Emotion Classification using TF-IDF + SVM and BERT Models

Selina Meng

MSDS 453 NLP

June 1st, 2025

1. Introduction and Problem Statement

Emotion classification has significant applications in customer service automation, sentiment analysis for market research, and real-time monitoring of mental health signals in social media. For this project, I focused on using two powerful NLP methodologies: traditional feature engineering with Support Vector Machines (SVM) and advanced transformer-based models like BERT, to classify emotions in text data. The dataset includes a collection of English sentences labeled with one of six emotion categories, ranging from 0 to 5: sadness, joy, love, anger, fear, and surprise. My primary research question was how the performance of a traditional TF-IDF + SVM pipeline compares to a fine-tuned BERT model on emotion classification.

The objective was to evaluate both models on the same dataset, which was split into training, validation, and test, comparing their precision, recall, and F1-scores to analyze strengths and limitations. The key business context for this investigation is the growing need for accurate emotion classification to guide decisions, from analyzing product feedback to improving customer support and making it faster and more personalized.

2. Research Design and Modeling Method

I began with data cleaning and preprocessing. This included removing URLs, mentions, punctuation, and stop words to ensure high-quality input for the traditional model. For the BERT model, the original text was kept unchanged because its pretrained tokenizer can handle raw text input directly. Additionally, I applied label encoding to convert the emotion labels into numerical format compatible with BERT.

For the TF-IDF + SVM, I tokenized the cleaned text using a TfidfVectorizer with a maximum of 5,000 features. This ensured a balance between computational efficiency and vocabulary richness. The data was split into three subsets: training (80%), validation (10%),

and test (10%) to prevent information leakage. I trained the SVM on the training data and evaluated it on the validation and test sets to measure generalization.

For the BERT model, I utilized the pretrained bert-base-uncased transformer from Hugging Face. The model was fine-tuned for 3 epochs using the AdamW optimizer and a learning rate of 0.00005. Data was tokenized with a maximum sequence length of 128 and attention masks, which help BERT focus on relevant tokens. The loss function was cross-entropy loss, appropriate for multi-class classification. I also implemented a DataLoader to efficiently handle mini-batch training and random shuffling. Unlike SVM, which operate on the entire dataset in matrix form, BERT requires feeding data in small batches for gradient-based optimization.

The reason for comparing TF-IDF + SVM with BERT lies in their fundamental differences: TF-IDF + SVM uses sparse bag-of-words features, whereas BERT learns dense, contextualized embeddings. This contrast helps assess the effectiveness of transformer-based representations in capturing emotional differences compared to traditional frequency-based features. As mentioned by Prabhu (2024), while SVMs have historically been used with simple feature extraction methods, integrating them with large language models like BERT can reveal interesting performance contrasts and demonstrate the power of contextual embeddings over sparse representations.

In addition to model training, I evaluated each approach using classification reports and confusion matrices. These evaluation metrics provided detailed insights into precision, recall, and F1-scores across all emotion classes. Confusion matrices helped visualize misclassifications, which I used to support my deeper analysis.

3. Results

TF-IDF + SVM Model

The SVM model reached an accuracy of 87% on the test set (Figure 2), with F1-score of 0.81. Although it showed strong performance for the majority classes, it struggled significantly with the minority classes, such as love and surprise, which were the least represented emotions in the dataset at 8.2% and 3.6%, respectively. In the report (see Table 1 for emotion label mapping), label 2's F1-score dropped to 0.70, and label 5's F1-score was only 0.63. Similarly, on the validation set (Figure 1), it achieved an overall accuracy of 87% and an F1-score of 0.84.

The confusion matrices (Figures 3, 4) for SVM revealed more misclassifications than BERT, especially for minority classes. This aligns with the inherent limitations of sparse feature representations in TF-IDF, which lack the contextual nuance needed for sophisticated emotional distinctions.

BERT Model

On the validation set (Figure 6), BERT achieved an overall accuracy of 93%, with an F1-score of 0.91. The precision, recall, and F1-score across classes were more balanced, although label 5 still showed a slightly lower F1-score of 0.85.

