Applied Machine Learning: Mini-Project 3

JAN TIEGGES, RISHABH THANEY, JONATHAN COLAÇO CARR

ABSTRACT: This report provides a comparative analysis of Naive Bayes and BERT models for emotion classification from text data. Using the Emotion dataset, our study contrasts the performance of a Naive Bayes model implemented from scratch against a pre-trained BERT model, with and without fine-tuning. The Naive Bayes model records only an 83.85% accuracy on the test set, which we attribute in part to class imbalance in the Emotion dataset. In contrast, our best BERT model acheives an accuracy of 92.65%. We supplement these results with a discussion of the Naive Bayes' learned features and an analysis of the attention matrices in the BERT models. Overall, our findings are consistent with the theory of sequence models and the Naive Bayes algorithm presented in class.

1. Introduction

Understanding and interpreting emotions from textual data is a critical challenge in the fields of deep learning and natural language processing. The primary objective of this project was to gain practical experience in implementing and comparing two distinct models for emotion classification: the Naive Bayes and BERT (Bidirectional Encoder Representations from Transformers) [DCLT19] models. We compared these models on the Emotion Dataset [SLH+18], which has been widely used to train emotion classification models [GA, QS23].

Our experiments offered both a comparative analysis of the Naive Bayes and BERT models as well as insights into the learned features for both models. We compared the Naive Bayes with three different BERT models: one "base" model and two models which were fine-tuned on the Emotion dataset. Our best BERT model significantly outperformed the Naive Bayes model (83.83% vs. 92.65% test accuracy) which is consistent with the power of Transformer-based models discussed in class [PSb]. Our best BERT model is not quite as accurate as other known models[1] which were trained on a much larger portion of the Emotion dataset. Our fine-tuning did not improve the performance of the base model, which is perhaps unsurprising given that the base model was already fine-tuned on an emotion classification dataset. To interpret the Naive Bayes model, we derived the most important features for each class using log probabilities. For the BERT model, we considered the attention matrices for various heads, layers, and test samples. While the attention heads are not as interpretable as the Naive Bayes features, they are comparable to known patterns [CKLM19].

2. System Models

In this section we briefly review the basic elements of the Naive Bayes and BERT models.

Naive Bayes. The Naive Bayes model is a probabilistic classifier that uses Bayes' Theorem to estimate the probability of a label y given a feature vector $x \in \mathbb{R}^D$. The key assumption of the Naive Bayes model is that each feature is conditionally independent given the label. That is, the Naive Bayes model assumes

$$p(x|y) = \prod_{d=1}^{D} p(x_d|y). \tag{2.1}$$

Under this assumption, the Naive Bayes model tries to estimate the likelihood of each label given an input vector. In practice, it achieves this by counting the relative frequency of the features conditioned on each label.

Bert. The original BERT model [DCLT19] is a deep bidirectional Transformer encoder derived from the Transformer architecture [VSP+17]. BERT's model weights are pre-trained using vast amounts of unlabelled training data, including the BooksCorpus (800M words) [ZKZ+15] and English Wikipedia (2,500M words).

[1] https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion

A key component of both the BERT and Transformer models is a multi-headed, self attention mechanism, which computes the similarity between input query vectors and a set of keys. This mechanism is meant to give an improved representation of context [PSb].

We considered three adaptations of the original BERT model for our emotion classification task. First, we used Bhadresh Savani's Bert-based emotion classification model[2] as a base model, which we refer to as the BERT Out of the Box (OOB) model. We then fine-tuned the BERT OOB model using two strategies: one which fine-tuned all of the BERT OOB model parameters and one which fine-tuned only the parameters in the final layer.

3. Dataset

In this section we summarize the dataset considered. The `data_analysis.ipyb` notebook in our code folder contains a more in-depth analysis.

The Emotion Dataset. The Emotion Dataset [SLH+18] is a text classification dataset with 417 000 samples. Each sample consists of a single sentence labelled with one of six emotion labels (sadness, joy, love, anger, fear or surprise). These are six of the eight emotions described in Plutchik's Wheel of Emotions [Plu01]. While the original paper's data considered of all eight emotions, the Hugging Face dataset has only six of them. It is unclear why only six of the eight emotions appear in the Hugging Face dataset. Following the procedure in [Go09], Saravia et. al constructed the Emotion dataset by identifying candidate sentences on Twitter and using a machine learning algorithm to classify candidate sentences with emotions. Figure 1 shows the proportion of each emotion class in the dataset as well as some sample data points. Due to computational constraints, we only used 20 000 samples from the dataset. Of the 20,000 samples, we used 16,000 for training, 2,000 for validation and 2,000 for testing.

