Mini-Project 3:

Classification of Textual Data

Lily Gostovic - 260958505
Mohammad Abdullah - 260980866
Justin Novick - 260965106

A report describing the findings of
Mini-Project 3 of COMP551

COMP551: Applied Machine Learning
McGill University
Montreal, Canada
November 16, 2023

1 Abstract

This paper explores the application of machine learning models for Emotion Detection in textual data. The models considered include a Naive Bayes model, a pre-trained BERT model, and three variations of finetuned BERT models. The primary focus is on comparing their performance in terms of accuracy by class and training time. The Emotion dataset, sourced from English Twitter messages, serves as the basis for evaluation. The models are analyzed for their ability to distinguish emotions such as sadness, joy, love, anger, fear, and surprise. The results indicate that the `BERT-8` model, with eight hidden layers, achieves a comparable accuracy of 92.75% to `BERT-12AW` (93.25%) while demonstrating a 30-40% reduction in training time. Attention matrix analysis provides insights into the decision-making process of `BERT-8`, revealing strengths and weaknesses in classifying emotions. Pre-training effectiveness, the advantages brought by pre-training to this specific task, and the performance disparities between deep learning and traditional machine learning approaches are also discussed. It is concluded that `BERT-8` emerges as the optimal model, offering a balance between accuracy and efficiency for Emotion Detection tasks.

2 Introduction

Emotion detection is an ever-growing and important sub-problem of Natural Language Processing. It is the problem of identifying the emotion that a particular block of text is conveying. The Bidirectional Encoder Representations from Transformers (BERT) model developed by Google in 2018 was trained to "jointly predict a masked word from its context and to classify whether two sentences are consecutive or not" (Jawahar, Sagot, and Seddah 2019). The pre-trained BERT model can be finetuned to tackle more specific Natural Language Processing (NLP) tasks such as question and answering and language inference (Jawahar, Sagot, and Seddah 2019). This project examines the BERT model's ability to extend to tackle emotion detection problems and how it compares in performance to Naive Bayes machine learning models. Refer to (Turc et al. 2019) to read more about the concept of finetuning machine learning models that were pre-trained on broader tasks.

This project compares and contrasts the performance of Naive Bayes and BERT machine learning models on their performance in emotion detection. The Naive Bayes model was implemented from scratch using NumPy. Four variations of BERT models were compared in performance. The variations of BERT were created by finetuning the weights, structure, and hyper-parameters of a BERT model downloaded from HuggingFace. The purpose of this project is to compare the performance of the five machine learning models and draw conclusions on how to create the most optimal model to perform emotion detection tasks and how attention plays a role in the transformer model.

The five aforementioned models were evaluated for performance mainly on two metrics: accuracy by class, and training time. It was found that `BERT-12AW` had the best overall accuracy at 93.25% with `BERT-8` following close behind with an overall accuracy of 92.75%. However, when comparing the training time of these two models, `BERT-8` was approximately 30-40% quicker. This was attributed to the fact that `BERT-8` has less neurons to train due to the fact that it only has 66% the amount of layers as `BERT-12AW`. Therefore, it was concluded that `BERT-8` (post finetuning) is the most optimal model of the five presented in this project for Emotion Detection tasks.

Note: Some of the numbers throughout the document may be slightly off due to multiple training runs, but the trends in the numbers remain consistent.

3 System Models

This project examined the performance of five models: a Naive Bayes model implemented from scratch, a pre-trained BERT model (bert-base-uncased) from Hugging Face, and three different models that are essentially iterations of differently finetuning the downloaded BERT model. The first model is a `Naive Bayes` model implemented from scratch using the NumPy library. This model achieved an accuracy of 76.55% on the Emotion test dataset. The next model is a pre-trained `BERT` model. This model has 12 hidden layers, each with 768 neurons. This model was pre-trained on various Natural Language Processing tasks, including but not limited to Emotion detection tasks. This model achieved an accuracy of 34.85% on the Emotion test dataset. The third model, `BERT-12AW`, finetuned all the weights in `BERT` to achieve a higher accuracy of 93.25% on the Emotion test dataset. The fourth model, `BERT-12L4`, finetuned only the weights in the last four layers on the `BERT` model. It achieved an

Model	Training Time
BERT-12AW	19:19
BERT-12L4	19:56
BERT-8	13:14

Table 1: Comparing Training Time on Finetuned Models

Model	Accuracy
Naive Bayes	76.55%
BERT	34.85%
BERT-12AW	93.25%
BERT-12L4	80.30%
BERT-8	92.75%

Table 2: Comparing Accuracies on Different Models

accuracy of 80.30% on the Emotion test dataset. The fifth model, BERT-8, used the results from BERT-12AW and BERT-12L4 to create the optimal finetuned model to predict the Emotion dataset. It was structurally changed to have only 8 hidden layers instead of the previous 12 hidden layers and trained all weights in all layers since it was concluded in observing BERT-12AW and BERT-12L4 that this is optimal. BERT-8 achieved an accuracy of 92.75% on the Emotion test dataset. The accuracies of the five models can be seen in the below table.