On the test set (Figure 7), BERT maintained an accuracy of 93%, with an F1-score of 0.89. It consistently delivered high precision for labels 0 and 1, with slight dips for label 2 (precision 0.76) and label 5 (precision 0.70). Confusion matrices (Figures 8, 9) confirmed that BERT's misclassifications primarily occurred between more similar emotional categories, like it will fear and surprise, as well as joy and love.

The training loss plot (Figure 5) for BERT over three epochs indicated rapid convergence, dropping from 0.4 to below 0.15, highlighting the model's ability to learn efficiently even with modest epochs.

4. Analysis and Interpretation

The results clearly demonstrate the superior performance of the BERT model compared to the TF-IDF + SVM in this emotion classification task. While both models achieved accuracy above 85%, BERT consistently outperformed SVM across all classes. BERT's F1-score on the test set reached 0.89, much higher than SVM's 0.81, highlighting BERT's ability to maintain balanced performance even for underrepresented emotions.

The confusion matrices provide insightful patterns that BERT was more adept at distinguishing similar emotions like joy and love, whereas SVM often confused those classes. This suggests that transformer-based contextual embeddings better capture subtle emotional nuances and semantic richness in language.

Both models showed confusion between labels 4 (fear) and 5 (surprise). These are inherently similar emotional states, and this confusion pattern suggests that even advanced models like BERT require richer contextual data or specialized augmentation techniques to distinguish such subtle differences.

A notable performance difference is on label 2 (love), which represents only 8.2% of the dataset. BERT maintained a high recall of 0.96 for this label, while SVM's recall dropped to 0.59. This underscores BERT's robust handling of minority classes, likely because it uses pretrained language representations that already understand many different language patterns from a huge collection of text.

Another important observation comes from the training loss plot (Figure 5). BERT's rapid convergence over three epochs, which decreases from approximately 0.4 to below 0.15, demonstrates the power of transfer learning. As noted by Sharma (2020), transfer learning in NLP leverages pretrained language models like BERT, which already encode rich language understanding from massive corpora. This significantly reduces the computational burden

and data requirements for fine-tuning, making BERT highly practical for real-world applications.

This comparative study shows that while traditional methods like TF-IDF + SVM are still useful for smaller-scale applications because of their simplicity and speed, transformer-based models like BERT have clear advantages in terms of precision and recall, especially for capturing subtle emotional nuances and handling rare classes.

5. Conclusions

The project compared a traditional TF-IDF + SVM pipeline with a deep learning-based BERT model for emotion classification, using the Twitter emotion dataset. The results consistently demonstrated BERT's superior performance across all key metrics, with an F1-score improvement of about 8 percentage points over SVM on the test set. This confirms the significant advantages of transformer-based models in capturing emotional nuances that traditional feature-based approaches struggle to model.

One of the most impressive findings was how BERT managed to accurately classify less common emotions like love and surprise, even with their low dataset representation. This shows that the contextual information BERT learns from pretraining helps it pick up on these subtle emotional cues, which is important for things like customer sentiment tracking and mental health analysis.

The confusion matrices and classification reports highlight that while SVM provides a quick and relatively accurate baseline, it lacks the contextual depth to handle nuanced emotional distinctions, leading to misclassifications in low-frequency emotions. BERT, in contrast, demonstrated consistent diagonal patterns in its confusion matrices, indicating a more precise prediction.

In this project, I also incorporated the use of ChatGPT as a generative AI tool to assist with code development. It streamlined the coding process by helping troubleshoot errors and suggest alternative approaches, showcasing how AI can enhance technical development.

From a business perspective, these findings suggest that investing in transformer-based models like BERT offers higher accuracy and robustness, especially for tasks where nuanced emotional insights are essential, such as targeted marketing or social media monitoring. This project highlights how modern NLP approaches like BERT move beyond simple frequency-based methods, enabling a truly contextualized understanding of language.

