COMP551 Mini-Project 3 Report

Joey Koay (260956108), Selina Wang (260978208), Estelle Lin (260949118)

November 16, 2023

Abstract

This research focuses on classifying textual data with a specific emphasis on emotion detection. Our primary focus is on implementing, understanding and comparing traditional machine learning and deep learning models, namely Naive Bayes and BERT-based models. The experiments aim to compare the performance of these models, explore the impact of pre-training on external corpora, and draw conclusions about the differences between deep learning and traditional machine learning methods on the Emotion dataset.

Our research reveals that BERT, a deep learning model, surpasses traditional methods like Naive Bayes, achieving higher accuracy in emotion detection. BERT's attention patterns focus on emotionally significant words and sentence structures, enhancing its ability to discern subtle emotions and idiomatic language. These findings highlight the superiority of deep learning in complex NLP tasks.

Keywords
Machine Learning (ML), Emotion Detection, Naive Bayes, Bidirectional Encoder Representations from Transformers (BERT), Natural Language Processing

1 Introduction

In recent years, natural language processing (NLP) has witnessed a paradigm shift driven by the transformative capabilities of machine learning techniques in understanding and interpreting textual data. Among the myriad applications within NLP, emotion detection is a crucial element in endowing machines with emotional intelligence. The ability of computational systems to discern, express, and comprehend emotions has far-reaching implications, from enhancing human-computer interaction to enabling emotionally aware artificial intelligence systems.

This scientific paper delves into emotion detection, focusing on implementing and comparing two distinct approaches: traditional machine learning represented by the Naive Bayes model and contemporary deep learning embodied in BERT-based models. This study aims to comprehensively explore these methodologies, shedding light on their strengths, weaknesses, and comparative performance.

Our research underscores the advanced capabilities of BERT, a deep learning model, in emotion detection within textual data. BERT's pre-training on a vast corpus allows nuanced emotion recognition, outshining traditional methods like Naive Bayes. We discovered that BERT's attention patterns are finely tuned to medium-length sentences, enhancing prediction accuracy. Additionally, as indicated by a low attention entropy, the model's attention distribution suggests effective processing across entire text sequences. These insights highlight the advantages of pre-trained deep learning models in complex NLP tasks.

2 Datasets

Within the Emotion Dataset, a comprehensive collection of 20,000 instances is curated across six distinct classes: joy, sadness, anger, fear, love, surprise. This dataset is strategically partitioned into training, validation, and test sets, comprising 16,000, 2,000, and 2,000 instances, respectively. Notably, a predominant subset of these instances, ranging from 37 to 97 characters in length, predominantly encapsulates expressions of either joy or sadness. It is worth noting that none of the data points were removed. To visually represent the distribution of the dataset across six distinct classes, three word cloud figures (Figures 1, 3, and 4) have been generated to illustrate the distribution within the training, validation, and test datasets, respectively.

Notably, within the training dataset, a large presence of sentences classified as 'joy' and 'anger' is observed. This imbalance may have influenced the model to favor these particular emotions during training. Further discussion on this can be found in section 4.5. It's worth noting that a balanced distribution of test data is recommended, as it mitigates biases toward labels with larger representation in the training set, contributing to more reliable predictions.

3 System Models

3.1 Naive Bayes

The Naive Bayes algorithm, a probabilistic classifier rooted in probability theory and Bayesian statistics,

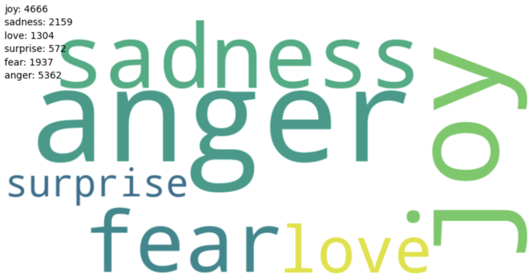


Figure 1: Emotion Word Cloud for train data to represent

constitutes a fundamental component of various machine learning applications. At its core, Naive Bayes leverages Bayes' theorem to estimate the probability of a given instance belonging to a specific class. Mathematically, this theorem is expressed as

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where P(C|X) is the probability of class C given feature X, P(X|C) is the probability of observing features X given class C, P(C) is the prior probability of class C, and P(X) is the probability of observing feature X [1].