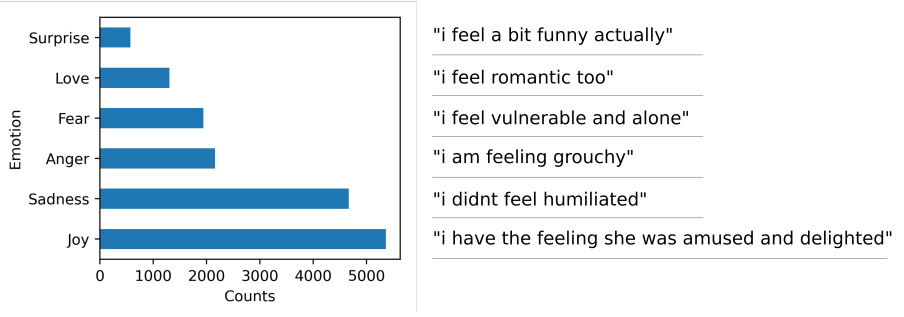

Figure 1. Proportion of each emotion label in the dataset (left) and a sample sentence from each emotion (right).

Ethical Concerns. This dataset poses various ethical concerns. First, it is unclear whether or not Twitter users gave informed consent for their tweets to be used for this dataset. Second, the emotion labelling is done by machine learning algorithms, rather than human labellers. Given the findings on the underperformance of NLP algorithms on non-native English speakers and alternative English vernaculars [BO17], it is likely that the labelled training data is worse for non-standard forms of English. To the best of our knowledge, the authors provided no interpretation or analysis of how the data was collected and made no comment on these risks.

Data Preprocessing. To preprocess the data for the Naive Bayes classifier, we used the CountVectorizer from scikit-learn [PVG+11]. We also experimented with removed the stop words and used bigrams, as we

[2] https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion

discuss in Section 4. For the BERT model, we used the pretrained BertTokenizer to convert the texts into input vectors.

4. Experiments

Our experiments were conducted in roughly three phases. In the first phase (4.1-4.3), we compared the performance of the Naive Bayes and BERT models by performing an extensive hyperparameter search and comparing the best variants of each model on the test set. In experiment 4.4, we interpreted the learned features of the Naive Bayes algorithm. Lastly in experiments 4.5 and 4.6 we investigated the attention matrices of our BERT models.

4.1 Naive Bayes Parameter Search. We examined various combinations of 4 parameters for the Naive Bayes Classifier, resulting in a total of 224 combinations. For the data preprocessing, we investigated the effect of using bigrams and removing stop words. For the model parameters, we considered the effect of smoothing in two ways discussed in class [PSa]: smoothing of the class distribution (α_{prior}) and smoothing of the occurrence of features per class ($\alpha_{likelihood}$). For the former we examined values between 0 and 100000 and for the smoothing of the likelihood values between 0 and 10. Note that the values for the smoothing of the class distribution are significantly higher, since there are hundreds of samples in each class, while the individual occurrences of features per class have significantly smaller numbers. Our experiments showed that the model with an α_{prior} of 100,000 and an $\alpha_{likelihood}$ of 0.2, as well as the removal of stopwords and the use of bigrams recorded the best accuracy on the validation set. It is interesting to note that the model appears to work best with a practically uniform prior (induced by the high smoothing parameter). The full results from the Naive Bayes hyperparameter search are available in our code.