References

dair-ai. (n.d.). Emotion dataset. Hugging Face. https://huggingface.co/datasets/dair-ai/emotion

OpenAI. (2024). GPT-4 (Mar 2024 version) [Large language model]. https://chat.openai.com/

Prabhu, S. (2024, March 20). Leveraging support vector machines (SVM) in large language models (LLMs). Medium. https://medium.com/@prabhuss73/leveraging-support-vector-machines-svm-in-large-language-models-llms-5d8ee165bf54

Sharma, A. (2020, July 24). Transfer learning for NLP: Fine-tuning BERT for text classification. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/

Appendices

Validation Classification Report:

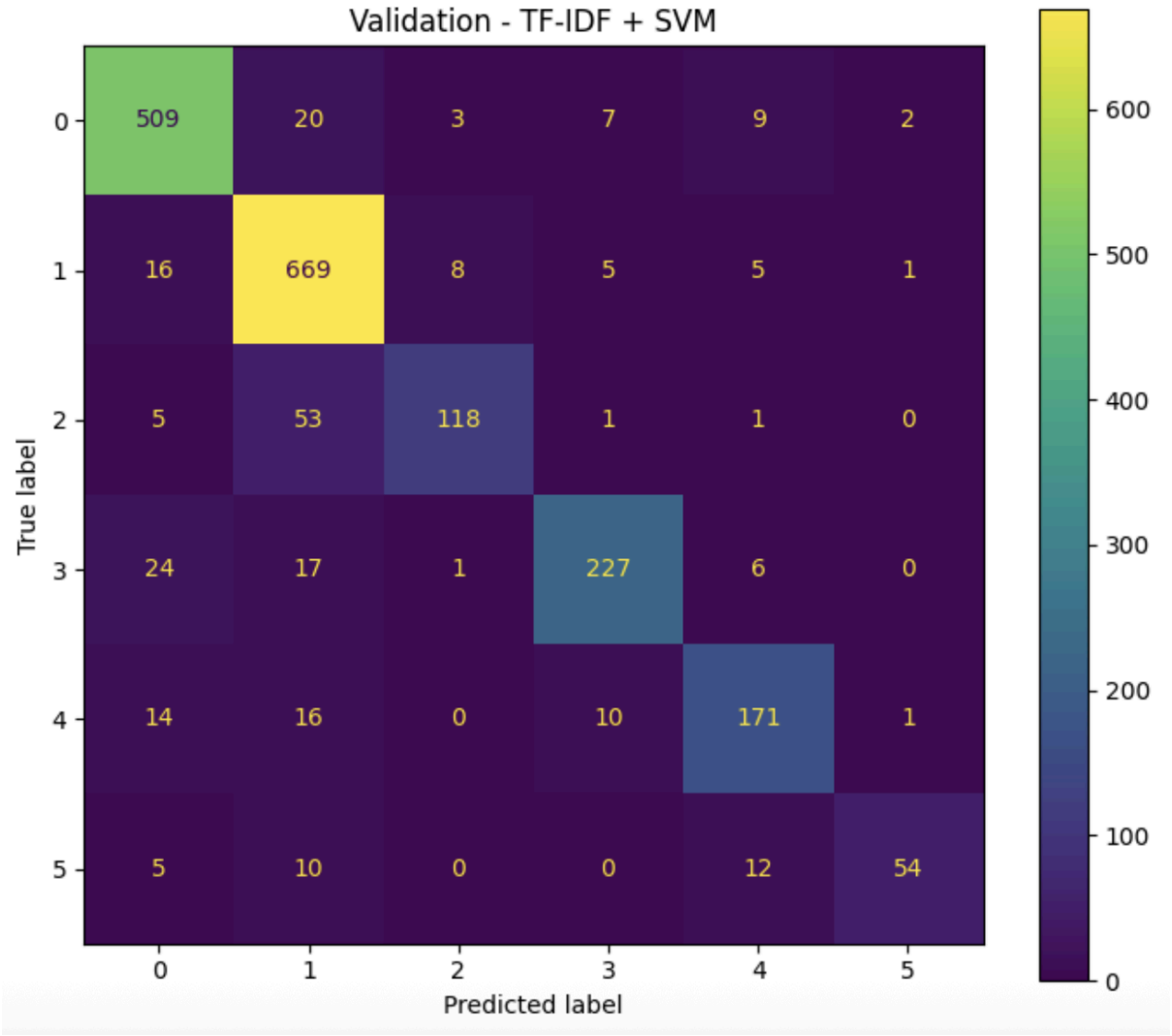
	precision	recall	f1-score	support
0	0.89	0.93	0.91	550
1	0.85	0.95	0.90	704
2	0.91	0.66	0.77	178
3	0.91	0.83	0.86	275
4	0.84	0.81	0.82	212
5	0.93	0.67	0.78	81
accuracy			0.87	2000
macro avg	0.89	0.81	0.84	2000
weighted avg	0.88	0.87	0.87	2000

