COMP 551 - HW3

Applied Machine Learning

Lancelot Normand
261155638
lancelot.normand@mail.mcgill.ca

Eduard Anton
261033247
eduard.anton@mail.mcgill.ca

Jessica Zhu
260957235
jessica.zhu@mail.mcgill.ca

Fall 2023

School of Computer Science

McGill

Contents

1 Abstract **3**

2 Introduction **3**

3 System Models **3**

4 Dataset **3**
4.1 Preprocessing Methods 4

5 Experiments **4**
5.1 Naive Bayes on Dataset 4
5.2 Naive Bayes with Laplace Smoothing 5
5.3 Naive Bayes with Balanced Dataset 5
5.4 Pretrained BERT on Dataset 5
5.5 Finetuned BERT on Dataset 5
5.6 Finetuned BERT with Regularization 6
5.7 Finetuned BERT on Last Layers 6
5.8 Finetuned BERT with Balanced Dataset 6

6 Results **7**
6.1 Performance Analysis of Naive Bayes and BERT 7
6.2 Transformers for NLP 7

7 Conclusion **7**

8 Appendix **9**
8.1 Performance of naive Bayes, BERT-based models and our implementation 9
8.2 Naive Bayes 9
8.3 BERT-based models 9
8.4 Attention 10
8.4.1 Predicted: sadness, Truth: sadness (Layer 11, Head 8) 10
8.4.2 Predicted: anger, Truth: anger (Layer 11, Head 8) 10
8.4.3 Predicted: love, Truth: joy (Layer 11, Head 8) 10
8.4.4 Predicted: surprise, Truth: joy (Layer 11, Head 8) 10

List of Figures

1 Class distribution of Emotion train dataset 4
2 Class distribution of Emotion test Dataset 4
3 Balanced Training Set for BERT 6
4 Finetuned BERT with Balanced Dataset 6
5 Naive Bayes No Smoothing 9
6 Naive Bayes Smoothing of 1 9
7 Naive Bayes Smoothing of 0.5 9
8 Naive Bayes Balanced Dataset 9
9 BERT Pretrained 9
10 BERT Finetuned 9
11 BERT Finetuned with Regularization 9
12 BERT Fintetuned on Last Layers 9
13 I didn't feel at all deprived this morning having it in my chai this morning 10
14 I am trying not to feel bitter but how else can I feel when it seems my desire is pretty much impossible. 10
15 I feel very strongly about supporting charities that help children 10
16 I loved the feeling I got during an amazing slalom run whether it was in training or in a race. 10

List of Tables

1 Training and validation loss during BERT fine-tuning 6
2 Performance comparison of our implementation of Naive Bayes, Vanilla Naive Bayes, and Bert 9

1 Abstract

Abstract Emotion recognition fram text is an essential component of natural language processing. This research delves into the critical realm of emotion recognition from text within natural language processing (NLP). Investigating the Dair-ai emotion dataset, we assess the effectiveness of Naive Bayes and BERT-based deep learning models for text classification based on emotions. Our experimentation encompasses fine-tuning adjustments on both Naive Bayes and BERT, coupled with a thorough analysis of preprocessing techniques. Aligning with existing literature, our findings underscore the superior performance of deep learning models, particularly BERT, over Naive Bayes in the context of emotion classification tasks.

2 Introduction

People convey emotions through language and the internet has made sharing one's thoughts so much easier. Everyday, 500 million tweets are sent and Twitter is just one of many social media platforms [1]. With the amount of data available, textual data can be analysed for emotion patterns. Early emotion recognition systems employed Naives Bayes classifiers [2]. One challenge of applying neural networks to solve NLP tasks is the sequential processing of the input. But the proposal by Vaswani et al. (2017) for an attention-based transformer architecture provided a solution to this problem [3]. Most of the state-of-the-art models for emotion recognition are transformer-based [4]. Transformer is an encoder-decoder model leveraging attention mechanisms to allow for parallel input processing [3]. Various benchmark datasets for emotion recognition have been gathered over the years. The Emotion dataset consists of data from emotion-conveying tweets and was put together by Saravia et al. (2018), who proposed a CNN-based model for recognizing emotion patterns [5]. The proposed model achieved a F1-score of 0.79 and an accuracy of 81% [5]. A 2023 study fine-tuned a pre-trained BERT model by building a vocabulary of emotional words from BERT's training corpus and adjusting the embedding of emotional words and neutral words [6]. For the Emotion dataset, they report an accuracy of 0.8782 for pretrained BERT-based CNN and an accuracy of 0.9286 with the fine-tuned BERT-based CNN [6]. In this paper, we aim to investigate how Naive Bayes algorithms and BERT-based deep learning models perform on the Emotion dataset.

