Assignment

by Ahtasham Karim

Using BERT to Conduct Fine-grained Sentiment Analysis on Twitter Texts

SUBMITTED BY:	AHTSHAM KARIM	ID:	21031263

Introduction

In the middle of the worldwide Covid-19 epidemic, social media sites like Twitter, Facebook, and Instagram have grown in prominence as venues for people to express their real-time ideas and opinions. This research looks at the opinions posted on social media with a focus on the emotional nuanced reactions made by people Social media sites unintentionally reveal people's thoughts and nuanced emotional characteristics, which results in the categorization of individuals based on their distinctive viewpoints. This research uses Twitter as its main platform using sophisticated Natural Language Processing (NLP) and Machine Learning Bert Model approaches to evaluate real-time data from Twitter [1]. To assess the polarity of the text and classify sentiments as positive, neutral, or negative, this research uses sentiment analysis. The identification of public attitudes and feelings during the problems posed by the pandemic is made possible by this analytical technique, which provides insightful information about the overall mood. This study emphasizes both the value of natural language processing (NLP) in understanding human communication and the challenges of interpreting emotions in informal language. The research in this study employs advanced sentiment analysis techniques and Natural Language Processing to ascertain the general opinion about the outbreak. This activity greatly enhances our ability to understand public opinion amid the exceptional global crisis [2].

Purpose of Dataset

The dataset consists of tweets from Twitter that have been manually classified and labeled. The dataset is intended to perform tasks like as sentiment analysis, exposed classification, and related classification tasks on tweets in order to assess and classify them based on their content [3]. Table 2 provides the sentiment of all labels. The dataset's data occurrence plots are shown in Figure 2, and a review of word length is shown in Figure 3.

Exploratory analysis: It's crucial to do an exploratory study to thoroughly grasp the features of the dataset before beginning the text classification procedure. Analyzing label distribution is the first step in determining the balance between distinct categories. Frequency-based visualizations help assess the distribution of categorization categories and provide an extensive overview of the dataset's makeup. Figure 1 shows tweets and the emotions they express. The dataset division is shown in Table 1. The Sentiment scores and accompanying labels are shown in Table 2.

	Reviews	Sentiment
0	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	2
1	advice Talk to your neighbours family to excha...	3
2	Coronavirus Australia: Woolworths to give elde...	3
3	My food stock is not the only one which is emp...	3
4	Me, ready to go at supermarket during the #COV...	0

Fig.1 Shows Coronavirus Tweets Sentiments

Table.1 Shows Division of the Dataset

TRAIN: DATASET	NUM_ROWS:	37452
VAL: DATASET	NUM_ROWS	3705
TEST: DATASET	NUM_ROWS:	3798

Table 2. Shows Sentiment and Their Label

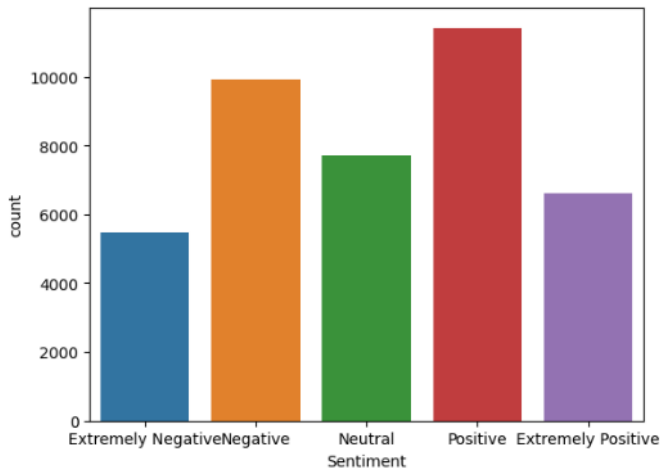| Extremely Negative	Negative	Natural	Positive	Extremely Positive
0,0	1,0	2,0	3,0	4,0

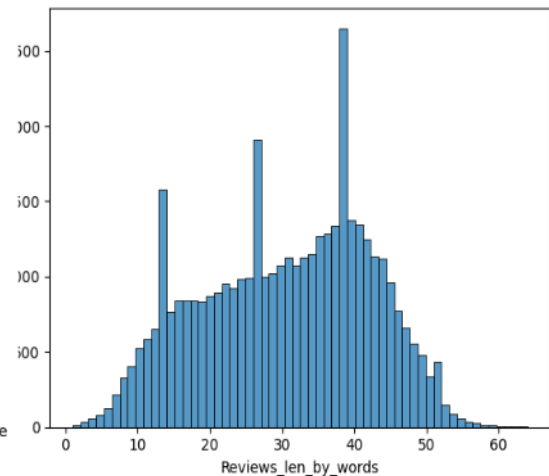Fig.2 Shows sentiment occurrence in the dataset

Fig.3 Shows the length of Tweets

Transformer Model

The Transformer model is a neural network architecture designed for processing sequential data, such as text and time-series information. This is achieved through the execution of parallel attention-based computations on input tokens, which effectively assess contextual information and interconnections. This architectural paradigm has greatly transformed machine translation and natural language comprehension, resulting in notable progress in these fields [4].