➔ Figure 1 (SVM)

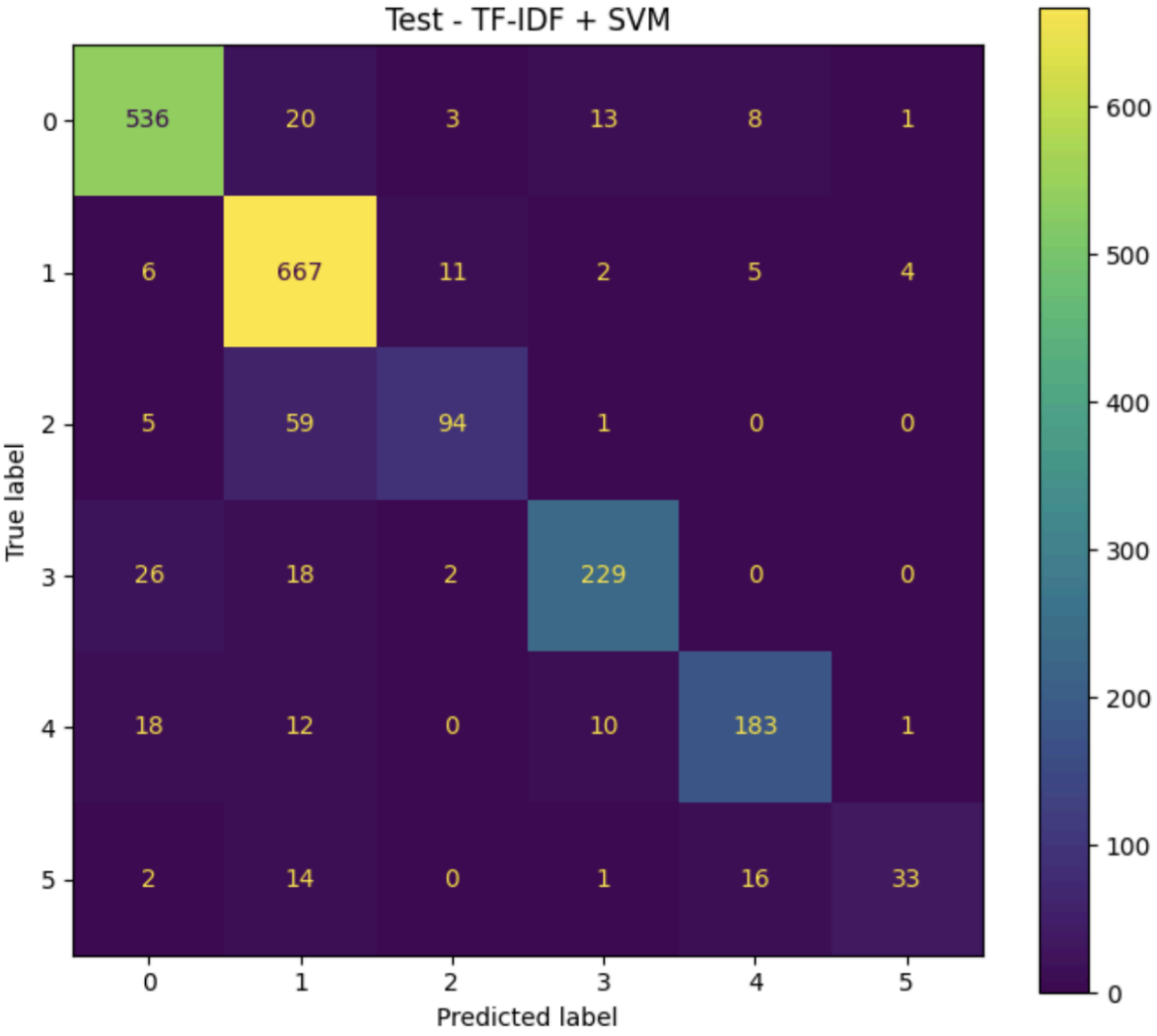
Test Classification Report:

	precision	recall	f1-score	support
0	0.90	0.92	0.91	581
1	0.84	0.96	0.90	695
2	0.85	0.59	0.70	159
3	0.89	0.83	0.86	275
4	0.86	0.82	0.84	224
5	0.85	0.50	0.63	66
accuracy			0.87	2000
macro avg	0.87	0.77	0.81	2000
weighted avg	0.87	0.87	0.87	2000