3 System Models

Two main models are used in this mini-project: Naive Bayes and BERT. Naive Bayes is a traditional machine learning model commonly used for text classification. The principle of this algorithm is based on the application of Bayes theorem with a strong independence assumption. BERT is an encoder-only transformer-based model that is trained in a bidirectional manner and achieves state-of-the-art results on a variety of NLP tasks[7]. It provides a pre-trained model which can be further fine-tuned to obtain better results for a specific task and is often employed in combination with other deep learning models (ex. CNN) [8].

4 Dataset

We conducted text classification using the dair-ai/emotion dataset [5] for emotion analysis. Originating from Twitter messages, this dataset comprises numerous instances of sentences, each belonging to one of six emotions: sadness, joy, love, anger, fear, and surprise. The dataset is thoughtfully divided into three splits (training, validation, and test) with 16,000,

2,000, and 2,000 instances, respectively. As showcased in Fig. 1 and Fig. 2, the distribution of classes within the dataset is uneven, with joy and sadness exhibiting the highest instances and surprise having the fewest across all splits. This inherent class imbalance presents a unique challenge in training robust models that can effectively handle the varying frequencies of different emotional expressions.

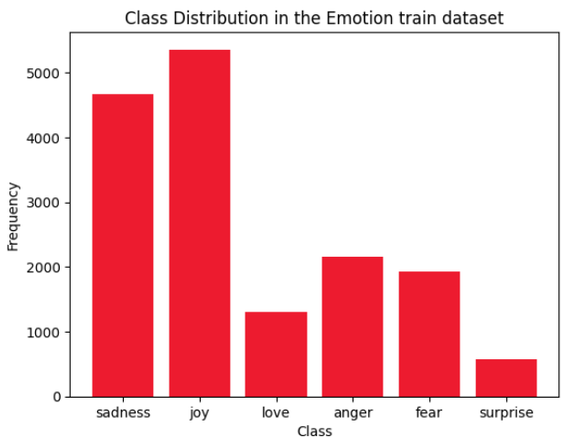

Figure 1: Class distribution of Emotion train dataset

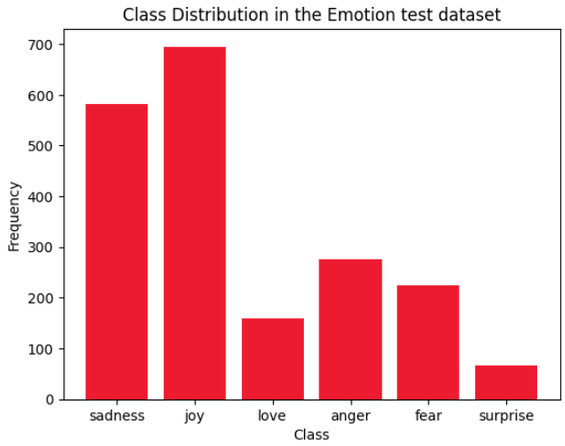

Figure 2: Class distribution of Emotion test Dataset

4.1 Preprocessing Methods

A preprocessed dataset from the original emotion dataset was generated using lemmatization, removal of irrelevant stop words and expansion of contractions. Lemmatization involves removing the inflection endings (ex. suffix added from tense changes, -s ending for plurals of words, -er ending for the comparative form) in the dataset. The dataset was lemmatized using the WordNetLemmatizer() function in the ntlk library. Stop words are common words in a sentence that do not affect the overall meaning of the sentence. For example, words like "I", "the", "a" are common stop words. The list of English stopwords provided by the ntlk library contains negative words such as "not" and "no". The removal of these negative words would alter the emotion of a sentence so negative words were removed from the list of stopwords. Contractions are the abbreviated form of a combination of words (ex. I'm, I've). Failure of ML models to recognize contractions and its fully written form as equivalent can negatively affect the model performance. Therefore, we expand any contractions using the 'contractions' library in Python. However, it was observed that using the preprocessed dataset for training actually slightly worsens the performance of the Naive Bayes and BERT-based models. So we still opted to use the original dataset.

5 Experiments

5.1 Naive Bayes on Dataset

In our exploration of emotion classification using the Naive Bayes algorithm with a Categorical distribution for the prior and a Multinomial distribution for the likelihood, we achieved an average accuracy of 71.1% on the test set. However, a closer examination reveals a nuanced perspective on the model's performance. The dominance of instances from the sadness and joy classes in the test dataset skews the average accuracy, and in reality, the model's performance across all classes yields an average accuracy of 59.33%. Looking into the confusion matrix in Fig. 5, intriguing patterns emerge, particularly in the misclassification of love as joy and surprise as fear, sadness or joy.