It can be seen in Table 1 that the model that achieves the highest accuracy on the Emotion dataset is BERT-12AW, which achieved the highest accuracy of 93.25%. It is important to note however, that BERT-8 was only slightly less accurate, achieving an accuracy of 92.75% while taking significantly less time to train, due to the smaller size and complexity of the model.

4 Experiments and Conclusions

4.1 The Emotion Dataset

The Emotion dataset (Saravia et al. 2018) was analyzed in this project. The dataset was loaded into Google Colab and then vectorized and tokenized respectively for training the Naive Bayes and BERT based models respectively. The Emotion dataset (Saravia et al. 2018) contains text snippets with labels corresponding to a specific emotional category which the text is meant to convey. The labels are as follows: 0=sadness, 1=joy, 2=love, 3=anger, 4=fear, 5=surprised. The dataset comprises a training set of 16,000 instances and a test set of 2,000 instances. Examples of instances of the dataset can be found in the appendix.
An interesting metric to consider with this dataset is the class distribution of the data in both the training and test sets. Figures 1 and 2 show the class distributions of both datasets. As can be seen, there is significantly more

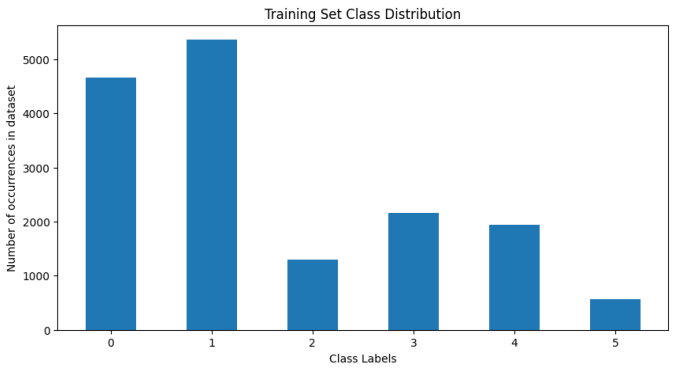

Figure 1: Training Set Class Distribution

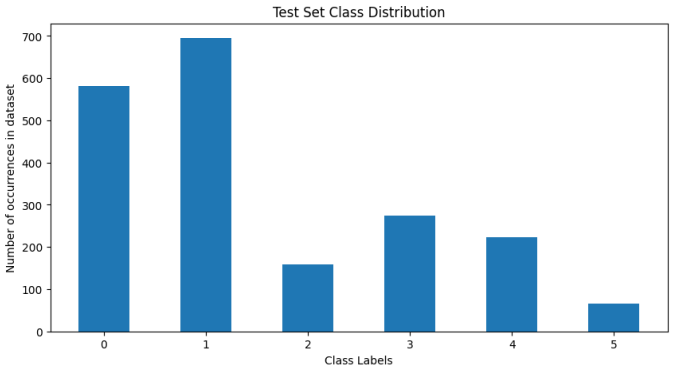

Figure 2: Test Set Class Distribution

data points for labels 0 and 1, sadness, and joy. This must be taken into consideration when training since it is possible that an overfit model may tend to favour and give higher probability to these classes due to their greater presence in the training set. To remedy this problem while running a Naive Bayes model, the use of Laplace Smoothing was employed. Doing so, by including a constant of 1 in the count of each class during training, helps balance the probabilities, of each class appearing and furthermore makes the model slightly less biased towards the over-represented classes.

4.1.1 Ethics of the Emotion Dataset

When using machine learning it is always important to consider the ethics of the data being used for training. The Emotion dataset was populated using English Twitter messages. It is unclear whether Twitter users gave consent to have their Tweets used for machine learning purposes. If they did not give consent, then it is an ethical question whether this data should be used for machine learning purposes, especially since Tweets may convey the writer's deep personal feelings which are being superficially used for the purposes of training machine learning models.

