MiniProject 3: Classification of Textual Data

Jiahang Wang Li Jiang Liting Chen

***Abstract*—The ongoing advancement in Natural Language Processing (NLP) has opened new frontiers in text classification. This project compares traditional machine learning and state-of-the-art deep learning techniques—Naive Bayes and BERT—on the emotion classification task using the 'Emotion' dataset from Hugging Face. Our findings reveal the superior capability of BERT models in capturing the nuances of emotional expressions within text, a feat less pronounced in traditional methods. Through a series of experiments, including attention analysis and activation pattern examination, we dissect the underlying reasons for BERT's effectiveness, emphasizing the role of pretraining in performance enhancement.**

***Index Terms*—Emotion Classification, Text Classification, Naive Bayes, BERT**

I. Introduction

The ascension of deep learning in Natural Language Processing (NLP) has profoundly transformed how machines understand human language. Among various NLP tasks, emotion classification in textual data stands out due to its complexity and practical importance. Traditional machine learning methods, such as Naive Bayes, have laid the groundwork, offering interpretable models but often lacking the depth to grasp language nuances. On the other hand, the advent of transformer-based models like BERT has marked a leap forward. With their inherent capacity to capture contextualized word meanings, these models have set new standards for text classification. This report delves into a comparative analysis of these methodologies, employing the 'Emotion' dataset from Hugging Face to benchmark their performances. We conduct a series of experiments to investigate the impact of BERT's pretraining on an external corpus and its subsequent fine-tuning on model performance. The analysis includes attention visualization to understand the model's focus points and layer-wise activation exploration to uncover the representation learning at different stages. Through this comparative analysis, we aim to deduce insights into the advanced language processing capabilities of deep learning over traditional methods and the role of pretraining in enhancing model performance for complex classification tasks like emotion detection.

II. Related Work

The classification of textual data remains a pivotal and dynamic area in the domain of machine learning and NLP. Advancements in this field are driven by an increasing volume of complex texts that require sophisticated techniques to parse and interpret. The foundation of text classification systems is largely built upon a pipeline consisting of feature extraction, dimensionality reduction, classifier selection, and evaluation, as described in [3].

Feature extraction is the initial step, where techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe transform unstructured text into structured feature space [4]. Following feature extraction, the choice of classification algorithm is crucial. Traditional methods, ensemble-based techniques, and simple algorithms like logistic regression and Naive Bayes have been extensively used and studied for their efficiency and low memory requirements [5]. BERT-based methods have revolutionized the field of text classification due to their deep understanding of language context and nuance [6]. Pre-trained on vast amounts of text data, BERT's bidirectional training structure allows it to capture intricate patterns in language that were previously inaccessible to machine learning models [1]. In text classification tasks, BERT has been fine-tuned to achieve state-of-the-art results by further training on domain-specific data, optionally incorporating multi-task learning, and adjusting to the specifics of the target task. Techniques to enhance its performance include handling long texts, selecting appropriate layers, adjusting learning rates, and mitigating problems like catastrophic forgetting and low-shot learning [7]. BERT takes sequences of tokens and outputs representations that are then used for classification tasks, with the final hidden state of the [CLS] token serving as the aggregate sequence representation. This has led to BERT-based models achieving new benchmarks across various text classification datasets [2].

III. System Models

A. Data Preprocessing

Naive Bayes: In implementing the Naive Bayes classifier, our preprocessing pipeline plays a crucial role. The raw text data from the dataset is first transformed into a structured numerical format. This transformation is achieved using the 'Bag of Words' representation, facilitated by the CountVectorizer function from the scikit-learn library. This method converts the text documents into a matrix of token counts, effectively capturing the frequency of each word in the documents while disregarding the order of words. This representation is particularly suited for the Naive Bayes algorithm, which relies on the frequency of features (in this case, words) to make predictions.

BERT: For the implementation using BERT, a different approach to preprocessing is required. We utilize the transformers package from Hugging Face, particularly designed for tasks involving BERT. This package provides functionalities to tokenize the input text, converting them into a sequence of tokens that can be processed by the BERT model. These tokens are then transformed into numerical features that represent the input text in a format understandable by BERT. This process is crucial for capturing the contextual relationships between words in the text, which is a key strength of the BERT model.

B. Naive Bayes Model

We develop a custom Python class from scratch, leveraging the numpy library for numerical computations. The class constructor initializes model parameters and sets up essential properties. The fit function is designed to train the model using training data and hyperparameters, adjusting model parameters accordingly. Prediction is handled by a predict function that estimates labels for given input features. We also include an evaluate_acc function within the class to calculate the accuracy of the model by comparing predicted labels against true labels.