5.2 Naive Bayes with Laplace Smoothing

In our pursuit of refining the Naive Bayes model's performance, we implemented Laplace smoothing to alleviate the issue of zero probabilities and enhance the model's adaptability to unseen data. As seen in Fig. 6, employing a Laplace smoothing value of 1, we observed an improvement in the average accuracy on the test dataset, reaching 76.55%. However, a trade-off emerged, as the average accuracy over all classes drops to 54%. Notably, the model exhibits enhanced predictive capabilities for most classes, yet issues persist in accurately identifying instances related to love and surprise, with the latter yielding an accuracy of 0%. A subsequent attempt with a Laplace smoothing value of 0.5 yielded a highest average accuracy of 80% on the test dataset, accompanied by an the best average class accuracy of 62.91%, as presented Fig. 7. While this represents progress, the model's struggle to effectively discern instances associated with love and surprise persists.

5.3 Naive Bayes with Balanced Dataset

Next, we implemented a preprocessing step on the dair-ai/emotion dataset, ensuring an equal number of instances for each class. By selecting 572 instances for every class, matching the minimum number found in the surprise class, we tried to address the challenge posed by the initial class imbalance. Indeed, Naive Bayes can in some cases perform reasonably well on smaller datasets [3], and by selecting a smaller but more balanced dataset we attempted to improve on the generalization of our model. The results were evident, with the model achieving a final average accuracy of 64.65% on the test dataset but with an average class accuracy of now 70.85%. As presented in Fig. 8, this preprocessing approach notably reduced the confusion that was previously evident, particularly for the love and surprise classes. Surprisingly, both classes surpassed the performance of the initial champions, joy, and sadness, which experienced a significant drop in accuracy.

5.4 Pretrained BERT on Dataset

Subsequently, we turned to deep learning and employed a pretrained BERT model, specifically bert-base-uncased-emotion, which had been finetuned from bert-base-uncased [7] on an emotion dataset. The results obtained from the model are outstanding as demonstrated in Fig. 9, showcasing a remarkable average accuracy of 92.65% on the test dataset and an impressive 87.98% accuracy when averaged across all classes. The model's ability to discern and classify emotions surpasses that of the Naive Bayes counterpart. However, a nuanced analysis reveals persistent issues in distinguishing between love and surprise classes, indicating further needed model refinement.

5.5 Finetuned BERT on Dataset

To refine the performance of the model for emotion classification, we undertook further fine-tuning on the bert-base-uncased-emotion model using the emotion dataset. Maintaining the initial fine-tuning parameters, a learning rate of 2e-5 and 8 epochs were used, but a batch size of 32 instead of 64 was considered. The results in Fig. 10 yielded a nearly unchanged accuracy of 92.4% on the test dataset and 87.67% in accuracy when averaged across all classes. However, a noteworthy observation emerged during the validation phase, where we detected a gradual increase in validation loss during training as seen in Table 1, signaling a potential case of overfitting.

Epoch	1	2	3	4	5	6	7	8
Training loss	0.097	0.074	0.051	0.033	0.023	0.019	0.014	0.010
Validation loss	0.197	0.227	0.258	0.296	0.329	0.356	0.339	0.342

Table 1: Training and validation loss during BERT fine-tuning

5.6 Finetuned BERT with Regularization

To mitigate overfitting in the fine-tuning of the BERT model, we implemented a set of strategic adjustments. We increased the dropout rate on all layers from 10% to 15%, providing a more robust regularization mechanism. We also opted to reduce the learning rate (2e-6) and the number of epochs to prevent catastrophic forgetting [9], where pre-trained knowledge is lost during fine-tuning. Simultaneously, we decreased the batch size, aligning with recommendations from [10] that suggest smaller batch sizes can enhance accuracy. The results indicate a significant reduction in overfitting, with the accuracy on the test dataset slightly improved at 92.8%, and the average class accuracies reaching 88.5% as seen in Fig. 11.

5.7 Finetuned BERT on Last Layers

Building upon the encouraging results of our previous experiments, we chose to align with the methodology proposed in Paper [9], focusing our optimization efforts exclusively on the classifier layers and the last two attention layers of the BERT model. The underlying premise puts forward the idea that these layers are already adept at comprehending English sentences, with the last layers handling a more substantial influence on emotion classification. Unfortunately, as seen on Fig. 12, this approach yielded a consistent accuracy of 92.8% on the test dataset, and a modest improvement was observed in the average class accuracy, which rose to 89%. Although, the most notable impact surfaced in the training efficiency, with the model's training time trimmed from 22:56 minutes to 10:22 on an NVIDIA T4.

5.8 Finetuned BERT with Balanced Dataset

To address the challenge of lower performance in the love and surprise classes, we opted to rebalance our dataset. This involved selectively retaining only half of the training instances for the joy and sadness classes and duplicating instances for the surprise class as seen in Fig. 3. The results of this rebalancing yielded an almost unchanged test accuracy of 92.35%. However, the more significant outcome was the improvement in the average class accuracy, rising to an impressive 90.67% as indicated in Fig. 4. We also notice that the performance gap, previously a cause for uncertainties, between the surprise and fear, and the love and joy classes significantly decreased while other other classes kept a similar performance.

