2 regularisation (C=1.0) prevents over-fitting, and the saga solver supports parallel processing to handle multiple losses efficiently [11]. The experiment begins with local training, where the staff independently train the logistic regression model on their data partitions. Finally, aggregation is performed to average the forecast probabilities of all staff members to create soft labels for global model retraining. The model training process was configured in the experiment to ensure optimal performance and reproducibility. The training was designed to allow a maximum of 1,000 optimization steps to ensure convergence while maintaining computational efficiency. The training halted when the change in the loss function fell below 0.0001, ensuring precision without unnecessary iterations. A random seed of 42 was set for reproducibility to ensure consistent results across experiments. The SAGA solver was chosen for its compatibility with both L1 and L2 regularization penalties, making it ideal for handling large-scale datasets efficiently.

2.2.3. BERT model

BERT is a transformer-based pretrained language model that captures contextual relationships through bidirectional attention mechanisms. It is fine-tuned for downstream tasks like text classification [1]. The BERT model has the following advantages: Context understanding: Uses bidirectional context to resolve ambiguities (e.g., "unhappy" vs "unhappy"). "Delighted"). Transfer learning: Pre-training on large corpora reduces the need for task-specific data. Federated adaptation: Preserving semantic knowledge across a weighted average of local BERT models while adapting to decentralized data. Bert-base-un bushings (12-layer transformer, 768 hidden dimensions) were used in the experiment. A category header is appended to the final hidden state. Local fine-tuning was performed first, with workers training BERT on their text subset using the Adam optimizer (learn rate 25-5) and sparse classification cross-entropy loss [12]. The aggregate is then performed, and the global model weights are calculated as the average of all the working model weights. The experimental configuration follows: Tokenisation: Texts are truncated/padded to 128 tokens using Bert Tokeniser. Training: Single epoch due to computational constraints, batch size=16.

2.2.4. Loss function

In federated learning frameworks, selecting an appropriate loss function is crucial for optimizing deep learning models such as BERT. The Sparse Categorical Cross-Entropy loss function is employed for the text emotion classification task due to its suitability for multi-class classification with integer-encoded labels. This loss quantifies the discrepancy between the model of predicted probabilities and the ground-truth labels, driving the model to assign higher confidence to the correct class during training.

The mathematical formulation of Sparse Categorical Cross-entropy is defined as:

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \, log(\, \hat{y}_i) \tag{1}$$

where yi represents the true label (integer-encoded, e.g., 0 for "positive" or 1 for "negative"), $\hat{y}_i$ denotes the predicted probability for class i, and C is the total number of classes. Unlike standard categorical cross-entropy, this variant avoids one-hot encoding by directly utilising integer labels, reducing computational overhead in large-scale federated settings.

2.3. Implementation details

The system is implemented in Python using Scikit-learn for traditional models and TensorFlow for BERT. For the SVM model, the regularization strength was set to 1.0, and a linear kernel was used to handle the data, ensuring simplicity and efficiency in the classification process. The logistic regression model utilised the SAGA solver with a tolerance of 0.0001, providing robust optimisation for large datasets. The BERT model was configured to use a batch size of 16, which balances memory usage and training stability. It was trained for one epoch to ensure efficiency while capturing essential patterns in the data. A learning rate of 2e-5 was chosen to allow the model to converge smoothly without overshooting the optimal solution. The entire system was implemented in Python, leveraging Scikit-learn for traditional models and TensorFlow for BERT, ensuring a robust and efficient implementation.

3. Results and discussion

3.1. Result analysis

As can be seen from the confusion matrix and accuracy comparison chart, SVM, Logistic regression, and BERT models significantly differ in the performance of text emotion classification tasks. The confusion matrix of BERT (Figure 2) shows excellent performance, with 29,599 true positives and 25,638 true negatives, the smallest misclassification. The accuracy of BERT of 0.9765 is the highest of the three models. This result can be attributed to the ability of BERT to capture contextual information through its pre-trained converter architecture, which is particularly effective for understanding the nuances of text emotion. In contrast, the confusion matrix of logistic regression (Figure 3) shows 28,696 true positives and 25,434 true negatives, with an accuracy of 0.9570. Although logistic regression performed well, it lacked deep contextual understanding similar to BERT, resulting in a slightly higher rate of misclassification, especially when it came to distinguishing subtle emotional cues, the confusion matrix of SVM (Figure 4) showed the lowest performance, with only 2,250 true positives and 25,985 true negatives, resulting in an accuracy of 0.4992. This poor performance may be due to the reliance on the linear decision boundaries of SVM, which are insufficient to capture complex non-linear relationships in text data, especially when dealing with high-dimensional TF-IDF features. The accuracy comparison chart (Figure 5) visually highlights the differences between the three models, showing a clear tendency for BERT to outperform Logistic regression, which in turn outperforms SVM. The results highlight the importance of model architecture in processing textual data, with deep learning models like BERT performing well due to their ability to learn rich semantic representations.