4.2 Attention Matrix Analysis

The following experiments conduct an analysis of attention matrices for different input documents, to help understand the way different tokens from such input documents contribute to the decision-making process of the model. Such analysis is useful for identifying learning patterns and biases. The experiments below were conducted with BERT-8, due to its balance between training time and accuracy. After visualizing all the attention matrices for samples of both correctly and incorrectly classified documents, it became evident that the second head of the second layer played a large role in influencing the class decision. On the other hand, heads from the first layer, and many of the middle layers seemed to mainly pay attention to more abstract features, as seen by an approximately balanced strength from tokens to other tokens. In Figure 4, the attention mechanism for an incorrectly classified document is shown, and in Figure 3 the attention mechanism for a correctly classified document is shown (both pertaining to the second layer's second head). Note that layer and head numbers are 0-indexed. In the incorrectly classified document, it seems that every token had approximately equal attention bearing on the classification when the sentence had multiple subjects. This could potentially be problematic, given that words like "it" and "with" are usually only paired with other words to give meaning with respect to classification, in the English language. We should not see these words getting as much influence as words like "passionate". The latter could have in part caused the incorrect classification seen. On the other hand, in the correctly classified single-subject document, we can see some preposition words (i.e. "so"), and unemotional verbs (i.e. asking) getting less attention to the class than words such as "feeling". This is undoubtedly a working part of the correct classification. The differences in the sentences seem to reveal this mechanism's difficulty in dealing with multiple subjects. More examples of this can be seen in the attached Google Colab.

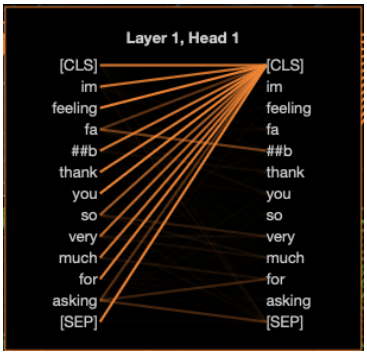

Figure 3: Correctly Classified

Figure 4: Incorrectly Classified

4.3 Training Time Analysis

Table 2 illustrates the average training time for each finetuned model. As can be seen, `BERT-8` takes over 30% less time to train than `BERT-12AW` and `BERT-12L4`. This can be attributed to the fact that the model has significantly less neurons to train since it has on 66% the amount of layers as the other two finetuned models.

4.4 Comparing Performance by Class

Figures 7-9 show the accuracy of each model by class. Each graph shows the percentage of each class the respective model correctly classified. Figure 7 shows that the `Naive Bayes` model performs very well in classifying sadness

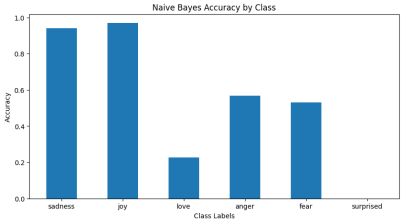

Figure 7: Naive Bayes Class Accuracy

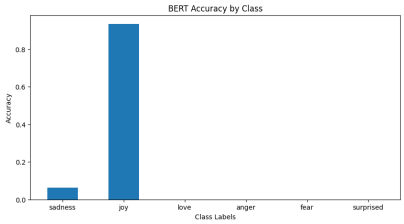

Figure 8: BERT Class Accuracy

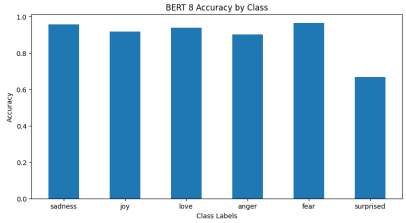

Figure 9: BERT-8 Class Accuracy

and joy, identifying close to 100% of all instances. However, it performs less well on love and surprise, only correctly identifying 20% and 0% respectively. This visualization is especially interesting when considering the `BERT` model which successfully identifies 100% of the joy instances, but predicts close to 0% correct for all other classes. This suggests that the model has a high bias to predicting the sadness class and is not extended well to handle the other classes. This can be explained by the fact that the `BERT` model was not trained on the Emotion dataset upon which it is being evaluated. Finally, looking at the `BERT-8` model, it has a high level of accuracy of above 95% for each class except for the surprise class which is slightly lower around 70%. This is likely due to the fact that the training set contains far fewer instances of this class, resulting in a slightly underfit model to handle this class. The Class Accuracy graphs for `BERT-12AW` and `BERT-12L4` can be seen in Figures 10-11 in the appendix. Overall, it can be seen that the `BERT-8` model performs the best of all five models in terms of best overall class accuracy.

4.5 Confusion Matrix Analysis

This section will compare the confusion matrices for various models, specifically the `BERT-12L4` and `BERT-12AW`. The confusion matrices of the other models can be found in Figures 14-16 in the appendix, however they will not be discussed in this section since they do not reveal as much information as Figures 12 and Figure 13.