4.2 BERT Fine-Tuning Parameter Search. Next, we performed a hyperparameter search to determine which parameters to use when fine-tuning the BERT model. We experimented with fine tuning all weights (FTA) as well as fine tuning only the weights in the last layer (FTL). For both the BERT FTA and BERT FTL models, we considered fine-tuning with different batch sizes (32, 64 and 128), learning rates (1×10^{-5}, 2×10^{-5} and 3×10^{-5}) and weight decays (0.01 and 0.001), resulting in 36 total combinations. These hyperparameter ranges were selected around the hyperparameters used to train our BERT OOB model[3]. Due to resource constraints, we only fine-tuned our model for 3 epochs. The results of the hyperparameter searches are shown in Tables A.2 and A.3. While the batch size and learning rate had little effect on the validation accuracy, one consistent trend for both the FTA and FTL models was that the smaller weight decay (i.e. less regularization) lead to slight improvements.

4.3 Final Comparison of all models. After our hyperparameter testing, we selected the best models (according to validation accuracy) and compared their performance on the test set. The results are shown in Table 1.

Metric	Naive Bayes	BERT FTA	BERT FTL	BERT OOB
Accuracy	0.8385	0.9250	0.9265	0.9265
Weighted F1 score	0.8343	0.9285	0.9265	0.9265

TABLE 1 *Model Performance Comparison on the Emotion Dataset.*

Both the BERT FTL and the BERT OOB have the highest accuracy of 92.65%, while the BERT FTA has the highest weighted F1 score of 92.85%. The differences between the various BERT models are almost

[3] https://github.com/bhadreshpsavani/ExploringSentimentalAnalysis/blob/main/SentimentalAnalysisWithDistilbert.ipynb

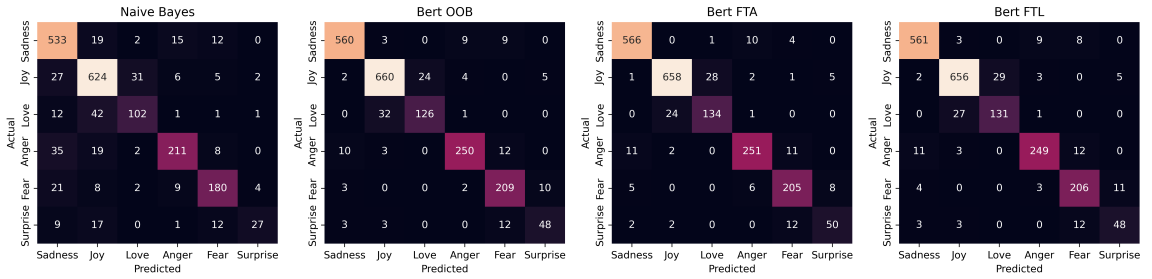

Figure 2. Confusion matrices for each of the models on the test set.

non-existent, which indicates that the fine-tuning had little effect on improving the BERT OOB model performance. All of the Bert models outperformed the Naive Bayes model, both in terms of accuracy and F1 score.

In addition to comparing the final accuracy and F1 scores for each model, we also investigated how their performance varied across the different emotion classes. As shown in Figure 2, all four of our models are prone to mistaking *joy* for *love*, *fear* for *surprise*, and *anger* for *sadness*. These gaps were anticipated after our data analysis, which revealed that many of the sample labels were ambiguous. For instance, in Figure 1 the sentence "I didn't feel humiliated" was labelled as a *sad* sentence.

4.4 Naive Bayes Interpretation. Next, we examined the behavior of the Naive Bayes Classifier in more detail. As shown in Table A.4, the Naive Bayes model performs significantly worse for the classes *surprise* and *love*, which are the classes with the smallest number of samples in the dataset. This observation is consistent with the theory presented in class, which highlighted the Naive Bayes' sensitivity to class imbalances, and the fact that it can be biased towards the majority classes [PSa].

To gain further insight on our Naive Bayes algorithm, we then considered the most important features per class, sorted by their log probabilities (available in our code). We noticed that words such as "feel", "feeling", "like" and "I'm" were identified as prominent features in several classes. This indicates that these words are often used to express feelings, but are not necessarily attributed to a particular emotion. However, for some emotion classes, the model was able to recognize key words. For example, the model associated the word "love" to the emotion class *love* and the words "amazed", "impressed", "overwhelmed", "strange" and "surprised" to the emotion class *surprise*.