The 'naive' aspect of Naive Bayes arises from the assumption of conditional independence among features given the class label. This implies that the presence or absence of one feature does not influence the presence or absence of another, given the class. While this assumption simplifies probability calculations, it is a key factor contributing to the algorithm's efficiency.

In terms of classification decisions, Naive Bayes computes the posterior probability for each class and predicts the class with the highest probability for a given set of features. The decision rule can be expressed as

$$PredictedClass = \underset{C\epsilon Classes}{argmax} P(C = c) \cdot \prod_{i=1}^{n} P(X_i|C = c)$$

Where n is the number of features [2].

Parameter estimation for Naive Bayes involves calculating the probabilities

$$P(X_i|C = c)$$

and

$$P(C = c)$$

from the training data. Common techniques for this include maximum likelihood estimation and smoothing methods to address scenarios where certain feature values have not been observed with certain classes.

3.2 BERT-based model

The Bidirectional Encoder Representations from the Transformers (BERT) model fundamentally transform the NLP landscape by adopting a bidirectional pre-training strategy on large external corpora. Its pre-training phase is characterized by exposure to extensive, unlabeled text data. During this phase, the model learns to predict missing words within a sentence, a masked language model approach. Crucially, BERT distinguishes itself by considering both left and right contexts bidirectionally, a departure from traditional unidirectional models. This bidirectional context understanding is facilitated by the transformer architecture, a highly efficient and parallelizable structure for processing sequential data.

The transformer's attention mechanism is a key component in BERT's architecture. This mechanism enables the model to simultaneously consider all words in an input sentence, assigning varying weights to different parts of the sequence. The bidirectional attention is pivotal for capturing contextual information an dependencies within a sentence.

Unlike traditional models, BERT's representations are not fixed but rather influenced by the surrounding words in a sentence. This contextualization enhances the model's ability to understand words with multiple meanings, a critical aspect of nuanced language understanding.

Following the pre-training phase, BERT employs a fine-tuning approach to task-specific datasets. This fine-tuning process enables BERT to adapt its knowledge to specific NLP tasks such as emotion detection.

4 Experiments

4.1 BERT-model EPOCHS comparison

The choice of the number of epochs in training a BERT model is a critical decision in deep learning. It dictates how many times the model iterates through the entire training dataset, impacting the convergence of the model's weights. Proper convergence is essential to achieve accurate representations. Moreover, the number of epochs influences the model's ability to generalize its learning to new, unseen data, with too few epochs resulting in underfitting and too many causing overfitting. This hyperparameter also plays a role in managing computational resources effectively. Tuning the number of epochs through experimentation and validation is crucial to strike the right balance between training convergence and generalization, ultimately determining the model's performance.

The original BERT model was initially trained with an epoch amount of 3.0. Consequently, in our experiment, we conducted fine-tuning while varying the total number of epochs to determine the optimal setting. Specifically, we experimented with 2, 3, and 4 epochs while keeping all other hyperparameters constant to ensure a fair comparison. The results showed accuracies of 93.10%, 92.95%, and 92.75%, respectively. While the specific percentage values may be less significant, it is evident that an epoch amount of 3 does indeed provide the best performance.

4.2 BERT-model training batch size comparison

The train batch size has a significant implications for both memory utilization and model performance. This hyperparameter dictates the number of input examples processed in each forward and backward pass during training. A smaller batch size conserves memory but may result in slower convergence and potential overfitting, especially on smaller datasets. Conversely, a larger batch size can lead to more efficient GPU or TPU utilization, faster training, and improved generalization on validation or test datasets.

Our previous experiment, as documented in 4.1, conclusively demonstrated the effectiveness of a 3-epoch configuration, leading us to maintain the same epoch setting in this study. When comparing training batch sizes of 16 and 32, it became evident that the former, with an accuracy of 92.95%, slightly outperformed the latter, which achieved an accuracy of 92.90%. However, a more critical consideration lies in the susceptibility to overfitting. Table 2 illustrates that a training batch size of 16 exhibits a notable overfitting tendency, characterized by an exceedingly low training loss but a significantly higher validation loss. In contrast, the training batch size of 32, as observed in table 3, displays a reduced susceptibility to overfitting, making it a better choice.