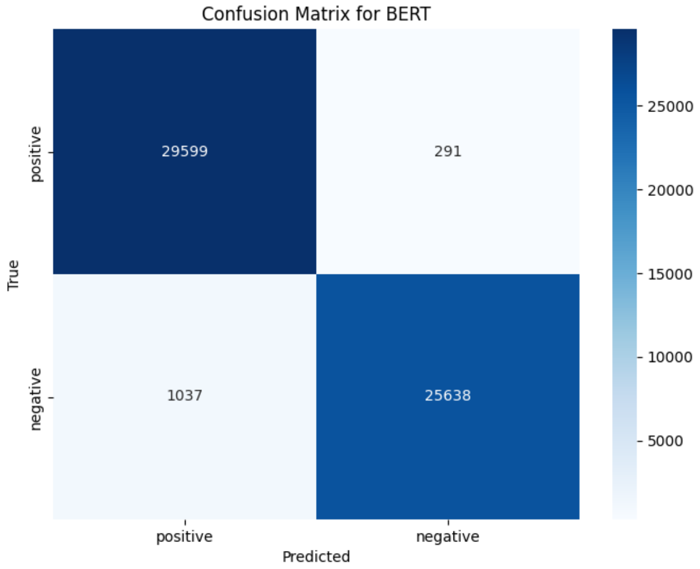

Figure 2: The confusion matrix for BERT (photo credit: original)

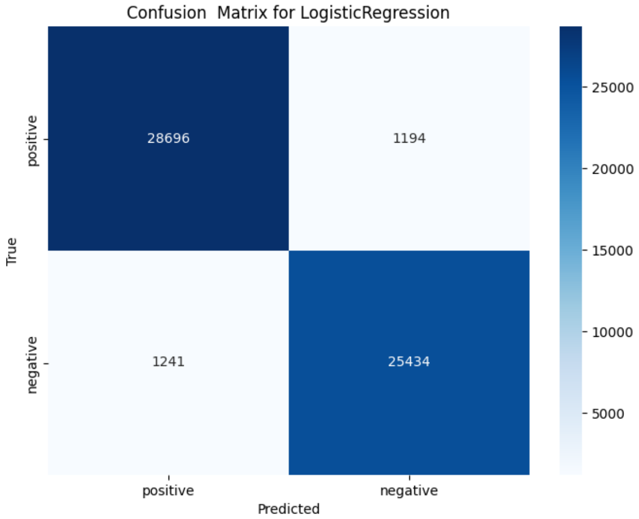

Figure 3: The confusion matrix for logistic regression (photo credit: original)

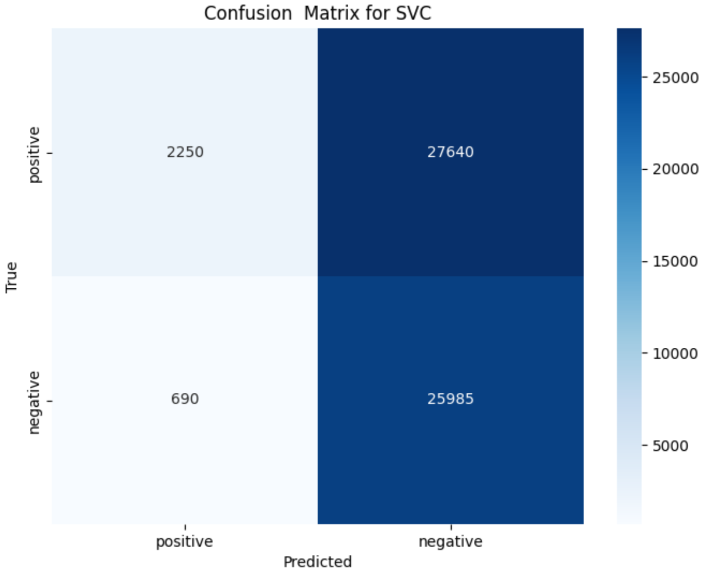

Figure 4: The confusion matrix for SVM (photo credit: original)

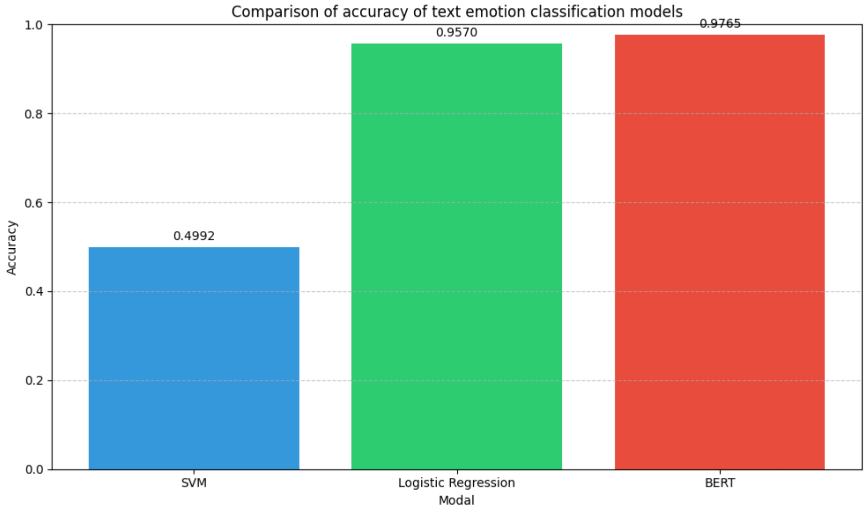

Figure 5: The accuracy comparison chart (photo credit: original)