C. Bert-based Model

For the BERT-based model, we utilize pre-trained weights available from Hugging Face. This allows us to leverage the powerful language representations BERT has learned. We experiment with fine-tuning strategies, including adjusting the entire model or only the last few layers, to enhance performance on the emotion prediction task. By fine-tuning on our specific dataset, we tailor the pre-trained model to better recognize the nuances in emotional expression within our text corpus.

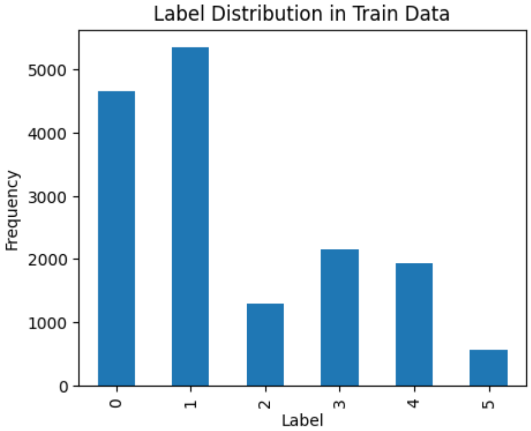

Fig. 1. Label Distribution in Train Data

IV. Experiments

In this section, we present a comprehensive suite of experiments designed to evaluate and compare the performance of traditional and deep learning NLP models on the task of emotion classification. Using the ’Emotion’ dataset, we first describe the dataset and then provide a detailed analysis of the performance of a Naive Bayes classifier constructed from scratch and a series of BERT-based models with varying degrees of fine-tuning. Our attention analysis within BERT sheds light on the model’s focus points and decision-making process. Lastly, we examine the layer-wise activations of BERT to understand the evolution of text representation within the model, offering insights into the advantages of pretraining on large language corpora.

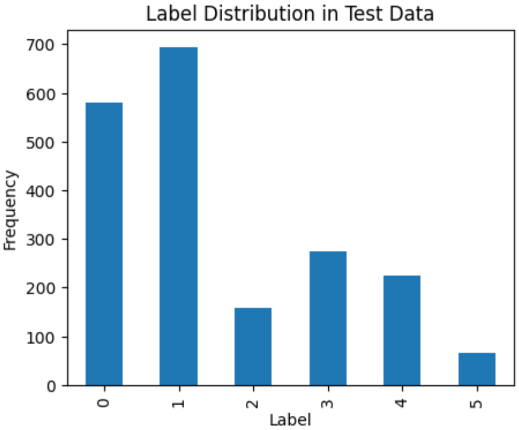

Fig. 2. Label Distribution in Test Data

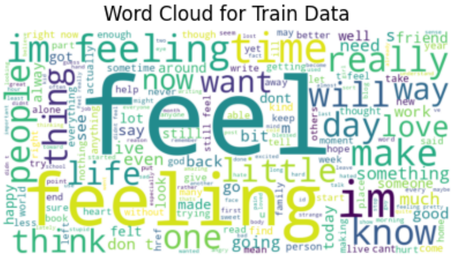

Fig. 3. Word Cloud for Train Data

A. Dataset

The dataset employed in this project is the ’Emotion’ dataset hosted on Hugging Face’s datasets repository, specifically designed for the task of emotion recognition in textual data. It comprises various text snippets, each annotated with an emotion label that categorizes the sentiment of the text. This rich dataset is instrumental for training machine learning models to understand and predict emotions conveyed in written language.

The exploratory data analysis visualizations are presented in Figures 1 to 4. Figure 1 depicts the label distribution of the training dataset, where labels correspond to different emotions. It’s evident that there is an imbalance in the label frequency, with Labels 0 (joy) and 1 (sadness) being the most common.

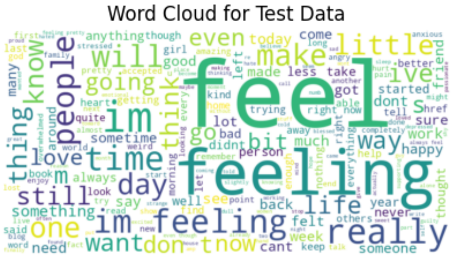

Fig. 4. World Cloud for Test Data

This suggests that certain emotions are more prevalent in the dataset, a factor that may influence the model's learning process. Figure 2 shows the label distribution for the test dataset, which follows a similar pattern to the training set, albeit with smaller quantities, as is typical for test data.