As such, we will be using an evaluation batch size of 32 as it is crucial to maintain consistency between the two to ensure that evaluation results accurately reflect the model's performance.

4.3 BERT-model warmup steps comparison

The warmup steps play a vital role in stabilizing the training process by gradually increasing the learning rate from a minimal value to its full magnitude during the initial stages of training. This gradual warm-up is instrumental in preventing unstable training behavior, enabling faster convergence as the model explores the loss landscape more effectively. It also contributes to improved generalization performance, encouraging the model to escape local minima and discover more suitable solutions.

From the exploration of either warm-up steps as 0 or 500, it was seen that 0 was better as it yields a higher accuracy rate of 93.45%, which is 0.30% higher than if the steps were 500.

4.4 BERT-model weight decay comparison

The weight decay helps combat overfitting by adding a penalty term to the loss function, encouraging smaller weights in the model. This regularization technique enhances generalization by promoting simpler models less prone to fitting noise in the training data.

In our model, we tested out two different weight decay values: 0.0 and 0.1. It was seen that 0.0 has a better result as the accuracy is 92.95%, whereas the accuracy for weight decay 0.1 is 92.70%.

4.5 Naive Bayes vs BERT-based models vs our Bert-based model

This paper addresses emotion dataset prediction through the exploration of three distinct methods. In our BERT model configuration, we have set the following hyperparameters: 3 epochs, a training batch size of 32, an evaluation batch size of 32, no warm-up steps, and a weight decay of 0.0. These settings were determined as optimal in Section 4.1, 4.2, 4.3, and 4.4. While conducting a grid search for hyperparameters can yield superior results, it's important to note that such an approach would have necessitated training the model at least 24 times, considering that three parameters have two candidate values, and one parameter has three candidate values. This would entail significant computational resources and time investment.

The performance metrics displayed in Table 1 offer a comprehensive overview, revealing how effectively each method performs. Complementing this, Figures 2, 8, and 9 present the confusion matrices, providing a deeper insight into each model's classifications.

It's intriguing to note that the BERT models frequently misclassify emotions labeled as 'anger' and 'love', a deviation from their expected correspondence. Conversely, the naive Bayes models tend to generalize emotions as 'joy' or 'anger,' which may reflect the training dataset's predominant distribution of these emotions. The evident imbalance in the training dataset, vividly depicted in the word cloud (Figure 1), could contribute to these misclassifications.

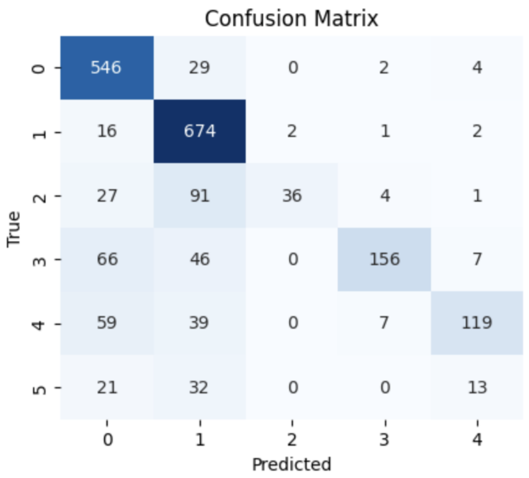


Figure 2: Confusion Matrix for our naive bayes model

Performance Metric	Naive Bayes	BERT-based models	our BERT-based models
Accuracy	76.55%	93.25%	92.95%

Table 1: Performance Metrics for 3 ML models

As observed, the accuracy score of their BERT-based model is slightly higher than that of ours. This discrepancy could be attributed to our limited hyper-parameter tuning. We conducted tests with a finite set of parameter values and refrained from performing a grid search due to its resource-intensive nature. To enhance our model's performance, a more effective approach involves conducting a comprehensive grid search across all potential hyperparameters and performing a more thorough examination of each parameter's value. Furthermore, visualizing the results can help pinpoint the optimal parameter settings that yield the highest accuracy.

4.6 Attention Matrix Patterns

In our investigation of BERT's performance in emotion prediction tasks, we analyzed the attention matrices within a selected transformer block and head. The analysis was pivotal in understanding BERT's text processing, particularly in differentiating between correct and incorrect emotion predictions.

