Maching Learning - Mini Project 2

Thinakone Louangdy

February 2024

1 Introduction

Given a set of data of tweets from a certain banking details and features, we want to be able to classified the sentimental analysis. The data that we will used is publicly available data set created by three graduate students at Stanford University which comprises approximately 1.600.000 annotated tweets.

However, why resolve this issue? The NLP or Natural Language Processing has been growing popular day by day in our eras. Our objective for this project is to be delve into the used of machine learning to textual context of the real world data.

We will employ machine learning techniques to find a solution to this issue. We were asked to select two machine learning models that we have learn in this course or outside to perform a classification on sentimental analysis.

We will primarily try to test our model in linear regression family and the most popular one from transformers.

2 Data

We have a dataset that contains various only 10% of the whole dataset that has already been labeled with sentimental label. Our data set contains label and text-data and special character as you can see at 1.

	sentiment_label	tweet_text
0	4	@elephantbird Hey dear, Happy Friday to You A...
1	4	Ughhh layin downnnn Waiting for zeina to co...
2	0	@greeniebach I reckon he'll play, even if he's...
3	0	@vaLewee I know! Saw it on the news!
4	0	very sad that http://www.fabchannel.com/ has c...

Figure 1: Target Distribution

The objective for this task is to be able to convert textual data into the meaning numerical features for our machine learning model. Luckily, our data has a balance dataset of two classes of positive(4) and negative(0) as in figure 2

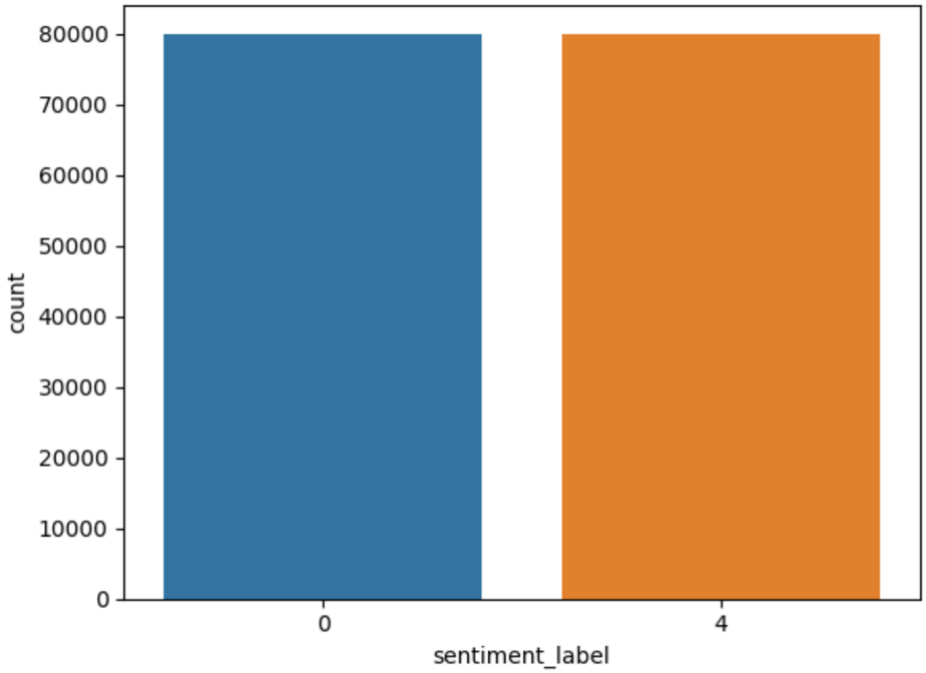


Figure 2: Target Distribution

3 Data Preprocessing

Our data set contains textual data from tweets but also already has labeled of sentiment, this could lesser our work task but the majority task were still remaining. We will handled textual data by utilizing the regex patterns which will captures the matching patterns in text then we will remove it for further data pre-processing. As we can see that there were some useless characters in out data such as dot(.), comma(,), and other special characters as well as the username from tweets, we won't need that for our model.

After that we will continued using a punkt package from **nltk** library to create a tokenization for our text, we also use stopwords and wordnet to perform lemmatization to change all the text to lowercase and remove stop word from our text data. We have also converted the label of sentiment_label to binary format (0 and 1) due to our model will perform a binary classification.

4 Modeling

For this classification problem, I selected logistic regression from linear classification model and BERT from the transformer family in order to take advantage of their complementing qualities.

- Model 1: Logistic Regression
  - Among the standard algorithms for linear classification, logistic regression stands out as being especially useful for text-based data. Its ease of use frequently serves as the standard by which other, more intricate models are evaluated.

  - To represent tweets as numerical vectors, TF-IDF (Term Frequency-Inverse Document Frequency) was utilized, which weighted words according to their significance.

- Model 2: BERT (Bidirectional Encoder Representations from Transformers)
  BERT is a powerful pre-trained language model that excels in understanding the nuances and context of language. The 'bert-base-uncased' variant was fine-tuned for this sentiment classification task.

5 Model Evaluation

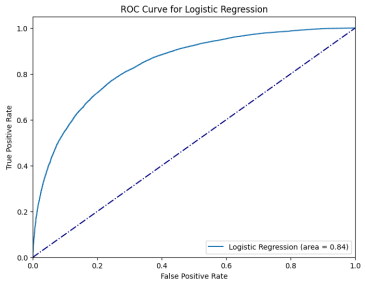
The BERT model marginally surpassed logistic regression in metrics such as accuracy, achieving 77.98%, compared to logistic regression's 76.27%. This modest improvement aligns with expectations, as BERT's design enables it to grasp more nuanced linguistic connections and contextual nuances.

The logistic regression model's ROC-AUC curve (see attached plot figure3) illustrates its performance in discriminating between positive and negative sentiment.

Machine Learning Model: Logistic Regression
Accuracy: 0.76278125
Classification Report:

	precision	recall	f1-score	support
0	0.77	0.74	0.76	16002
1	0.75	0.78	0.77	15998
accuracy			0.76	32000
macro avg	0.76	0.76	0.76	32000
weighted avg	0.76	0.76	0.76	32000

(a) Classification Report



(b) ROC Curve

Figure 3: Logistic Regression Performance Metrics

6 Conclusion

The tweet sentiment analysis demonstrates that, while pre-trained transformer models like BERT offer advanced capabilities in navigating the complexities of social media language, they also demand significantly more computing resources. Despite their potential for superior performance in classification tasks for text-data, the marginal improvement observed does not justify the increased computational cost under our current constraints. Unfortunately, our limited access to extensive computing resources restricts our ability to train these models further to realize their full potential. This underscores the importance of balancing the benefits of advanced NLP technologies with the practicalities of available resources.