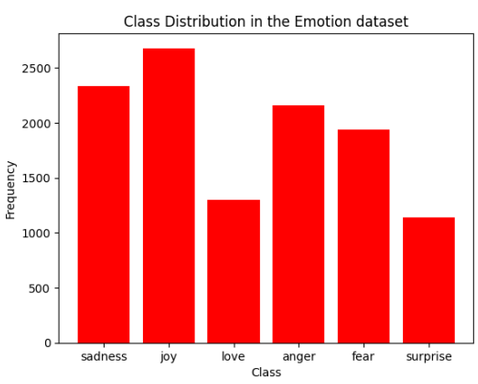

Figure 3: Balanced Training Set for BERT

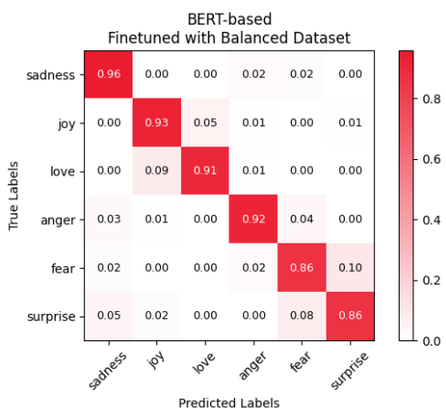

Figure 4: Finetuned BERT with Balanced Dataset

6 Results

6.1 Performance Analysis of Naive Bayes and BERT

Comparing the performance of Naive Bayes and our pretrained BERT model for emotion classification in Table 2 highlights a stark difference in their respective capabilities. BERT demonstrates a clear superiority with a test accuracy of 92.5% and an impressive class accuracy average of 90.67%, outperforming Naive Bayes which achieves a maximum test accuracy of 80% at a class accuracy of 62.91%. The marked difference in performance can be attributed to BERT's ability to leverage the complexity of its model architecture, harnessing large amounts of data and utilizing attention mechanisms. BERT's contextual understanding of words in a sentence allows it to capture intricate relationships and nuances that Naive Bayes, with its assumption of word independence, struggles to do. However, this performance comes at a cost, as BERT is magnitudes slower to run and train compared to Naive Bayes.

6.2 Transformers for NLP

Pretraining on an external corpus, as exemplified by our BERT model, proves advantageous for the emotion prediction task. Pretraining equips the model with a rich understanding of linguistic nuances and context from diverse textual sources as tested in section 5.7. This foundational knowledge empowers the model to discern and classify emotions more effectively when fine-tuned.

Fig. 13 and Fig. 14 show the attention matrices (layer 11, head 8) for two instances of correctly predicted sentences, while Fig. 15 and Fig. 16 are the attention matrices of two incorrectly predicted sentences, where brighter cells indicate larger attention weights. We observe for correctly predicted sentences, each query focuses to a smaller set of keys, with the majority of keys focusing more on words that evoke strong emotions. However, for incorrectly predicted sentences, the attention weights assigned to each key for each query tends to be smaller (though not always the case) and more spread out along multiple keys. Both of the incorrectly predicted sentences do not contain any emotional words strongly associated with one type of emotion. For example, Fig. 16 contains the word "loved" but in the context of the sentence, it would mean the person greatly enjoying something rather than feeling the emotion of love. When we read these sentences, we did find the emotion conveyed by the sentences to be ambiguous and were not able to correctly determine the label either.

7 Conclusion

In conclusion, our experiments in NLP emotion classification show the substantial performance advantage offered by deep learning pretrained models, exemplified by BERT, over more traditional models like Naive Bayes. The robust accuracy achieved by BERT, with a test accuracy of 92.5% and an impressive class accuracy average of 90.67%, far surpasses the capabilities of Naive Bayes, which attains a maximum test accuracy of 80% and a class accuracy of 62.91%. However, this heightened performance comes at a notable cost: expensive computation. The choice between accuracy and computational efficiency becomes a crucial consideration in selecting the right model. Future research could delve into exploring alternative models, such as transformer-based architectures beyond BERT and RNNs to find a balanced solution for NLP emotion classification tasks.

References

[1] LiveStats, *Twitter usage statistics - internet live stats*, 2009. [Online]. Available: https://www.internetlivestats.com/twitter-statistics/.

[2] N. Sebe, M. Lew, I. Cohen, A. Garg, and T. Huang, "Emotion recognition using a cauchy naive bayes classifier," in *2002 International Conference on Pattern Recognition*, vol. 1, 2002, 17–20 vol.1. DOI: 10.1109/ICPR.2002.1044578.