4.5 Attention Matrices for Best BERT Model. Following our analysis of the Naive Bayes model, we considered the attention matrices of our best BERT model, the BERT OOB model. We selected correctly classified and incorrectly classified examples and plotted the attention matrices at various layers and heads in Figures B.1 and B.2. Figure 3 shows the attention matrices of the BERT OOB model for different samples at Layer 9, Head 7. This particular attention head shows a clear "focused" attention pattern, where tokens attend to specific keywords. For instance, for the Surprise (+) sample, the model attends to the word "shocked" very closely. However, attention patterns varied widely across different layers and heads, as shown in Figures B.1 and B.2. One other notable observation was that for many of the attention heads, the input tokens primarily attended to the `[SEP]` token. This pattern is well-documented in BERT models [CKLM19], and it is hypothesized that the pattern arises when the function of a specific attention head is not needed.

4.6 Comparison of Attention Matrices accross all BERT Models. In our final experiment, we considered how the attention matrices varied for each of the three BERT models. As shown in Figure 4, the changes between the attention heads of the three models were insignificant. This was expected, since the BERT models showed very similar overall performance on the test set.

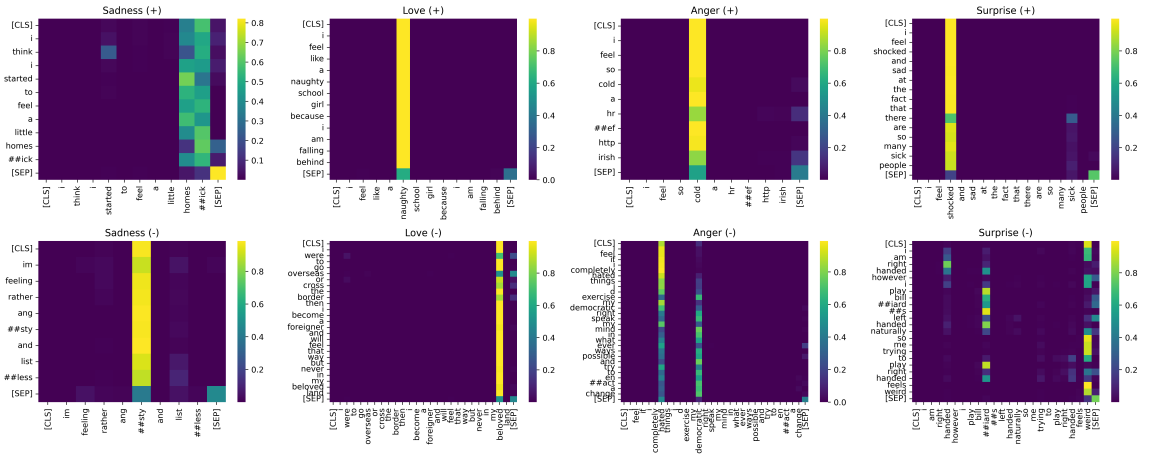

Figure 3. Bert OOB Attention Matrix (Head 7, Layer 9) for correctly classified (+) and incorrectly classified (-) samples.

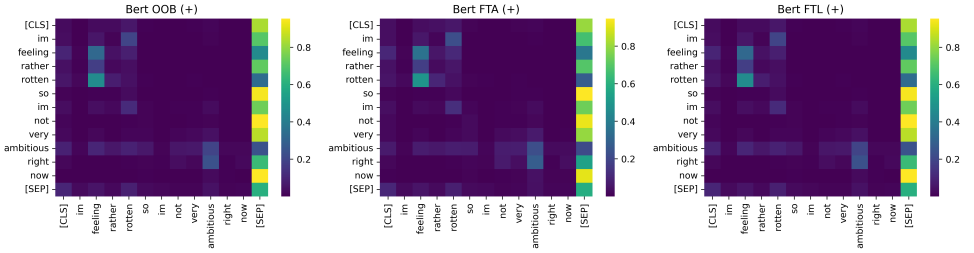

Figure 4. Attention Matrix (Layer 9, head 3) for the three Bert models.