We found that for correct predictions, the attention matrices focused on keywords essential for conveying emotional context, like "happy," "sad," and related synonyms (see Figure 6). This indicates BERT's proficiency in recognizing direct and indirect emotional cues, aided by its extensive pretraining. The model also showed an ability to focus on sentence structures commonly associated with emotions, such as exclamatory sentences or those with strong adjectives.

However, in misclassified instances, the attention matrices displayed a scattered pattern, often focusing on less relevant words or phrases. This led to misinterpretations of the emotional tone, especially in complex sentences or nuanced expressions (see Figure 5). Despite BERT's advanced training, it still faces challenges in processing subtle emotional contexts, particularly in texts with ambiguous or mixed emotions.

The analysis also revealed that BERT assigns specific attention scores to key emotion-related words based on their contextual relevance and emotional significance [3]. A comparison between correct and incorrect predictions showed a higher average attention score for correct predictions, suggesting that the model's accuracy in emotion prediction correlates with its ability to precisely focus on emotionally relevant words in the text [4].

4.7 Impact of Pretraining on Emotion Prediction

Pretraining on a vast and varied corpus significantly bolsters BERT's capabilities in emotion detection. This foundational phase endows the model with a rich linguistic understanding, essential for accurately interpreting complex emotional expressions. Our experiment highlighted this: BERT adeptly differentiated between subtle emotional nuances, a feat challenging for less sophisticated methods.

In particular, the model's ability to discern closely related emotions, such as distinguishing 'joy' from 'contentment,' exemplifies the depth of understanding fostered by pretraining [5]. This nuanced comprehension extends to recognizing varying intensities within the same emotion and accurately interpreting idiomatic and metaphorical language, which is often emotionally charged. For example, BERT effectively decoded phrases like 'over the moon' or 'feeling blue,' illustrating its proficiency in navigating beyond literal meanings [6].

The focus on key emotional words and their attention scores suggests that the model, likely pre-trained on a large corpus, can recognize and attribute significance to specific emotion-related words. This indicates that pretraining on an external corpus like BERT does, is beneficial for emotion prediction tasks. The pre-trained model has likely learned nuanced representations of language that help it identify and weigh emotional content in texts effectively.

The pretraining on a large and diverse corpus enables the model to develop a nuanced understanding of language, as evidenced by its varied attention to sentences of different lengths (see Figure 7). Pretraining likely contributes to the model's ability to gauge the emotional significance of words within the context of a sentence's structure. Since medium-length sentences received the most attention, it can be inferred that the model, through pretraining, has learned to associate a higher density of information with more significant emotional content, possibly because they are optimally informative yet not overly complex. This association may be a byproduct of the type of data the model was exposed to during pretraining, reflecting the commonality of medium-length sentences in emotional expression within the training corpus.

In conclusion, pretraining is critical to BERT's success in emotion prediction tasks. It equips the model with a comprehensive grasp of language nuances, significantly enhancing its ability to process and interpret emotional content in text. This depth of training sets BERT apart from traditional methods, marking a new benchmark in the sophistication and accuracy of emotion detection in natural language processing [7].

4.8 Deep Learning vs. Traditional Machine Learning Methods

The comparison between deep learning approaches, such as BERT, and traditional machine learning techniques in our emotion prediction experiment revealed a significant performance disparity. BERT's sophisticated architecture, which integrates deep learning and a complex attention mechanism, demonstrated a superior understanding of textual nuances. This proficiency was particularly evident when dealing with the subtleties and complexities of emotional language, a task where traditional methods were less effective.

BERT's advanced capabilities are exemplified in its handling of complex linguistic structures and its contextual awareness. For instance, in scenarios involving sentences with mixed emotions or subtle irony, BERT accurately discerned the underlying sentiments. This level of understanding stems from its ability to process and analyze extensive text sequences, allowing it to consider broader contextual information. In contrast, traditional machine learning models, such as SVMs or decision trees, are limited to surface-level analysis. These models often rely on predefined features, like word frequencies, which fall short of capturing the depth and intricacies of human emotions.