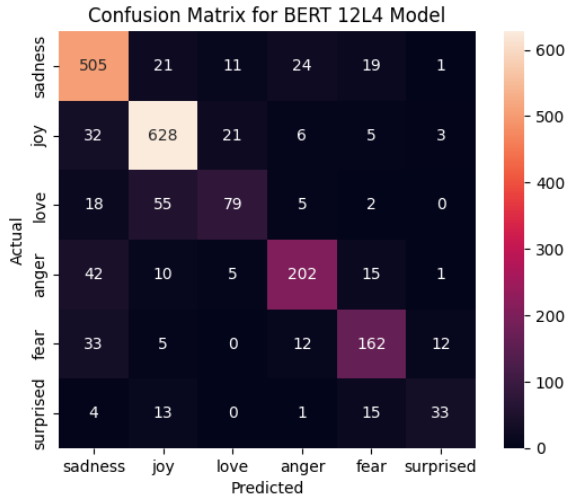

Figure 9: Confusion Matrix for BERT-12L4

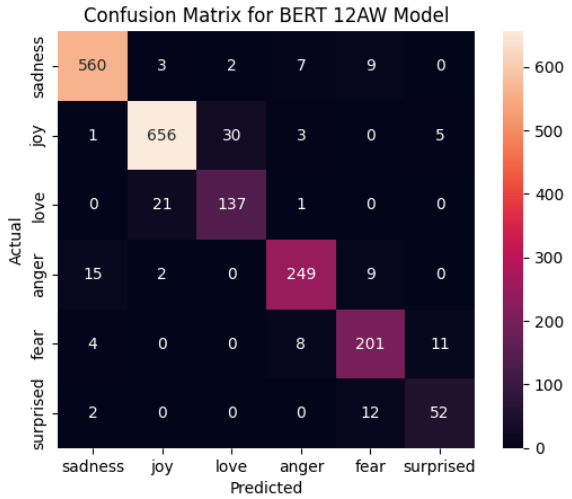

Figure 10: Confusion Matrix for BERT-12AW

The confusion matrix of `BERT-12L4` displays higher correlations between classes than `BERT-12AW`, indicating that it does a worse job of classifying the data. It can be seen in the `BERT-12L4` that the joy and love classes have correlation, as well as the anger, fear, and sadness classes. The classes that exhibit higher levels of correlation are also classes of emotions that humans tend to group together. This result indicates that the training done on this model was working in the correct direction, but did not have enough of a chance to fully grasp all the abstractions

needed to fully distinguish all classes. It is expected that the correlation matrix of `BERT-12AW` has less correlation between classes since it is the same model as `BERT-12L4`, but with more weights changed in training, therefore having had more of a chance to learn the model better. `BERT-12AW` still displays some correlation between classes humans would naturally relate such as joy and love, and fear and surprise, however, the higher amounts of training caused the correlations to decrease significantly, allowing the model to more accurately distinguish these related emotions.

5 Question Analysis

Pre-training on an external corpus, as commonly done with models like BERT, can be beneficial for emotion prediction tasks. Pre-training on a large, diverse dataset helps the model learn rich representations of language and contextual relationships. Figure 17 and Figure 18 show that the attention matrix of layer 0, head 1 for both the incorrectly and the correctly classified documents have a somewhat balanced attention weights. This enables the model to capture the patterns embedded within the tweets. Through this initial assignment of the attention weights, as the model understands the contextual meaning of the words in the sentence and then shifts the attention of each word to the corresponding class, as seen in Figure 3 and Figure 4. Therefore, we can deduce pre-training on an external corpus is beneficial for the emotion prediction task.

Pre-training in models like BERT significantly benefits emotion detection tasks by providing a foundational understanding of language and context with respect to the words in the sentence. The broad knowledge absorbed from the corpus through pre-training allows to model to understand subtle emotional references embedded in the text. This allows more efficient learning. Therefore, pre-training ensures a more robust and contextually aware model, essential for accurately interpreting emotional cues in text.

The comparison between deep learning models, like BERT, and traditional machine learning methods, such as Naive Bayes, in emotion detection tasks reveals distinct performance differences. Deep learning models demonstrate superior accuracy, primarily due to their advanced capabilities in understanding complex language patterns and context cues. They appropriately handle imbalanced data as well, as is shown here in Figure 1 and Figure 2. However, these advantages come at the cost of higher computational resource requirements and longer training times. In contrast, traditional machine learning methods, while less accurate in intricate tasks, are valued for their simplicity, efficiency, and lower computational demands, making them suitable simple tasks.