Figures 3 and 4 offer word clouds for the training and test data, respectively. These word clouds graphically represent the frequency of words, with the more predominant terms appearing larger. Common words like ”feel”, ”really”, and ”time” are emphasized, signaling their importance in the texts' emotional expressions. The word clouds not only provide a visual summary of the most recurrent terms but also highlight the challenge of distinguishing between different emotions based on commonly used words. This underlines the necessity for a nuanced approach to feature extraction in the subsequent stages of model development.

B. Naive Bayes Performance

The Naive Bayes classifier's performance was evaluated using several metrics, which are presented in Table I. The model achieved an accuracy of 76.9%, indicating a relatively high level of correct predictions across the multi-class emotion classification task. However, the precision and recall scores were 67.2% and 54.8%, respectively, suggesting that while the model is generally reliable when it predicts an emotion, it does not capture all instances of each emotion effectively. The F1 score, which balances precision and recall, was consequently moderate at 57.1%. These results highlight the Naive Bayes model's limitations in dealing with complex patterns within the data.

TABLE I
PERFORMANCE METRICS OF THE NAIVE BAYES CLASSIFIER ON THE EMOTION CLASSIFICATION TASK.

Metric	**Value**
Accuracy	0.769
Precision	0.672197
Recall	0.548084
F1 Score	0.570575

C. Accuracies of Different Models

In this experiment, we evaluated the emotion classification capabilities of two types of models: the aforementined Naive Bayes classifier and various configurations of a BERT-based model with pre-trained weights. The BERT models, sourced from pre-existing pre-trained weights, were fine-tuned to varying extents for the task at hand. The BERT-based models were fine-tuned in four different settings: without any fine-tuning, fine-tuning only the last layer, fine-tuning the last five layers, and finetuning all layers. The models' performance was measured by their accuracy in classifying the correct emotion from the given text data and is reported in Table II.

The BERT model fine-tuned on all layers emerges as the clear winner in the Emotion classification task. The potential reasons for its superior performance include the additional fine-tuning allowing the model to better adapt the pre-trained weights to the specific context and nuances of the emotion dataset. In contrast, the Naive Bayes model, while reasonably effective, lacks the sophisticated contextual understanding that BERT provides. The BERT model without fine-tuning performs poorly, likely due to a misalignment between the pre-trained model's understanding and the specific requirements of the emotion classification task and the difference between the pre-training data set and the evaluation data set we use.

TABLE II
COMPARISON OF MODEL PERFORMANCES ON THE EMOTION CLASSIFICATION TASK.

Model	**Accuracy (%)**
Naive Bayes	76.90
BERT (without tuning)	29.05
BERT (tune first layer)	34.75
BERT (tune last layer)	91.05
BERT (tune all layers)	**93.15**

The results suggest that pretraining on an external corpus, as BERT does, is generally beneficial for the Emotion prediction task. Pretraining allows the model to learn a rich representation of language, capturing contextual information and nuances that are not immediately evident from the training data alone. This foundational knowledge is then adapted through fine-tuning to the specific task, allowing for more nuanced understanding and prediction of emotions in text.

Comparing the performance differences between deep learning methods, like BERT, and traditional machine learning methods, such as Naive Bayes, we see that deep learning methods, which can integrate and learn from a vast array of syntactic and semantic relationships in the data, outperform traditional methods. The traditional methods, while effective to a degree, lack the capacity to model the complex, non-linear patterns that deep learning architectures can capture. This is evident from the significantly higher accuracy scores achieved by the fine-tuned BERT models in the experiment.

D. Attention Analysis in BERT Model

In this experiment, we aimed to understand the role of attention mechanisms within the BERT model when predicting emotions in text. We visualized the attention weights from one of the multi-headed transformer blocks of BERT, focusing on a specific attention head to assess its behavior on both correctly and incorrectly predicted documents.

As shown in Figures 5 and 6, the heatmaps generated reflect the attention distribution across tokens within the documents. The x-axis and y-axis represent the sequence positions, and the color intensity reflects the attention from one token to another, with brighter colors indicating higher attention. We observed a notable concentration of attention towards the beginning of the sequences. This implies that initial tokens might play a significant role in influencing the model's predictions. Moreover, the vertical lines in the heatmaps indicate that certain tokens are focal points, receiving attention from multiple other tokens, hinting at their contextual importance. Correctly predicted documents show focused attention patterns, suggesting the