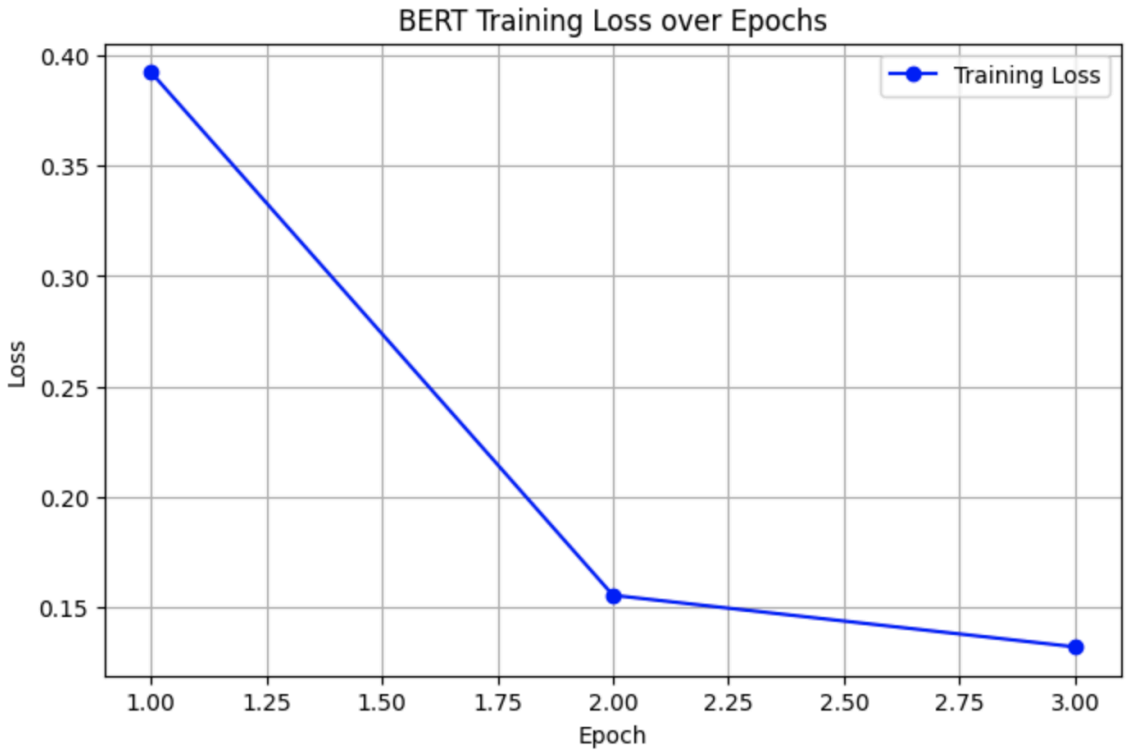
➔ Figure 2 (SVM)


Validation - TF-IDF + SVM
True label
Predicted label
0
1
2
3
4
5
509 20 3 7 9 2
16 669 8 5 5 1
5 53 118 1 1 0
24 17 1 227 6 0
14 16 0 10 171 1
5 10 0 0 12 54
600
500
400
300
200
100
0

➔ Figure 3


Test - TF-IDF + SVM
True label
Predicted label
0
1
2
3
4
5
536 20 3 13 8 1
6 667 11 2 5 4
5 59 94 1 0 0
26 18 2 229 0 0
18 12 0 10 183 1
2 14 0 1 16 33
600
500
400
300
200
100
0

➔ Figure 4



➔ Figure 5

	precision	recall	f1-score	support
0	0.96	0.96	0.96	550
1	0.97	0.93	0.95	704
2	0.81	0.98	0.89	178
3	0.91	0.93	0.92	275
4	0.92	0.83	0.87	212
5	0.83	0.88	0.85	81
accuracy			0.93	2000
macro avg	0.90	0.92	0.91	2000
weighted avg	0.93	0.93	0.93	2000