6 Discussion and Conclusion

This project explored the performance of various machine learning models on Emotion Detection problems. Five models were compared: 1. `Naive Bayes`, implemented from scratch using NumPy, 2. `BERT`, a pre-trained model with 12 hidden layers downloaded from Hugging Face, 3. `BERT-12AW`, an instance of `BERT` in which all weights were finetuned, 4. `BERT-12L4`, an instance of `BERT` in which the weights in the last four layers were tuned, and 5. `BERT-8` a pre-trained model with 8 hidden layers downloaded from Hugging Face which was finetuned on all weights and hyper-parameters. The performance of these models was evaluated on their accuracy by class and training time. It was observed that `BERT-12AW` had the highest accuracy, with the accuracy of `BERT-8` only 0.5% lower. The main metric which distinguished these two models was the training time. `BERT-8` displayed a 30% faster training time when compared to `BERT-12AW` due to its significant less amount of weights to train due to it having 25% less hidden layers. It is for these metrics that it was concluded that `BERT-8` is the optimal model of the ones presented in this project. It is important to realize however that there are many more possible models that can be created by finetuning the BERT model, therefore it is not guaranteed that `BERT-8` is the optimal BERT model for predicting the Emotion dataset.

7 Statement of Contributions

Novick worked on the Naive Bayes Model, Abdullah and Gostovic worked on the BERT Models. Everyone pitched in for their tasks on the report.

References

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah (July 2019). "What does BERT learn about the structure of language?" In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy. URL: https://hal.archives-ouvertes.fr/hal-02131630.

Saravia, Elvis et al. (Oct. 2018). "CARER: Contextualized Affect Representations for Emotion Recognition". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3687–3697. DOI: 10.18653/v1/D18-1404. URL: https://www.aclweb.org/anthology/D18-1404.

Turc, Iulia et al. (2019). "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models". In: *arXiv preprint arXiv:1908.08962v2*.

Appendix

Sample Data

1. `sadness`: ive been feeling a little burdened lately wasnt sure why that was
2. `sadness`: i feel like i have to make the suffering i m seeing mean something
3. `joy`: i have been with petronas for years i feel that petronas has performed well and made a huge profit
4. `joy`: i do feel that running is a divine experience and that i can expect to have some type of spiritual encounter
5. `love`: i feel romantic too
6. `love`: i can t let go of that sad feeling that i want to be accepted here in this first home of mine
7. `anger`: i am feeling grouchy
8. `anger`: i feel selfish as i read back to my former posts how i have never asked for prayers for others how i never considered that there may be others out there that deserve their prayers answered before my own
9. `fear`: i feel as confused about life as a teenager or as jaded as a year old man
10. `fear`: i will be able to lay on my bed in the dark and not feel terrified at least for a while
11. `surprised`: ive been taking or milligrams or times recommended amount and ive fallen asleep a lot faster but i also feel like so funny
12. `surprised`: i am now nearly finished the week detox and i feel amazing

Figure 5: Correctly classified document: *im feeling fab thank you so very much for asking*