[3] B. Seref and E. Bostanci, "Performance comparison of naïve bayes and complement naïve bayes algorithms," in *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*, Apr. 2019, pp. 131–138. DOI: 10.1109/ICEEE2019.2019.00033.

[4] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, e12189, 2020. DOI: https://doi.org/10.1002/eng2.12189. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.12189. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.12189.

[5] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "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, Oct. 2018, pp. 3687–3697. DOI: 10.18653/v1/D18-1404. [Online]. Available: https://www.aclweb.org/anthology/D18-1404.

[6] Z. Zhu and K. Mao, "Knowledge-based bert word embedding fine-tuning for emotion recognition," *Neurocomputing*, vol. 552, p. 126 488, 2023, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126488. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223006112.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805.

[8] J. Dong, F. He, Y. Guo, and H. Zhang, "A commodity review sentiment analysis based on bert-cnn model," in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 2020, pp. 143–147. DOI: 10.1109/ICCCS49078.2020.9118434.

[9] C. Sun, X. Qiu, Y. Xu, and X. Huang, *How to fine-tune bert for text classification?* 2020. arXiv: 1905.05583 [cs.CL].

[10] L. Luo and Y. Wang, *Emotionx-hsu: Adopting pre-trained bert for emotion classification*, 2019. arXiv: 1907.09669 [cs.CL].

8 Appendix

8.1 Performance of naive Bayes, BERT-based models and our implementation

	Naive Bayes				BERT				
	Normal	Laplace Smoothing			Pretrained	Our Finetuned models			
Accuracy	**Default**	**Smoothing 1**	**Smoothing 0.5**	**Balanced Dataset**	**Pretrained**	**Finetuned**	**Finetuned Regularization**	**Finetuned Last Layers**	**Finetuned Balanced Dataset**
Test	71.15	76.55	80.0	64.65	92.65	92.4	92.8	92.8	92.35
Sadness	77.8	94.0	91.9	62.5	96.38	96.90	97.41	96.90	95.86
Joy	84.3	97.0	93.4	57.0	94.96	93.81	93.81	94.24	92.66
Love	49.1	22.6	42.1	80.5	79.25	82.39	84.28	83.65	90.57
Anger	58.9	56.7	69.1	74.2	90.91	91.64	89.82	90.18	91.64
Fear	55.8	53.1	67.4	71.0	93.30	91.07	93.30	91.96	86.16
Surprise	30.3	00.0	13.6	83.3	72.73	69.70	72.73	77.27	86.36

Table 2: Performance comparison of our implementation of Naive Bayes, Vanilla Naive Bayes, and Bert