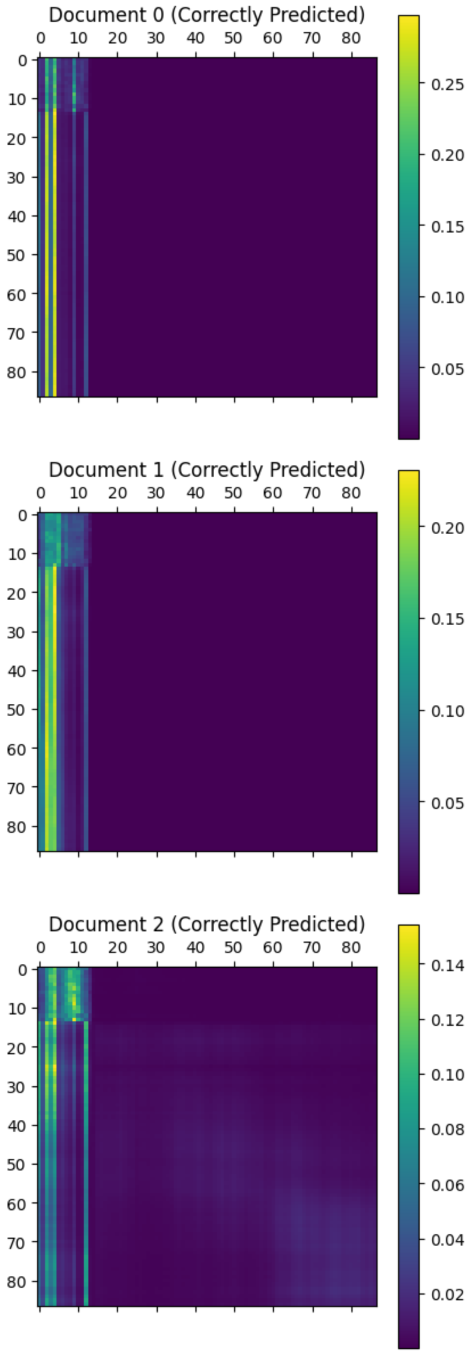

Fig. 5. Attention matrix for Correctly Predicted Instances

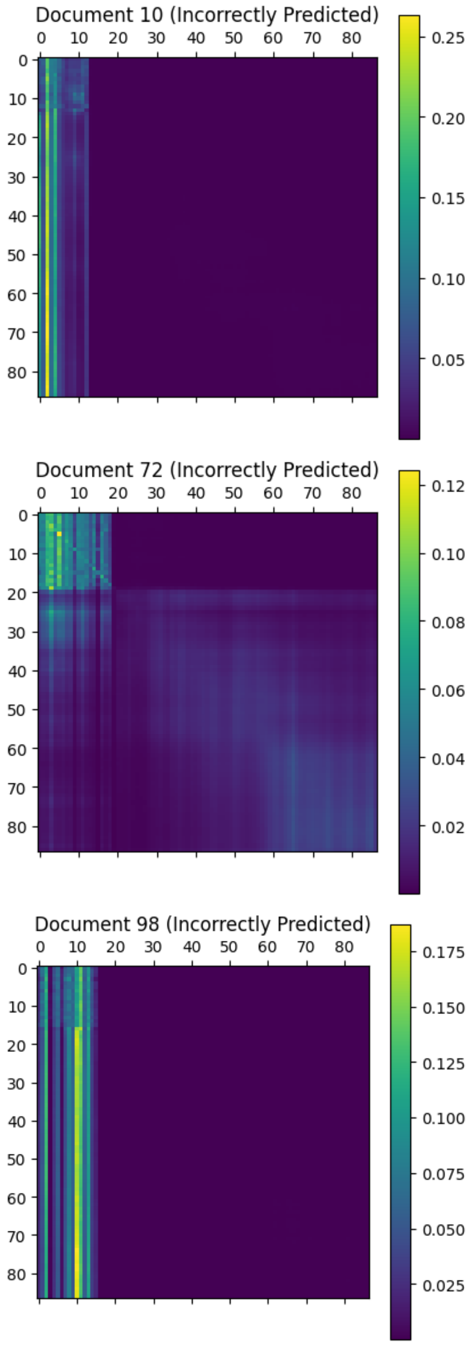

Fig. 6. Attention matrix for Incorrectly Predicted Instances

model is keying in on specific words that are strong indicators of the emotional content. Incorrect predictions display more diffused attention, indicating potential confusion or lack of strong features for the model to latch onto. The discrepancy could also stem from the inherent complexity of certain emotions or from subtleties in language that the model has not effectively learned to interpret.

Through these observations, we infer that the pre-training on extensive external corpora allows BERT to identify and focus on salient features within a text, which are crucial for emotion detection. This capability likely contributes to the model's performance and underscores the advantage of pre-trained models in capturing the essence of the text for nuanced tasks such as emotion prediction.