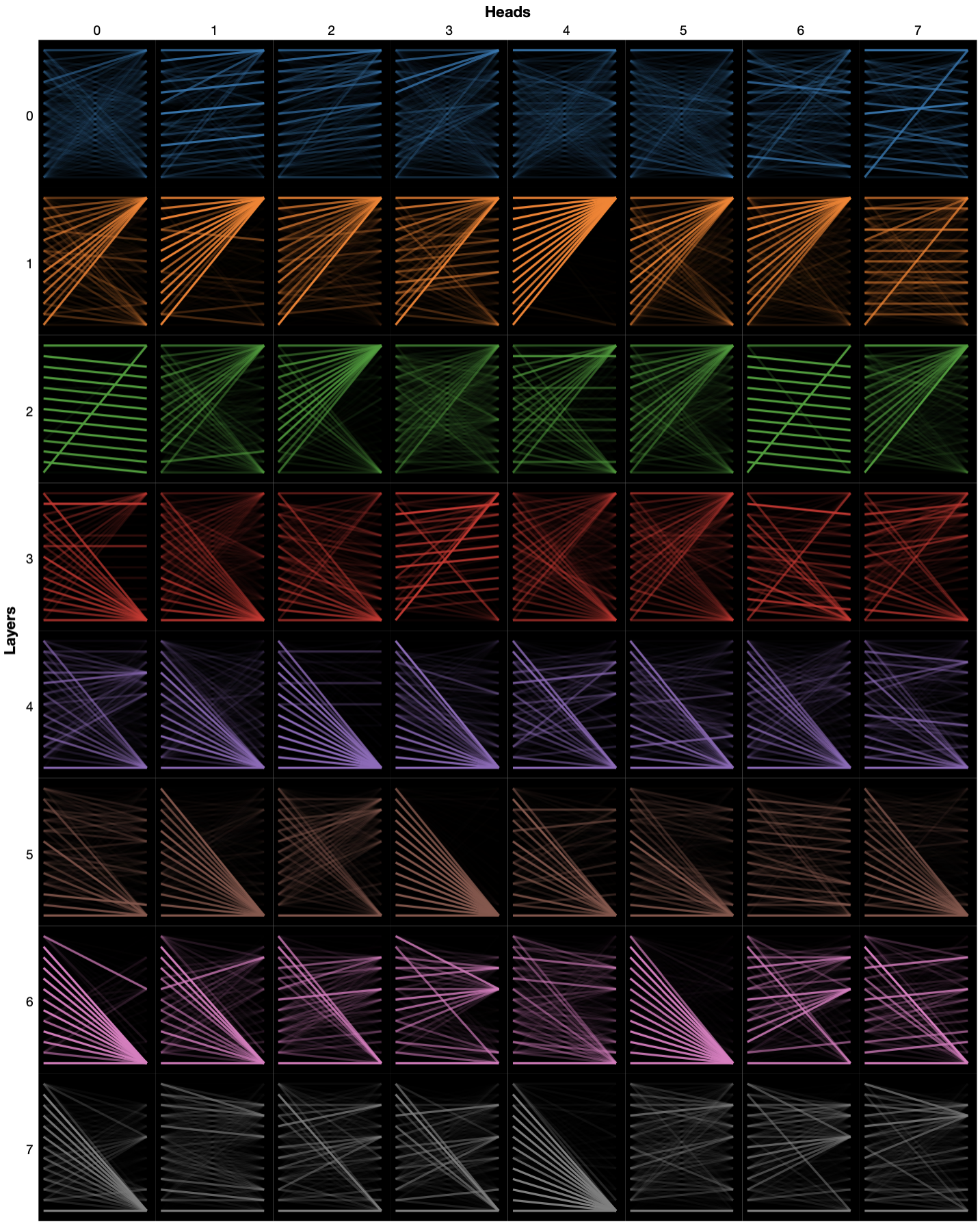

Figure 7: Incorrectly classified document: *i ever recognized what it was to feel passionate about something was with music*

