COMP 551 Project 3 Report: *Classification of Textual Data*

Students: Albert Caupin, Max-Henri Chanut and Christine Pan

Due Date: November 16th, 2023

Abstract

In this project, we were tasked the implement the Naive Bayes model from scratch and implement the BERT-based model with pertained weights through package and finetuning. The dataset used in this project is the "Emotion" dataset. Furthermore, we ran experiments on both models with this dataset, analyzed the results and compared the performances of the models, gaining us more experience implementing machine learning algorithms either from scratch or using deep learning libraries.

Introduction

The goal of this project is to implement the Naive Bayes model and the BERT model with which we train and test on a dataset, called "Emotion", which consists of English Twitter messages with six basic emotions. After acquiring the dataset, and processed and normalizing the data, we implemented the Naive Bayes model and the BERT model. The Naive Bayes model was implemented with the Multinomial distribution with no libraries, as it is the appropriate type of likelihood for features. On the other hand, the BERT based model was implemented using existing pre-trained BERT models online. We used the "bert-base-uncased" (Savani) model, which is a model with pre-trained weights with a package. We then proceeded to experiment these models to obtain performance result of the Naive Bayes, BERT-based models and our own BERT-based model on the Emotion dataset, and to analyze which model performed better. We then examined the attention matrix between the words and the class tokens of some correctly and incorrectly predicted documents. These experiments make us analyze and realize some conclusions about performances in deep learning and traditional machine learning, as well as how pretraining and finetuning are significant in a model's performance on a dataset.

System Models

In this project, two distinct system models were implemented to predict and classify emotion in English Twitter messages. The first model implemented is the Naive Bayes, a foundational probabilistic model grounded in Bayes' theorem. Specifically, the Multinomial Naive Bayes was implemented, since the Multinomial distribution is the appropriate likelihood for feature representation. This model, implemented from scratch , utilizes the assumption of feature independence to make predictions based on the probability density function of each class. The second system model introduced is the BERT-based model, a deep learning model for natural language understanding and sentiment analysis. The BERT model which stands for for Bidirectional Encoder Representations from Transformers, has demonstrated great performance in various NLP tasks due to the fact that it is pre-trained. For this project, a pre-trained BERT model, specifically the "bert-base-uncased-emotion" by Bhadresh Savani, was used. This model has been finetuned on emotion prediction tasks, making it well-suited for our target application. In addition to implementing the pretrained BERT model, an experiment on finetuning strategies was conducted. This involved experimenting with the finetuning of all layers of the BERT model versus fine-tuning only the last few layers. The objective was to discern the impact of different finetuning approaches on the model's performance and to provide insights into the advantages and disadvantages of each strategy.

Datasets

The dataset is "Emotion", which is a dataset (by DAIR.AI) of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness and surprise. Thus, the data field is a text, which is string feature, and the label which is the target value, is a classificaiton label, with possible values of 0 (sadness), 1 (joy), 2 (love), 3 (anger), 4 (fear) and 5 (surprise). This dataset is already split into "train", "validation" and "test", and those splittings are what we are using in this project and what we are reporting on. There are 16 000 data in the training set, 2000 data in the validation set and 2000 data in the test set, making it a total of 20 000 examples.

Preprocessing

In this section, we describe the data preprocessing steps undertaken for the Naive Bayes and BERT-based models in the context of emotion detection.

Naive Bayes Data Preprocessing

For the Naive Bayes model, the following steps outline the preprocessing methodology:

The Emotion dataset was loaded using the Hugging Face `load_dataset` function, and the data was split. The scikit-learn `CountVectorizer` was then utilized to fit transform the train dataset and transform the test dataset.

The resulting datasets served as the input features for training and testing the Naive Bayes model.

BERT Data Preprocessing

For the BERT-based model, the Hugging Face Transformers library was employed, utilizing the pre-trained model `bhadresh-savani/bert-base-uncased-emotion`. The following steps outline the preprocessing for BERT:

A tokenization function was defined to tokenize the text data using BERT tokenizer. The tokenization function was applied to the datasets using the `map` function.
A `DataCollatorWithPadding` was utilized for padding tokenized sequences to a common length, ensuring consistent input shapes.
The tokenized datasets were converted to TensorFlow datasets using the `to_tf_dataset` function, specifying the necessary columns, label columns, shuffling, collate function, and batch size for both the training, testing, and validation sets.

The resulting TensorFlow datasets (`tf_train_dataset`, `tf_test_dataset`, and `tf_validation_dataset`) contain tokenized and padded input features, ready for training and evaluating the BERT-based emotion detection model.

Experiment Settings

In this section, we detail the experimental setups conducted to evaluate the performance of Naive Bayes, BERT, and fine-tuned BERT models for emotion detection.