E. Analysis of BERT Layer Activations

This experiment investigated the activations within the [CLS] token's hidden state across various layers of a BERT model. By examining the activation patterns from the first, sixth, and twelfth layers, as shown in Fig. 7, we aimed to understand how the representations evolved and contributed to emotion prediction.

Layer 1 [CLS] token hidden state: The activations are relatively flat, indicating that at this early stage, the model's representations are likely raw and less abstract. This suggests that pretraining helps the model to start forming basic representations of the input data, which are further refined in subsequent layers.

Layer 6 [CLS] token hidden state: The plot shows more variation in activation patterns compared to Layer 1, which implies that by this middle layer, the model is beginning to develop more complex features. Pretraining has likely helped the model to learn patterns in language that are useful for interpreting emotional content, even if not directly trained on emotion-specific data.

Layer 12 [CLS] token hidden state: There is a high degree of variability in the activations, indicating complex interactions and a sophisticated level of abstraction. This complexity is crucial for tasks like emotion prediction because the model needs to synthesize nuanced features from the input text to understand and predict emotions accurately. Pretraining on a large corpus allows BERT to capture a wide range of language nuances, which is likely beneficial for identifying subtle emotional cues in the text.

The evolution of activation patterns across layers suggests that pre-trained BERT with the ability to process and synthesize complex features from text, which is likely to be beneficial for accurately predicting emotions.

V. Conclusion

The comparative analysis highlights the stark contrast between traditional machine learning and advanced deep learning techniques. Naive Bayes, although a solid baseline, is overshadowed by BERT's nuanced understanding of language, evidenced by its attention mechanism and layer activations. Pretraining on extensive corpora proves to be a significant factor in BERT's success, allowing it to outperform in tasks demanding deep semantic comprehension like emotion classification. This study not only underscores the strengths of pretraining and fine-tuning in transformer-based models but also points to the continual relevance of traditional methods in specific contexts.

In this project, Jiahang Wang and Li Jiang primarily concentrated on the coding aspects, while Liting Chen was chiefly responsible for drafting this report.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

[3] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

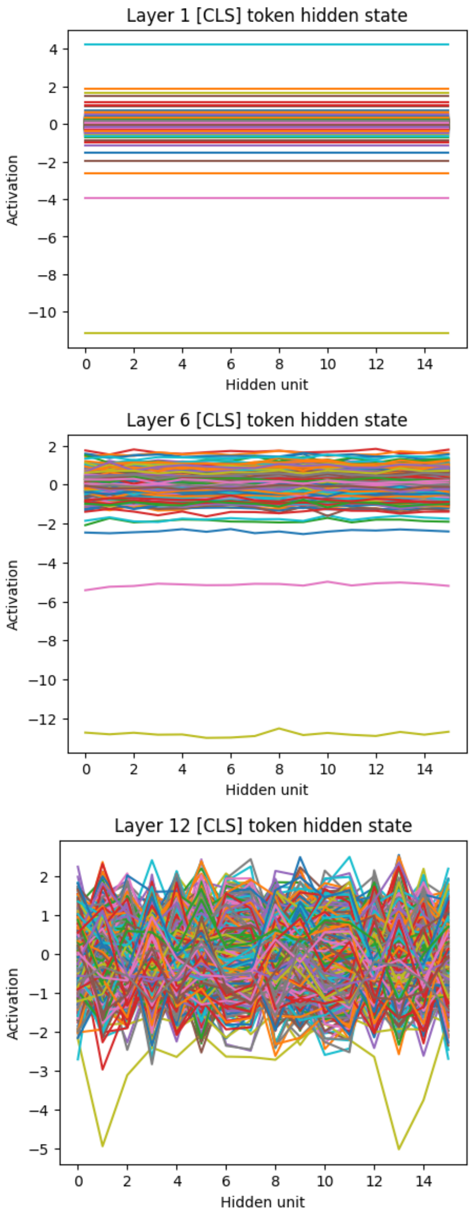

Fig. 7. BERT Layer Activations

[4] Mara Pistellato, Filippo Bergamasco, Andrea Albarelli, Luca Cosmo, Andrea Gasparetto, and Andrea Torsello. Robust phase unwrapping by probabilistic consensus. *Optics and Lasers in Engineering*, 121:428–440, 2019.

[5] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[6] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer, 2019.

[7] Chao Wang, Yi Hou, and Matthew Barth. Data-driven multi-step demand prediction for ride-hailing services using convolutional neural network. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 2 1*, pages 11–22. Springer International Publishing, 2020.