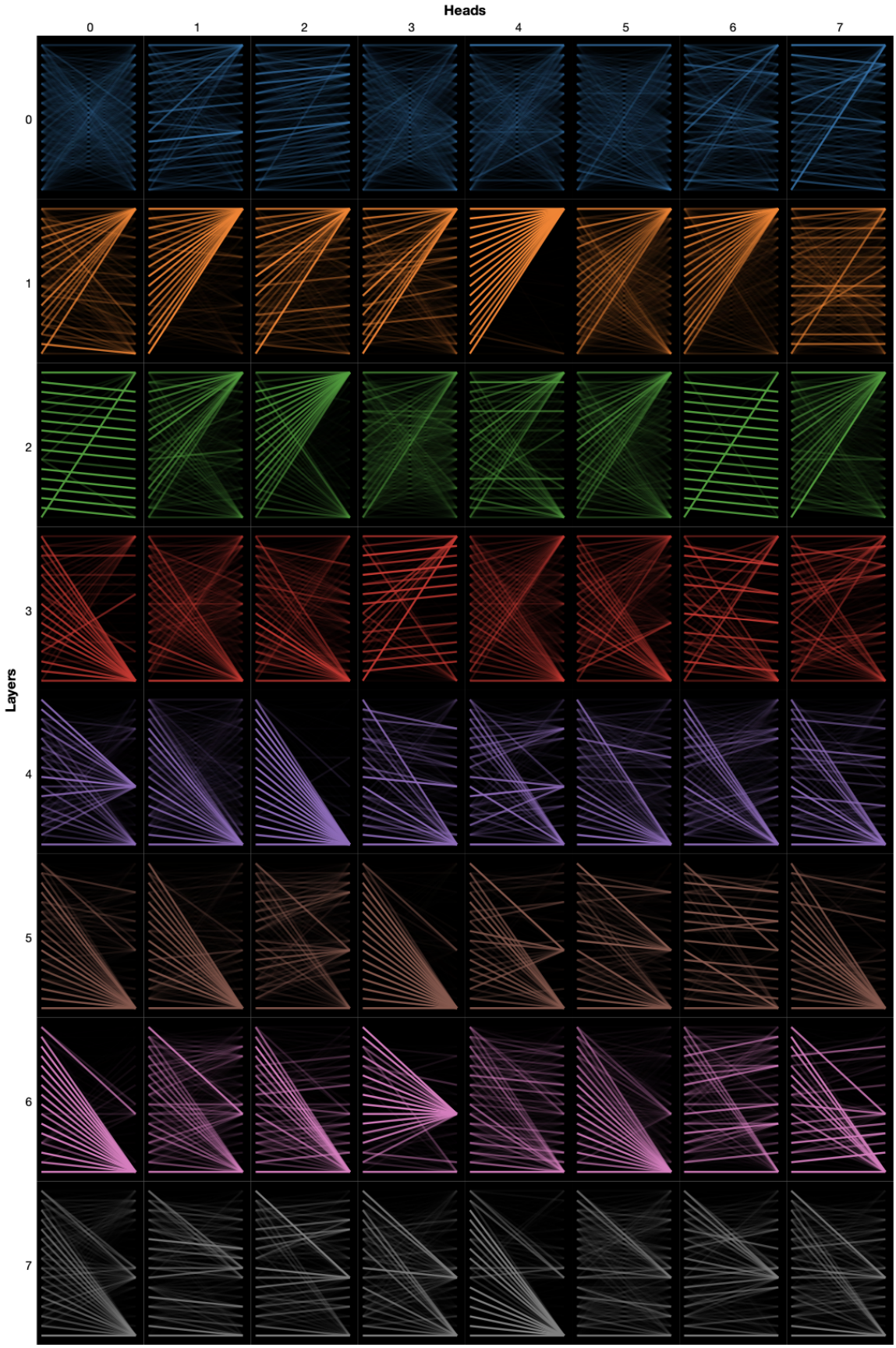

Figure 10: `BERT-12AW` **Class Accuracy Graph**

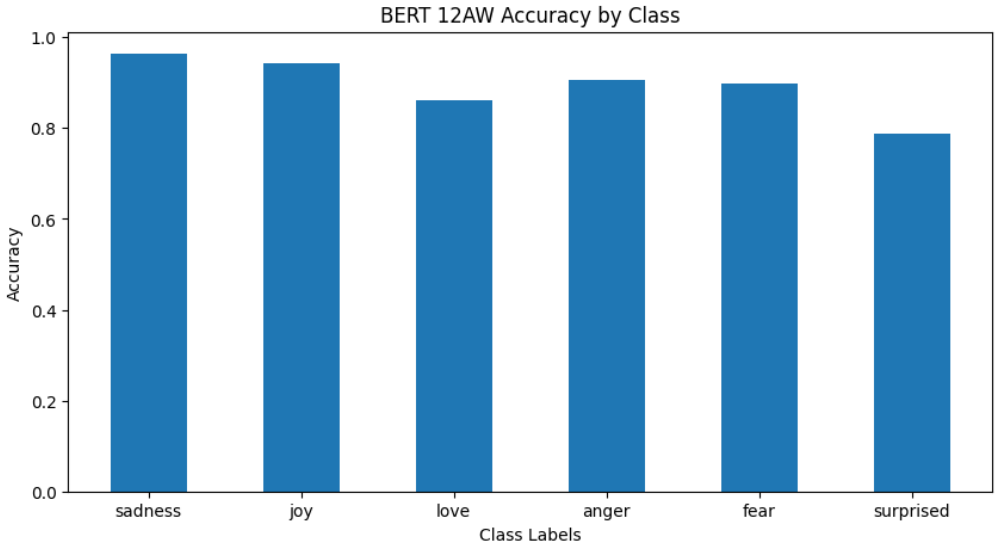

Figure 11: `BERT-12L4` **Class Accuracy Graph**

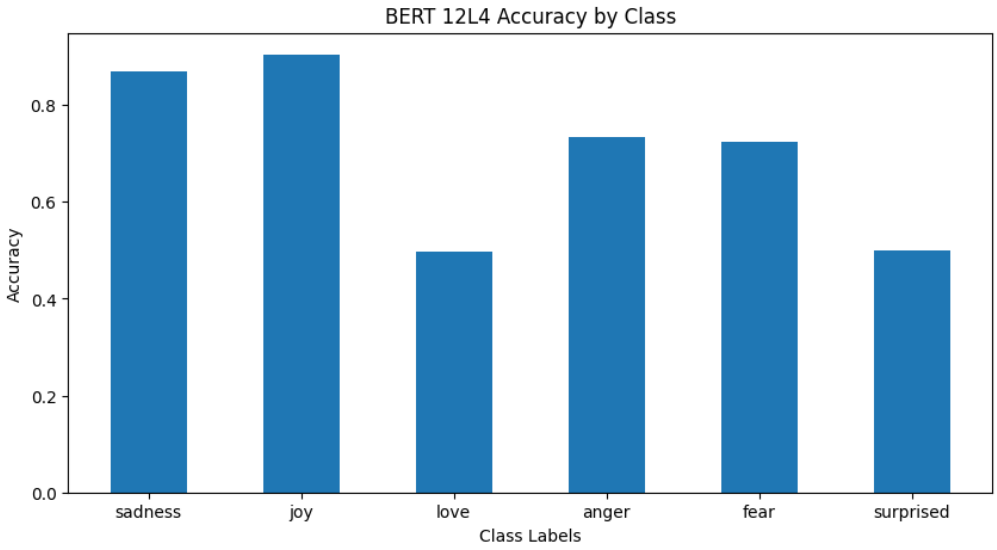

Figure 14: `Naive Bayes` **Confusion Matrix**

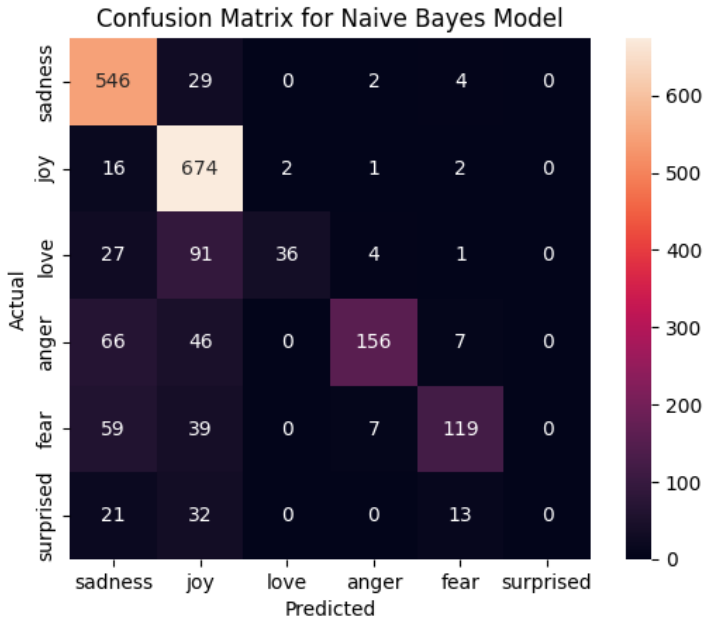