8.2 Naive Bayes

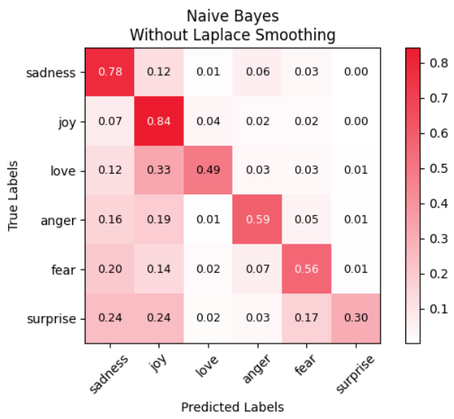

Figure 5: Naive Bayes No Smoothing

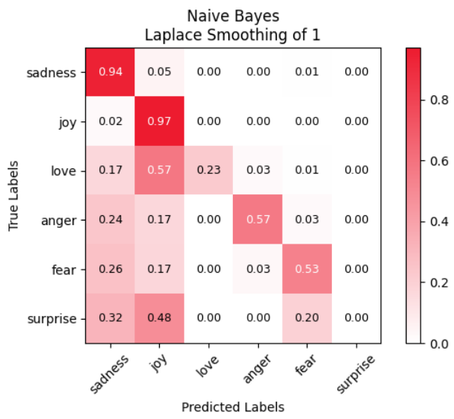

Figure 6: Naive Bayes Smoothing of 1

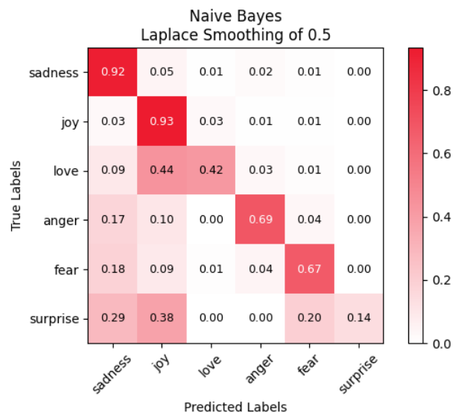

Figure 7: Naive Bayes Smoothing of 0.5

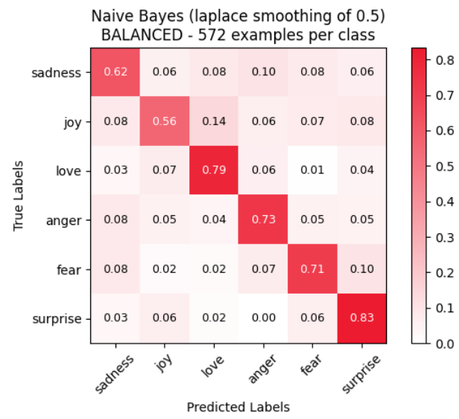

Figure 8: Naive Bayes Balanced Dataset

8.3 BERT-based models

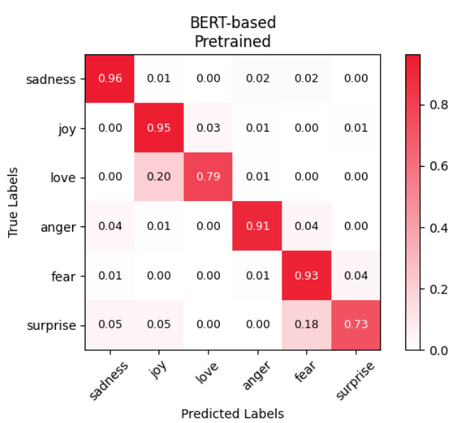

Figure 9: BERT Pretrained

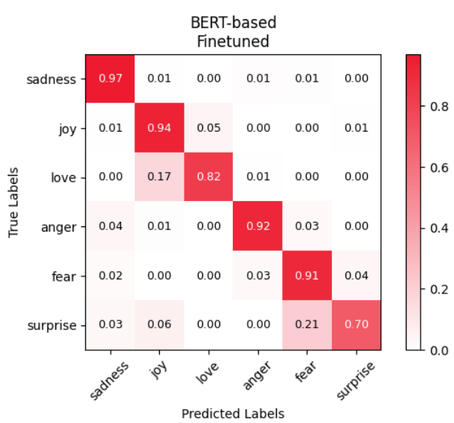

Figure 10: BERT Finetuned

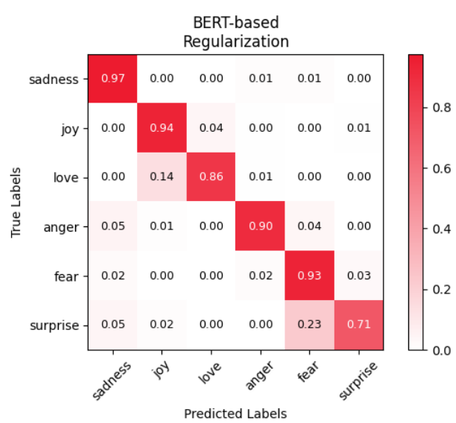