Naive Bayes

The Naive Bayes implemented is more precisely a Multinomial Naive Bayes model. The dataset in this project consists of discrete features that represent the words count and tokens of texts (after preprocessing the data such as vectorization and tokenization). Thus, Multinomial distribution is appropriate likelihood for this dataset's features since it is well suited for features that are counts or frequency. To explore the impact of the Laplace smoothing parameter (`myalpha`) on model performance for the Naive Bayes model, multiple experiments were conducted. The following accuracy results were obtained:

Experiment	Accuracy
myalpha=1	0.7915
myalpha=0.7	0.796
myalpha=0.6	0.8015
myalpha=0.5	0.7975

Table 1: Experiment Results

Based on these results, we selected myalpha=0.6 as the optimal Laplace smoothing parameter for the Naive Bayes model. This experiment aimed to determine the most effective smoothing parameter, considering its impact on accuracy in emotion detection.

BERT Model

The BERT model was fine-tuned for emotion detection using the Hugging Face Transformers library with the pre-trained model bhadresh-savani/bert-base-uncased-emotion. The default configuration of the BERT model was retained during fine-tuning, maintaining the following settings:

Parameter	Value
config.num_hidden_layers	12
config.num_attention_heads	12
config.hidden_dropout_prob	0.1
config.attention_probs_dropout_prob	0.1
config.hidden_size	768
config.intermediate_size	3072

Table 2: Default Configuration Parameters

Fine-Tuned BERT Models

To explore the impact of varying model complexity on emotion detection, we conducted fine-tuning experiments with modified configurations. Three different settings were explored:

1. **Double Parameters:** The number of hidden layers, attention heads, hidden dropout probability, attention dropout probability, hidden size, and intermediate size were doubled during fine-tuning.

2. **Half Parameters:** The number of hidden layers, attention heads, hidden dropout probability, attention dropout probability, hidden size, and intermediate size were halved during fine-tuning.

These experiments aim to investigate how variations in the BERT model's architecture and complexity impact its performance in emotion detection.

Results

In this section, we present the performance comparison of the Naive Bayes model, the BERT-based model, and the fine-tuned BERT models on the Emotion classification task. The table below summarizes the accuracy results:

Model	Accuracy
Naive Bayes (myalpha=0.6)	0.8015
BERT (Default Parameters)	0.9243
Fine-tuned BERT (Halfed Parameters)	**0.9345**

Table 3: Performance Comparison on Emotion Classification

The results indicate that the fine-tuned BERT model with halfed parameters achieved the highest accuracy (**0.9345**), outperforming Naive Bayes and the default BERT model. The potential reasons for this performance difference could be attributed to the increased model complexity, allowing the fine-tuned BERT model to capture more intricate patterns in the emotional text data. However, it's essential to note that higher model complexity may come at the cost of increased computational requirements. Here is the comparison of the BERT performance versus the finetuned BERT performance:

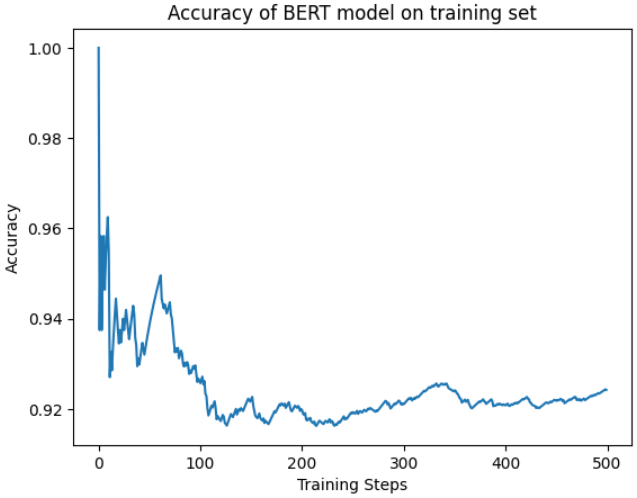

Figure 1: Performance of BERT model

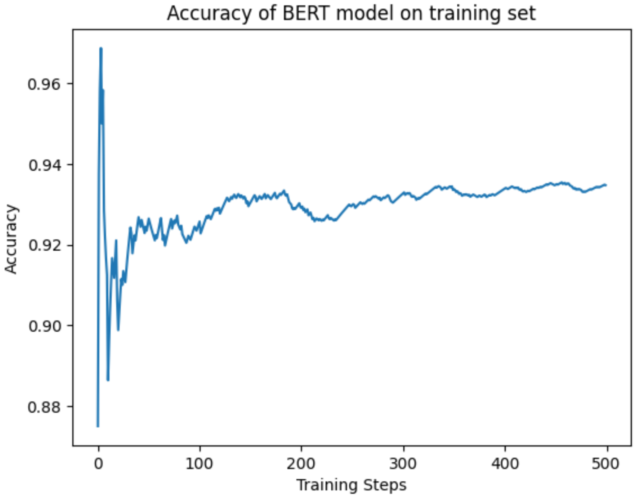

Figure 2: Performance of finetuned BERT model

Attention matrix

Below are attention matrices of our BERT model for some accurately and inaccurately labeled documents.