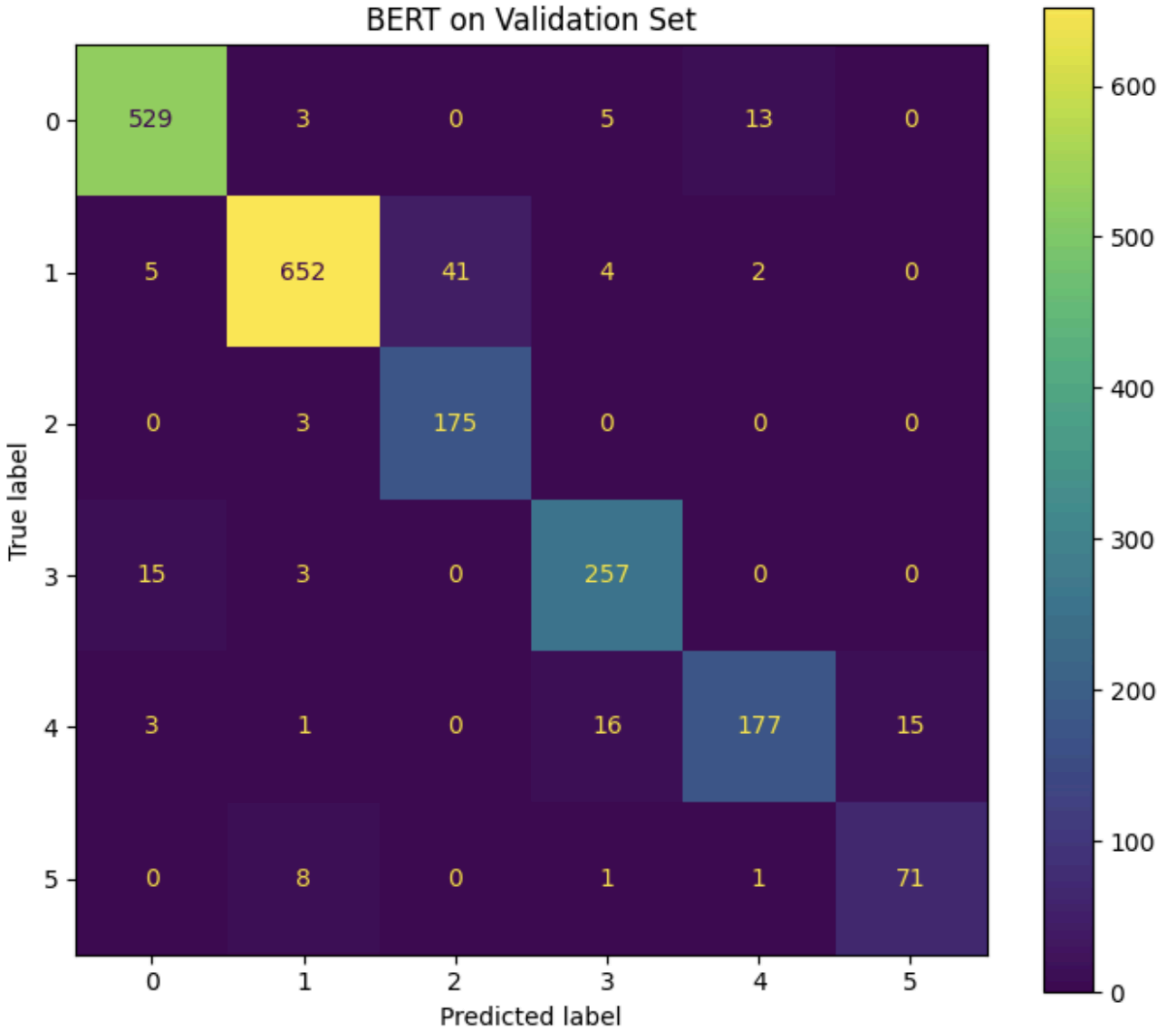
Validation Classification Report:

➔ Figure 6 (Bert)