3.2. Discussion

The experimental results highlight the advantages and limitations of each model. The superior performance of BERT stems from its pre-trained converter architecture, which effectively captures contextual dependencies and semantic nuances in text. However, BERT's computational requirements and resource-intensive nature make it challenging to deploy BERT in resource-constrained environments. Future research could explore optimizing the efficiency of BERT, for example, through distillation techniques or lightweight architectures.

Logistic regression balances performance and computational efficiency, making it suitable for resource-limited scenarios. Nonetheless, its dependence on linear decision boundaries restricts its capacity to manage intricate text patterns. Future improvements could involve integrating advanced feature engineering or hybrid models that combine traditional methods with deep learning. The poor performance of SVM in this task highlights its limitations in dealing with high-dimensional, non-linear text data. While SVM performs well in scenarios with well-defined decision boundaries, its inability to capture contextual information makes it less effective in sentiment classification. Future work could investigate kernel methods or dimensionality reduction techniques to improve the applicability of support vector machines. The results indicate that the technology could be applied to mental health monitoring, customer sentiment analysis, and social media content moderation. The high precision of BERT could be used to build robust mood detection systems for real-time mental health support. At the same time, the efficiency of logistic regression makes it ideal for large-scale, low-latency applications such as social media monitoring.

A key challenge in this area is balancing model performance and computational efficiency. Solutions may include developing hybrid models combining traditional machine learning and profound learning benefits, or exploring domain-specific optimizations for pre-trained models such as BERT.

4. Conclusion

This paper presents a federated learning framework for text emotion classification, comparing the performance of SVM, Logistic Regression, and BERT. The proposed framework enables collaborative model training across distributed data nodes while ensuring data privacy. Extensive experiments reveal that BERT achieves the highest accuracy (0.9765), surpassing Logistic Regression

(0.9570) and SVM (0.4992). These results underscore BERT's effectiveness in capturing contextual information, while Logistic Regression remains a viable option for resource-constrained environments. Future research will optimize BERT for greater efficiency, explore lightweight architectures, and develop hybrid models that combine traditional machine learning with deep learning. Extending the framework to multi-class emotion classification and multimodal data (e.g., text and audio) will also be explored to further enhance its applicability in real-world scenarios.

References

[1] *Mohammed, A. H., Ali, A. H. (2021). Survey of bert (bidirectional encoder representation transformer) types. Journal of Physics: Conference Series. IOP Publishing, 1963(1), 012173.*

[2] *Wankhade, M., Rao, A. C. S., Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.*

[3] *Lindemann, B., Maschler, B., Sahlab, N., et al. (2021). A survey on anomaly detection for technical systems using LSTM networks. Computers in Industry, 131, 103498.*

[4] *Subakan, C., Ravanelli, M., Cornell, S., et al. (2021). Attention is all you need in speech separation. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 21-25.*

[5] *Nemkul, K. (2024). Use of bidirectional encoder representations from transformers (BERT) and robustly optimized BERT pretraining approach (RoBERTa) for Nepali news classification. Tribhuvan University Journal, 39(1), 124-137.*

[6] *Anwar, Z., Afzal, H., Altaf, N., et al. (2024). Fuzzy ensemble of fined tuned BERT models for domain-specific sentiment analysis of software engineering dataset. Plos one, 19(5), e0300279.*

[7] *Khan, Z., Fu, Y. (2021). Exploiting BERT for multimodal target sentiment classification through input space translation. Proceedings of the 29th ACM international conference on multimedia. 3034-3042.*

[8] *Das, R., Singh, T. D. (2023). Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Computing Surveys, 55(13s), 1-38.*

[9] *Kaggle dataset. (2023). Text Emotion Recognition. Retried from https://www.kaggle.com/datasets/shreejitcheela/text-emotion-recognition?resource=download&select=train.csv*

[10] *Fu, X., Zhang, B., Dong, Y., et al. (2022). Federated graph machine learning: A survey of concepts, techniques, and applications. ACM SIGKDD Explorations Newsletter, 24(2), 32-47.*

[11] *Singh, N. K., Necoara, I. (2024). Unified analysis of stochastic gradient projection methods for convex optimization with functional constraints. European Control Conference, 3600-3605.*

[12] *Ji, Y., Zhou, Z., Liu, H., et al. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112-2120.*