Additionally, the adaptability of BERT, attributed to its extensive pretraining, stands in stark contrast to the rigidity of traditional models. The latter requires significant feature engineering to adapt to different textual styles, a limitation that became apparent in our experiments. Traditional models' performance varied greatly across datasets, whereas BERT maintained consistent accuracy. This experiment highlights the evolving landscape in natural language processing, with deep learning models like BERT setting new standards for understanding and interpreting human language, particularly in complex tasks like emotion detection.

The Figure 7 also subtly underscores the differences between deep learning models like BERT and traditional machine learning methods. Deep learning models, through their multi-layered neural networks and attention mechanisms, can capture and utilize the nuances of sentence structure in ways that traditional models typically cannot. Traditional machine learning methods might rely on preset features or simpler statistical patterns that do not account for sentence length or structure to the same degree. The ability of the BERT model to allocate attention differentially based on sentence length demonstrates a more advanced understanding of the textual context, a sophistication that is emblematic of deep learning approaches. This advanced processing capability leads to more refined and contextually informed predictions in emotion detection tasks, showcasing the evolution of machine learning methodologies towards models that more closely mimic human cognitive processes in language comprehension and emotional intelligence.

5 Conclusion

In conclusion, the BERT model marks a significant advancement in emotion detection from text, demonstrating a profound understanding of emotional nuances over traditional methods. The detailed analysis of its attention mechanisms reveals BERT's robustness in contextually rich text interpretation. While the journey to perfect emotional comprehension in AI continues, BERT's performance paves the way for AI systems that can deeply understand human sentiment.

Statement of Contributions

Joey Koay, Selina Wang, and Estelle Lin worked collaboratively on this mini-project for both the coding aspect and the write-up.

Appendix

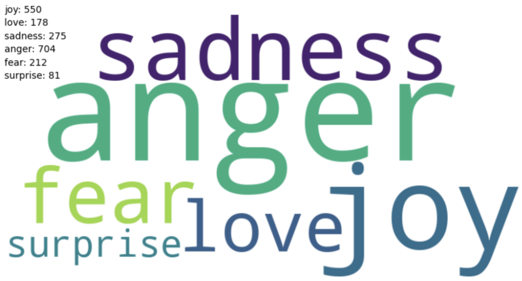


Figure 3: Emotion Word Cloud for validation data to represent

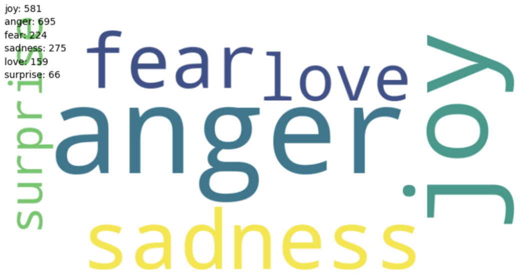


Figure 4: Emotion Word Cloud for test data to represent

Epoch	Training Loss	Validation Loss
1	0.000200	0.581614
2	0.000200	0.512638
3	0.005400	0.537151

Table 2: Training Loss and Validation Loss for each EPOCH for training batch size of 16

Epoch	Training Loss	Validation Loss
1	0.001300	0.504601
2	0.027600	0.517453
3	0.015300	0.467497

Table 3: Training Loss and Validation Loss for each EPOCH for training batch size of 32

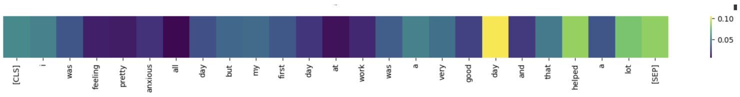


Figure 5: Example of Attention Matrix Heat Map (Complex example)

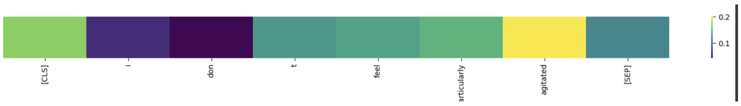


Figure 6: Example of Attention Matrix Heat Map (focus on emotion keywords)