5. Discussion and Conclusion

The results of our study provide interesting conclusions about the performance and functioning of Naive Bayes and BERT models for emotion classification. The fact that BERT generally outperforms Naive Bayes is in line with the theory discussed in class: BERT's Transformer-based architecture is inherently better suited to detect complex relationships between tokens, whereas the Naive Bayes assumes that each token is conditionally independent given the class labels. In particular, it is likely that BERT's pre-training on a large external corpus gives it a big advantage, since it helped the model identify basic linguistic patterns. This may be particularly important for our task, where the overall emotion in a sentence may depend on syntax and semantic structure, rather than simply which words appear. Given the significant gap between the Naive Bayes and BERT model performance, we conclude that deep learning models, especially those based on Transformers, are able to perform better than traditional machine learning models for text classification tasks. Although the attention matrices in our study showed patterns that are consistent with previous work [CKLM19], one disadvantage of the BERT model is that it is more difficult to interpret, and more challenging to determine the impact of individual parameters on the model's predictions.

Future research could investigate different BERT base models and their fine-tuning on the dataset. It is hard to speculate about BERT's ability to detect emotions in "real-world" speech, since the dataset is not well documented and is largely annotated by machines. Future efforts could focus on creating more robust benchmarking datasets and possibly including different languages and styles.

Statement of Contribution. All members contributed to experiments and report writing. RT and JCC performed the data analysis. JT, RT and JCC implemented the models and performed the grid search.

REFERENCES

BO17. Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english, 2017.

CKLM19. Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *ACL 2019*, page 276, 2019.

DCLT19. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

GA. Sugeeshwa SP Galhena and Ajantha S Atukorale. Exploring emotion classification in a labeled twitter dataset: A comparative study of machine learning approaches.

Go09. Alec Go. Twitter sentiment classification using distant supervision. 2009.

Plu01. Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

PSa. Isabeau Prémont-Schwarz. Naive bayes.

PSb. Isabeau Prémont-Schwarz. Neural networks for sequences.

PVG+11. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

QS23. Yuxing Qi and Zahratu Shabrina. Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, 13, 02 2023.