CUSTOM BERT MODEL

The `BertForClassification` model is derived from the BERT architecture in the Transformers library. The model is designed specifically for sequence classification tasks. The process commences by loading a BERT model, which is configured based on specific parameters [5]. A token classification head is created, consisting of a dropout layer and a linear classifier. During the forward pass, BERT processes the input text and utilizes the classifier on the contextual representation of the `[CLS]` token. A cross-entropy loss is calculated when labels are available. The model's output consists of classification logits, potential loss, hidden states extracted from BERT, and attention values. This code of assignment implements a BERT-based classification model for analyzing and categorizing text sequences.

AdamW Optimizer

AdamW is an enhanced version of the Adam optimizer that integrates weight decay into the optimization process to mitigate excessive parameter growth during training. This addresses a limitation in the Adam optimization algorithm, specifically the conflict between weight decay and the adaptive learning rate strategy. AdamW is a method that enhances the optimization process in neural network training by applying weight decay to parameters

individually. This approach leads to improved convergence and regularization, resulting in a more stable and effective optimization process [6].
.

Model Results

The model's evaluation shows promising results in classifying sentiments, with accurate precision, recall, and F1-score metrics reported for different sentiment classes. The 'BertForClassification' model, which is based on BERT, is effective in granular sentiment analysis and thorough text classification, as shown by its 70% accuracy are shown in Fig.4 and Fig.5 Shown confusion matrix. Which highlights its ability to classify a wide range of tweets. This establishes the model as an appropriate choice for sentiment analysis and similar classification tasks. Table 3. Showing training loss and validation loss. Fig. 6 model prediction.

Table 3: Shows Training loss and Val loss accuracy

Train loss	0.38689641446977757	Accuracy	0.9009131688561358
Val loss	1.156153346241777	Accuracy	0.732523616734143

```
                    precision   recall  f1-score   support

Extremely Negative     0.72       0.75     0.74        592
          Negative     0.70       0.63     0.67       1041
           Neutral     0.84       0.77     0.81        619
          Positive     0.68       0.57     0.62        947
Extremely Positive     0.61       0.90     0.73        599

          accuracy                         0.70       3798
         macro avg     0.71       0.73     0.71       3798
      weighted avg     0.71       0.70     0.70       3798
```
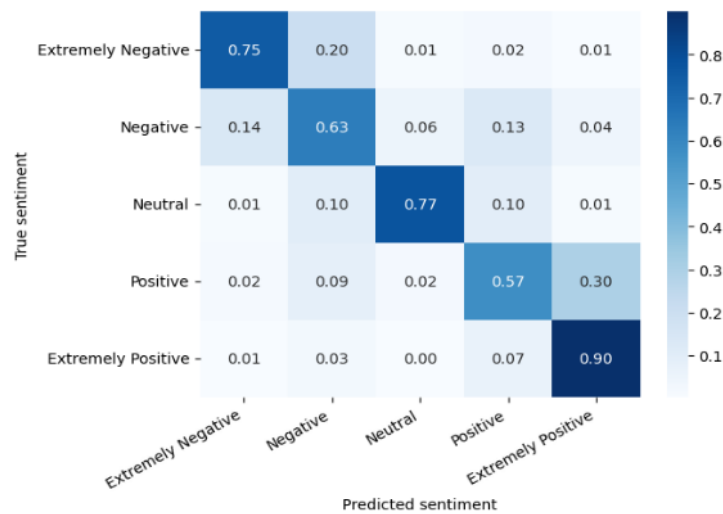
Fig.4. Shows different accuracy metrics

Fig.5. Shown confusion matrix

Fig.6. Shown model prediction on text

Google Colab Link:

https://colab.research.google.com/drive/1_By9R1RDGL7b8TsrBfKiR0l38RYQwds1?usp=sharing

GitHub Link:

References:

1) Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, A. (2020). Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. *J. Crit. Rev*, *7*(9), 2761-2774.
2) Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia* (pp. 212-216). IEEE.
3) https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification
4) Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908-15919.
5) Thapa, A. (2023). *Sentiment Analysis On Juvenile Delinquency Using BERT Embeddings* (Doctoral dissertation, Dublin, National College of Ireland).
6) Llugsi, R., El Yacoubi, S., Fontaine, A., & Lupera, P. (2021, October). Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural

Networks for the Andean city of Quito. In *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)* (pp. 1-6). IEEE.

Assignment