Figure 15: `BERT` **Confusion Matrix**

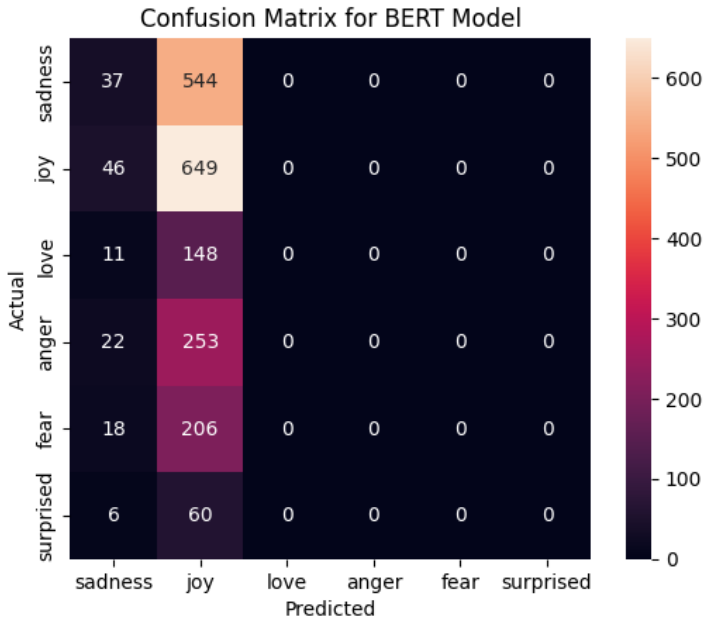

Figure 16: BERT-8 Confusion Matrix

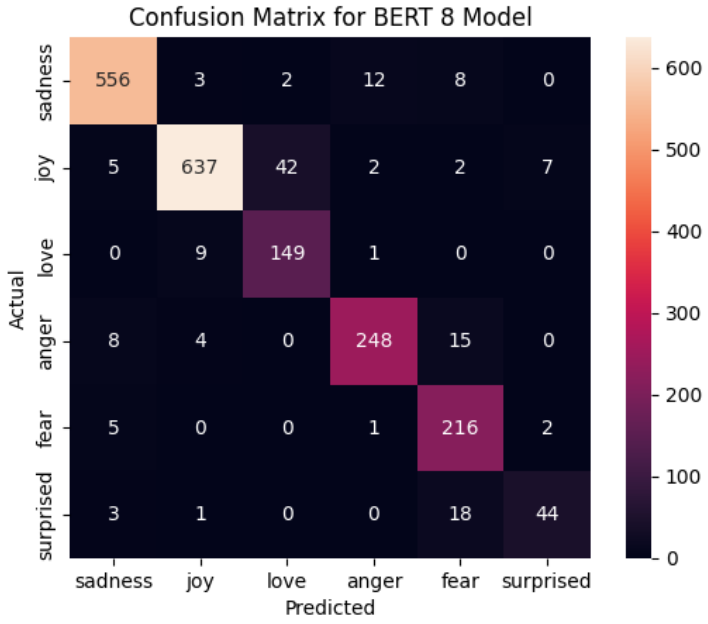

Figure 17: The correctly classified sentence, Layer 0, Head 1

Figure 18: The incorrectly classified sentence, Layer 0, Head 1