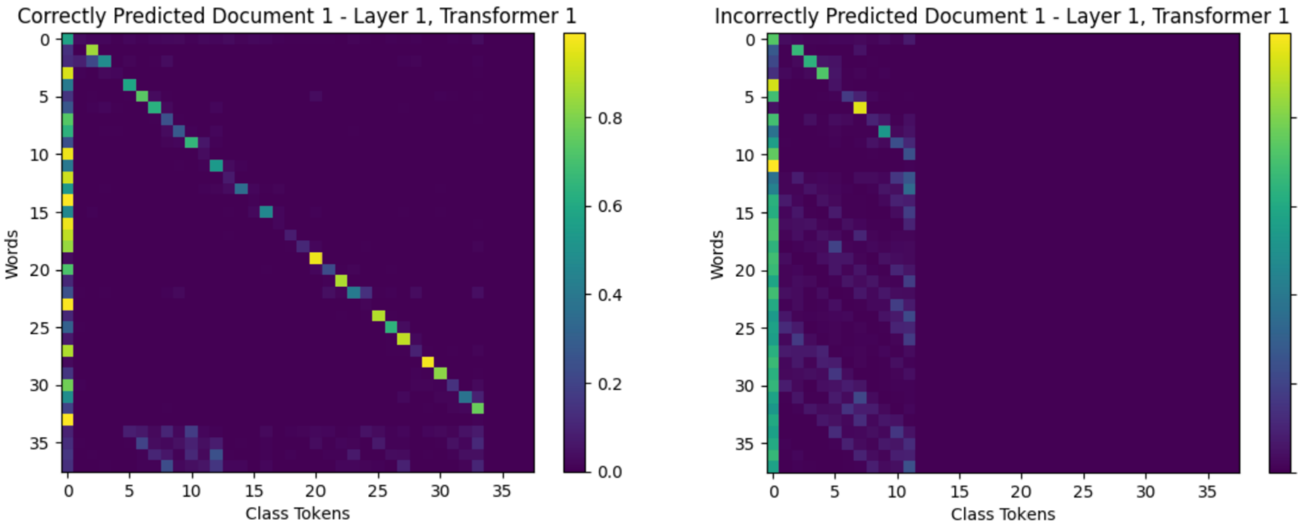

Figure 3: Attention matrices for correct and incorrect predictions

We can observe on the first attention map that there is a strong regular pattern. On the second one however, the model failed to compute a meaningful attention pattern and it led to a wrong prediction for the given document. Note that those attention matrices were used in the first layer of the transformer, which mean that they are more fitted to detect global patterns on the whole document rather than more localized ones.

Pre-training a large model allows us to gain a lot of time on the training by only needing to do the fine tuning. It is especially useful for emotion classification because it is part of a more general application which is natural language processing. Training on a more general dataset allows the model to learn wide variety of features of the natural language, learn how to analyse context and that can then be used as a base for more specific application such as emotion prediction. It also decreases the need for a large dataset during the finetuning part of the training, which is especially useful for emotion prediction since the data needs to be labeled, and thus is more costly to produce/obtain.

Deep learning algorithm are more efficient than traditional machine learning algorithms for complex tasks because the larger amount of parameters allow them to find more complex patterns and as such are more adaptive to new data. They come however at a greater computation cost.

Discussion and Conclusion

In this project, we explored both traditional machine learning and deep learning methodologies on emotion prediction in English Twitter messages. Two models, Naive Bayes and BERT-based models, were implemented, suggesting different strengths and weaknesses.
The Naive Bayes model, specifically the Multinomial Naive Bayes, demonstrated a reasonable performance with an accuracy of 80.15%. This traditional model is proved effective in predicting patterns within the discrete features of the Emotion dataset.

On the other hand, the BERT-based model, which is a pre-trained BERT model showed a much higher accuracy of 92.43% on emotion prediction tasks with its default configurations. Further experimentation we did was finetuning some of those configurations, such as the number of hidden layers, number of attention heads, the hidden dropout rate among others, suggested that a finetuned BERT model with halved parameters achieved the highest accuracy at 93.45

Through our experimentation we also measured and concluded that the attention matrix is a very powerful tool that allows models such as BERT to not only make prediction based on the content of each token individually but rather use the relationship between the variables to gather additional information about the structure and make predictions much more accurate, while also being able to handle more complex inputs successfully.

This project suggests that deep learning has a clear advantage over traditional machine learning with their accuracy rate. The experiments showed that Naive Bayes had limitations in predicting intricate patterns in the context of the Emotion dataset, while the BERT model has proven capable of understanding more complex nuances and more intricate patterns that is classifying emotion. Moreover, with finetuning a deep machine learning model like BERT showed that adjusting its architecture, configurations and complexity can lead to a greater performance accuracy, suggesting that deep learning models can be furthermore improved with customization and finetuning.

In conclusion, this project provided valuable insights on the effectiveness of both traditional and deep learning approaches for emotion prediction in English Twitter messages. The results highlighted the influence between model complexity, pre-training, and finetuning in achieving optimal performance. Future work could explore more advanced pre-trained models, experiment with different architectures, and go deeper into attention mechanisms to improve emotion prediction accuracy further.

Statement of Contributions

Max-Henri : Task 1, Task 2, write-up, Christine : Task 1, Task 2, write up, Albert : Task 3, write up