SLH+18. Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

VSP+17. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

ZKZ+15. Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

A. BERT Fine Tuning Grid Search

B. Extra Attention Matrix Plots

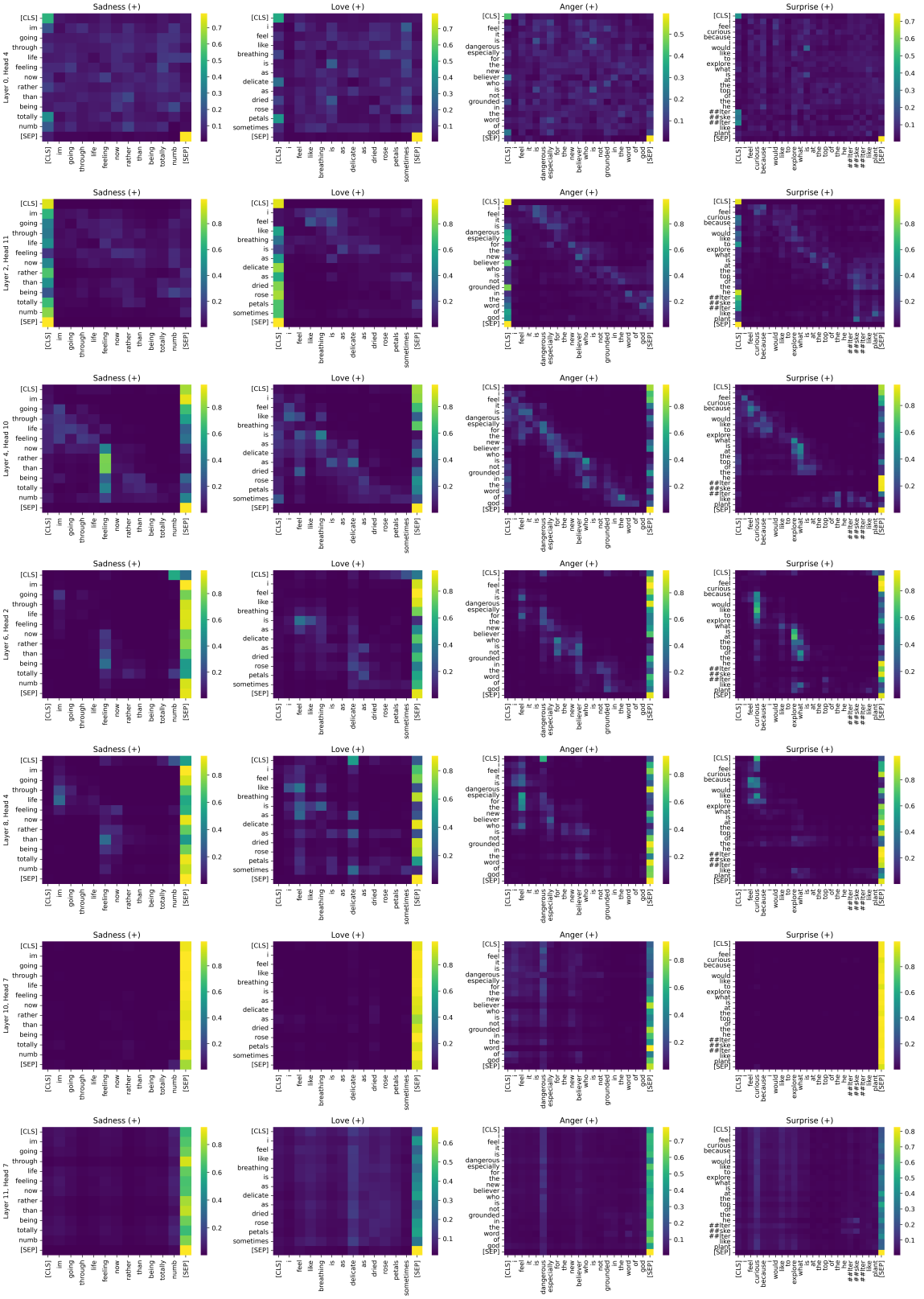

Figure B.1. Attention at different heads and layers for correctly classified samples.