Figure 11: BERT Finetuned with Regularization

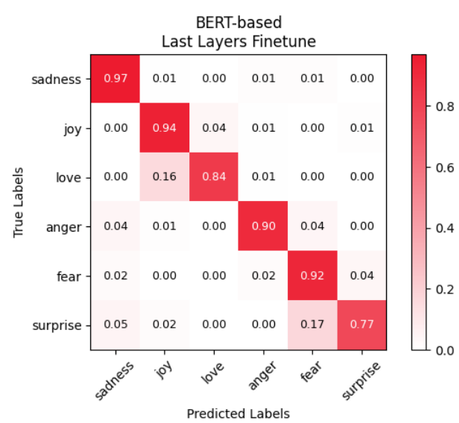

Figure 12: BERT Fintetuned on Last Layers

8.4 Attention

8.4.1 Predicted: sadness, Truth: sadness (Layer 11, Head 8)

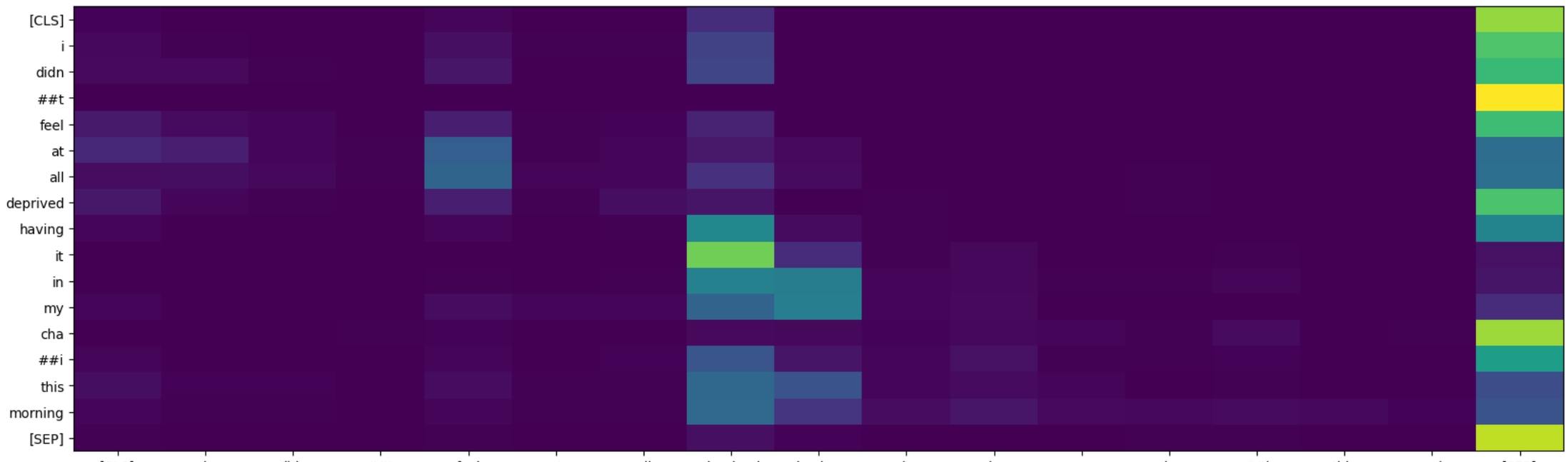

Figure 13: I didn't feel at all deprived this morning having it in my chai this morning

8.4.2 Predicted: anger, Truth: anger (Layer 11, Head 8)

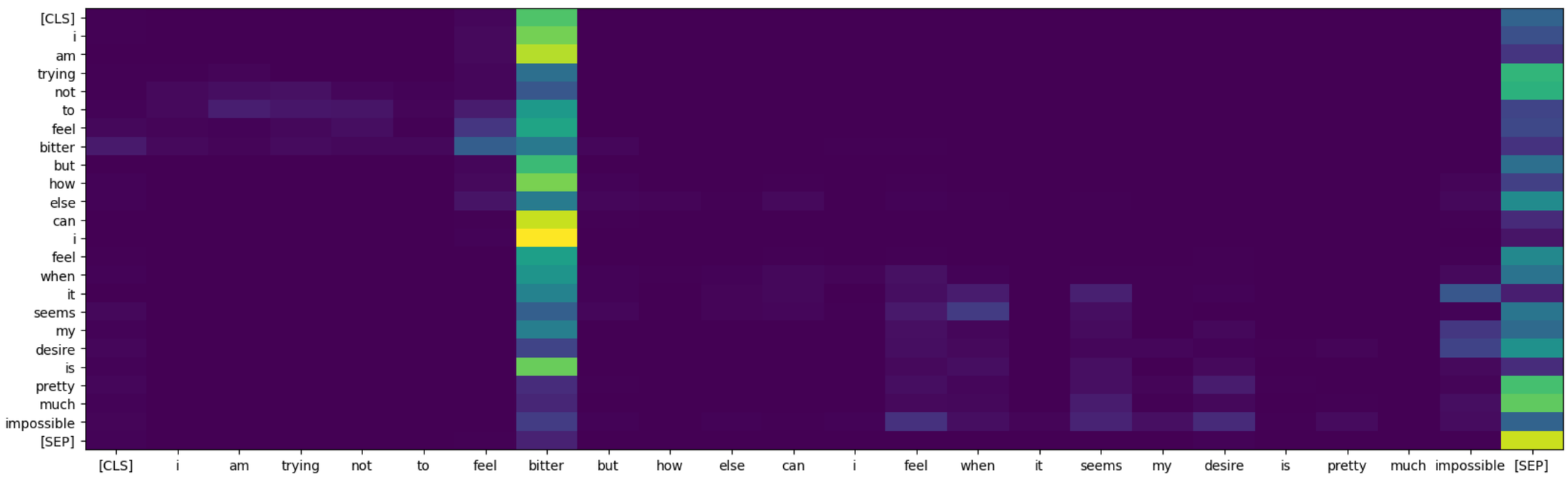

Figure 14: I am trying not to feel bitter but how else can I feel when it seems my desire is pretty much impossible.