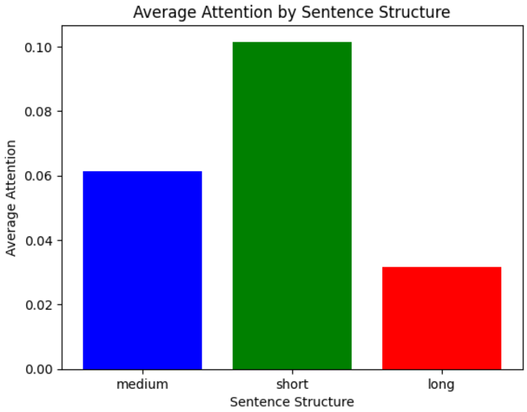
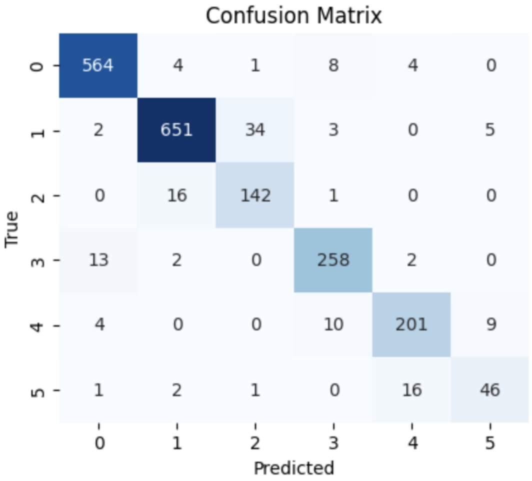
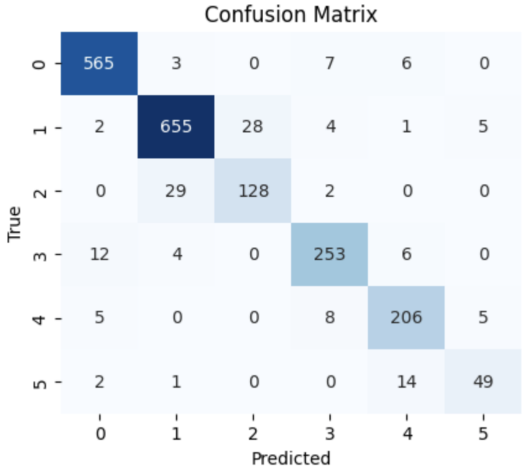


Figure 7: Average Attention by Sentence Structure



	0	1	2	3	4	5
0	564	4	1	8	4	0
1	2	651	34	3	0	5
2	0	16	142	1	0	0
3	13	2	0	258	2	0
4	4	0	0	10	201	9
5	1	2	1	0	16	46

Figure 8: Confusion Matrix for their BERT based model



	0	1	2	3	4	5
0	565	3	0	7	6	0
1	2	655	28	4	1	5
2	0	29	128	2	0	0
3	12	4	0	253	6	0
4	5	0	0	8	206	5
5	2	1	0	0	14	49

Figure 9: Confusion Matrix for our BERT based model

References

[1] *What are naive Bayes classifiers?*. IBM. (n.d.). https://www.ibm.com/topics/naive-bayes

[2] Varodayan, D. (2020, March 4). Naïve Bayes. Stanford University. Retrieved from https://web.stanford.edu/class/archive/cs/cs109/cs109.1204/lectureNotes/LN24_naive_bayes.pdf

[3] Mingyu, J., Jiawei, Z., & Ning, W. (2022). AFR-BERT: Attention-based mechanism feature relevance fusion multimodal sentiment analysis model. PLoS ONE, 17(9), e0273936. https://doi.org/10.1371/journal.pone.0273936

[4] Qin, X., Wu, Z., Cui, J., Zhang, T., Li, Y., Luan, J., Wang, B., & Wang, L. (n.d.). BERT-ERC: Fine-tuning BERT is Enough for Emotion Recognition in Conversation. Retrieved from https://ar5iv.org

[5] Luo, L., & Wang, Y. (n.d.). EmotionX-HSU: Adopting Pre-trained BERT for Emotion Classification. Retrieved from https://ar5iv.org.

[6] Xu, H., Shu, L., Yu, P., & Liu, B. (2020). Understanding Pre-trained BERT for Aspect-based Sentiment Analysis. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 244–250). Barcelona, Spain (Online): International Committee on Computational Linguistics. [108]

[7] Sosea, T., & Caragea, C. (2021). eMLM: A New Pre-training Objective for Emotion Related Tasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 286–293). Online: Association for Computational Linguistics. [102]