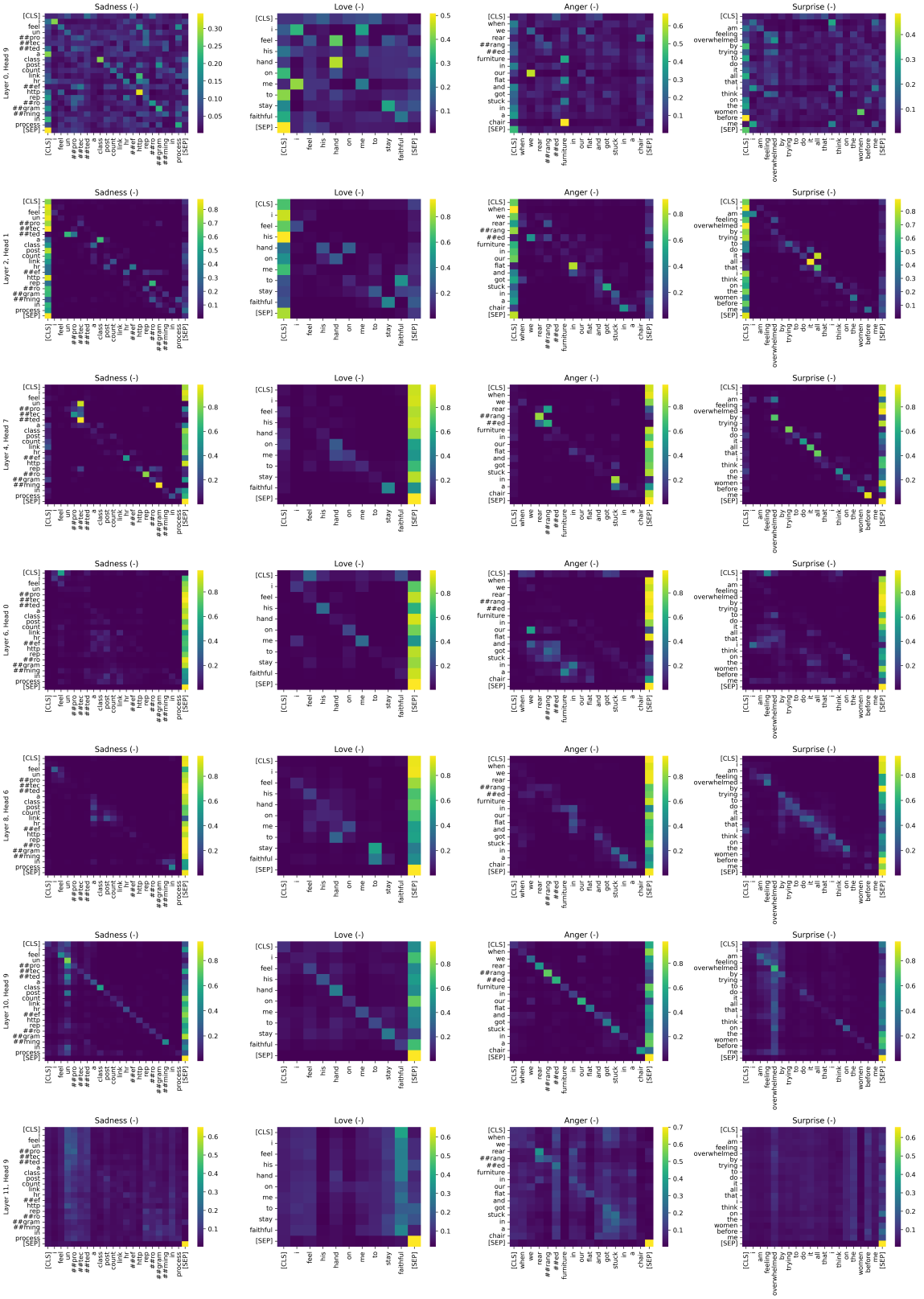

Figure B.2. Attention at different heads and layers for mis-classified samples.

Batch Size	Learning Rate	Weight Decay	Validation Accuracy
128	1×10^{-5}	0.01	0.943
128	2×10^{-5}	0.001	0.942
64	1×10^{-5}	0.001	0.942
32	1×10^{-5}	0.001	0.942
32	3×10^{-5}	0.001	0.942
64	2×10^{-5}	0.001	0.9415
64	3×10^{-5}	0.001	0.9415
128	1×10^{-5}	0.001	0.9415
128	3×10^{-5}	0.001	0.9415
32	2×10^{-5}	0.001	0.941
64	2×10^{-5}	0.01	0.941
64	1×10^{-5}	0.01	0.941
32	1×10^{-5}	0.01	0.9405
128	2×10^{-5}	0.01	0.94
64	3×10^{-5}	0.01	0.94
128	3×10^{-5}	0.01	0.9375
32	2×10^{-5}	0.01	0.937
32	3×10^{-5}	0.01	0.935

TABLE A.2 *Grid search results for the BERT FTA model.*

Batch Size	Learning Rate	Weight Decay	Validation Accuracy
32	1×10^{-5}	0.001	0.942
128	2×10^{-5}	0.001	0.942
64	1×10^{-5}	0.001	0.942
32	3×10^{-5}	0.001	0.942
32	1×10^{-5}	0.001	0.942
128	2×10^{-5}	0.01	0.942
128	2×10^{-5}	0.001	0.942
64	1×10^{-5}	0.001	0.942
32	1×10^{-5}	0.01	0.942
64	3×10^{-5}	0.01	0.9415
32	2×10^{-5}	0.001	0.9415
32	3×10^{-5}	0.001	0.9415
64	2×10^{-5}	0.001	0.9415
128	1×10^{-5}	0.001	0.9415
128	3×10^{-5}	0.001	0.9415
128	3×10^{-5}	0.001	0.9415
128	1×10^{-5}	0.01	0.9415
128	1×10^{-5}	0.001	0.9415
64	3×10^{-5}	0.001	0.9415
64	2×10^{-5}	0.001	0.9415
32	3×10^{-5}	0.01	0.9415
128	3×10^{-5}	0.01	0.9415
64	1×10^{-5}	0.01	0.9415
64	2×10^{-5}	0.01	0.9415
32	2×10^{-5}	0.01	0.941
32	2×10^{-5}	0.001	0.941
64	3×10^{-5}	0.001	0.941

TABLE A.3 *Grid search results for the BERT FTL model.*

Class	**Precision**	**Recall**	**F1-Score**
Sadness	0.8367	0.9174	0.8752
Joy	0.8560	0.8978	0.8764
Love	0.7338	0.6415	0.6846
Anger	0.8683	0.7673	0.8147
Fear	0.8257	0.8036	0.8145
Surprise	0.7941	0.4091	0.5400

TABLE A.4 *Classification Metrics of Naive Bayes for individual classes*