8.4.3 Predicted: love, Truth: joy (Layer 11, Head 8)

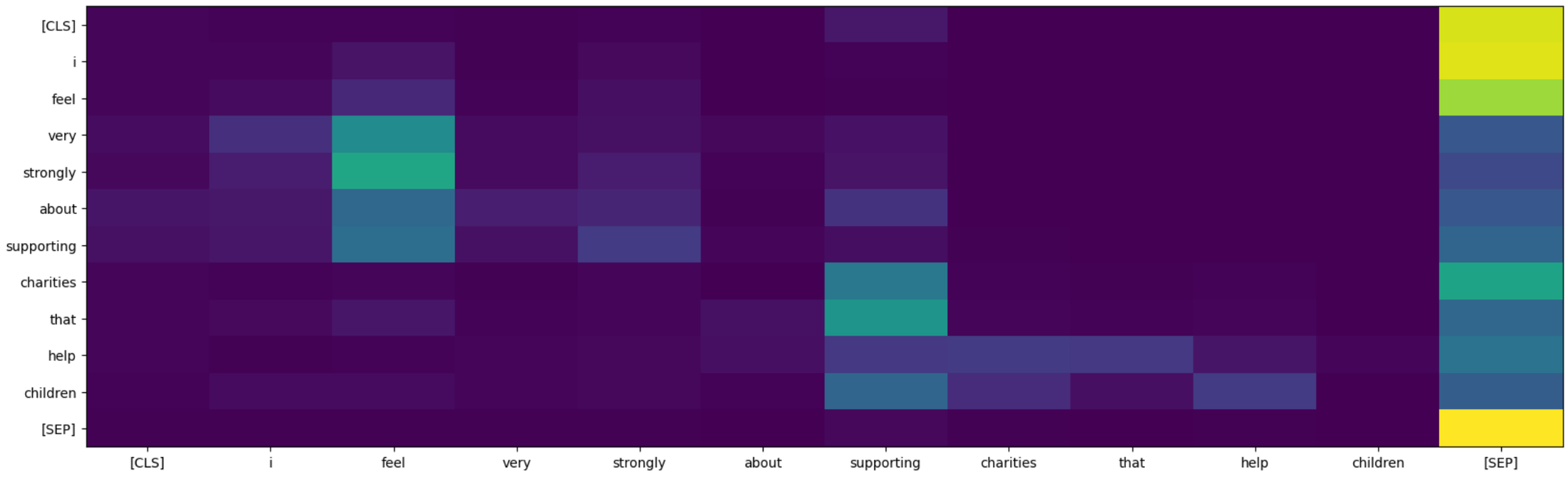

Figure 15: I feel very strongly about supporting charities that help children

8.4.4 Predicted: surprise, Truth: joy (Layer 11, Head 8)

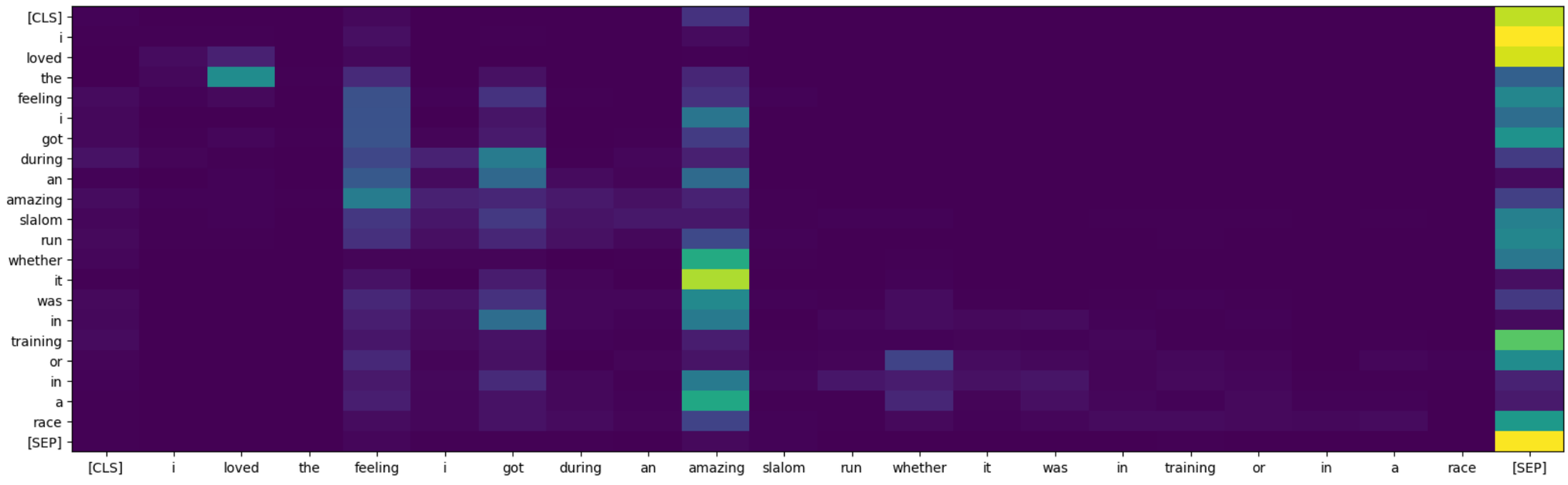

Figure 16: I loved the feeling I got during an amazing slalom run whether it was in training or in a race.