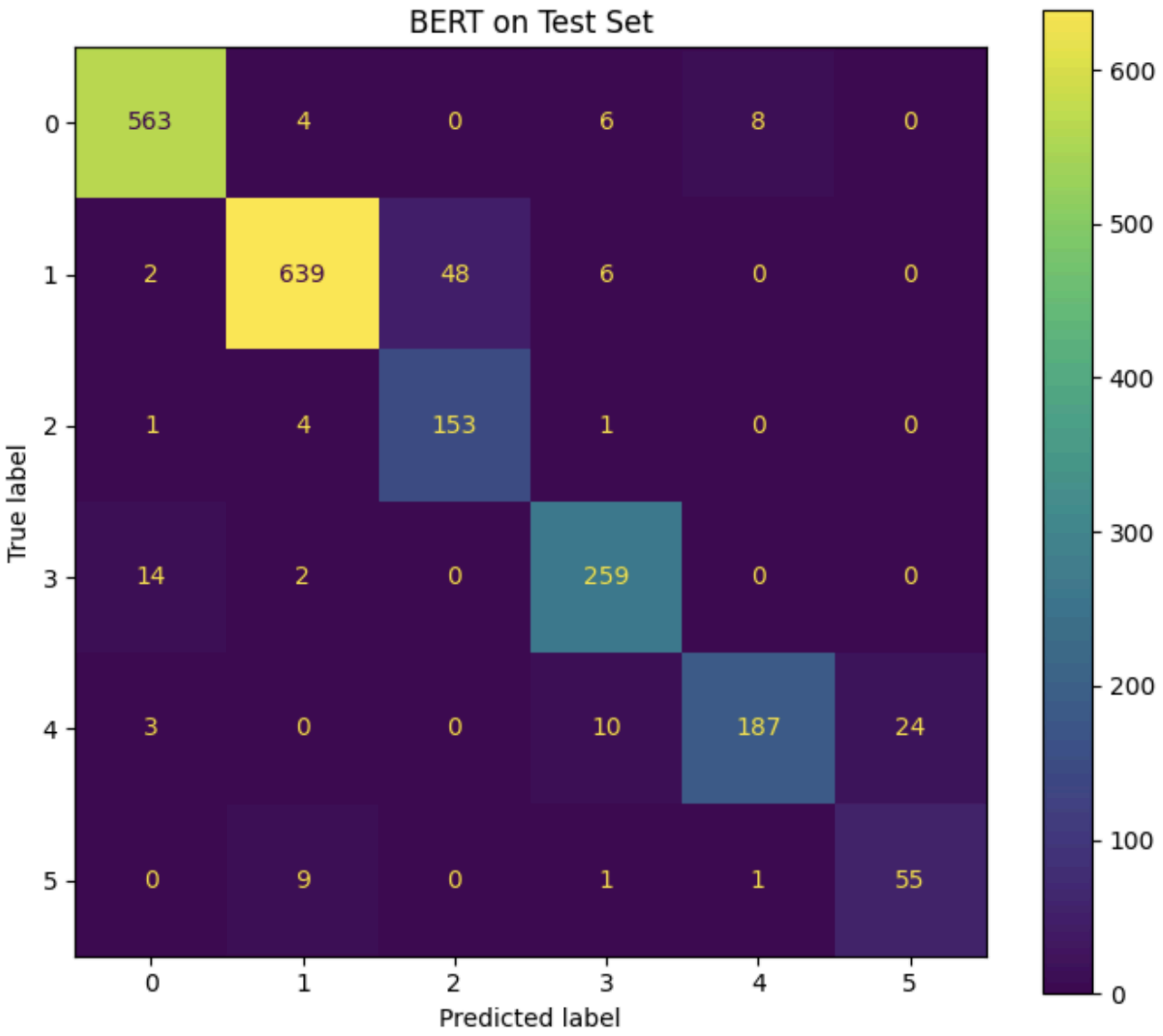
Test Classification Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	581
1	0.97	0.92	0.94	695
2	0.76	0.96	0.85	159
3	0.92	0.94	0.93	275
4	0.95	0.83	0.89	224
5	0.70	0.83	0.76	66
accuracy			0.93	2000
macro avg	0.88	0.91	0.89	2000
weighted avg	0.93	0.93	0.93	2000

➔ Figure 7 (Bert)



➔ Figure 8



➔ Figure 9

Label	Emotion
0	sadness
1	joy
2	love
3	anger
4	fear
5	surprise

➔ Table 1