Phase-5-capstone-project

1) INTRODUCTION

This is the phase 5 capstone project repository.In this repository , we will be performing sentiment analysis on sentiments by Redda.it users.The datset is sourced from Kaggle.This project was done by a team of five people , namely :

1) Anthony Ngatia.
2) Elsie Nduta.
3) Jessyca Aperi.
4) Joy Kipkemboi.
5) Naomi Rotich.

2) BUSSINESS UNDERSTANDING.

2.1) Business overview.

Getting feedback from application users is a crucial aspect of growth as it gives a deeper understanding of user sentiment, improves content moderation, and informs product and service improvements. Our project utilizes the Google AI GoEmotions dataset to expand emotion classification datasets, improving chatbot sensitivity, online behavior detection, and customer support. By training neural networks and SVM models to analyze text tonality, we advance emotion analysis in NLP, benefiting stakeholders such as chatbot system providers, online platforms, and customer support departments. Our project's enhanced emotion analysis addresses this real-world problem of limited sensitivity and understanding, leading to more empathetic interactions, improved content moderation, and optimized customer support, ultimately enhancing user experiences.

2.2) Business objectives.

Expand emotion classification datasets by training models to analyze text tonality using the Google AI GoEmotions dataset.

Specific Objectives

a) Develop a model to classify text data into different sentiment categories, such as positive, negative, neutral, or mixed sentiments.
b) To establish the most appropriate model amongst the ones chosen, to accurately predict sentiments.
c) Improve customer support by recognizing and addressing user emotions in textual

communication.

Business Success Criteria To make accurate predictions on sentiments from the text data.

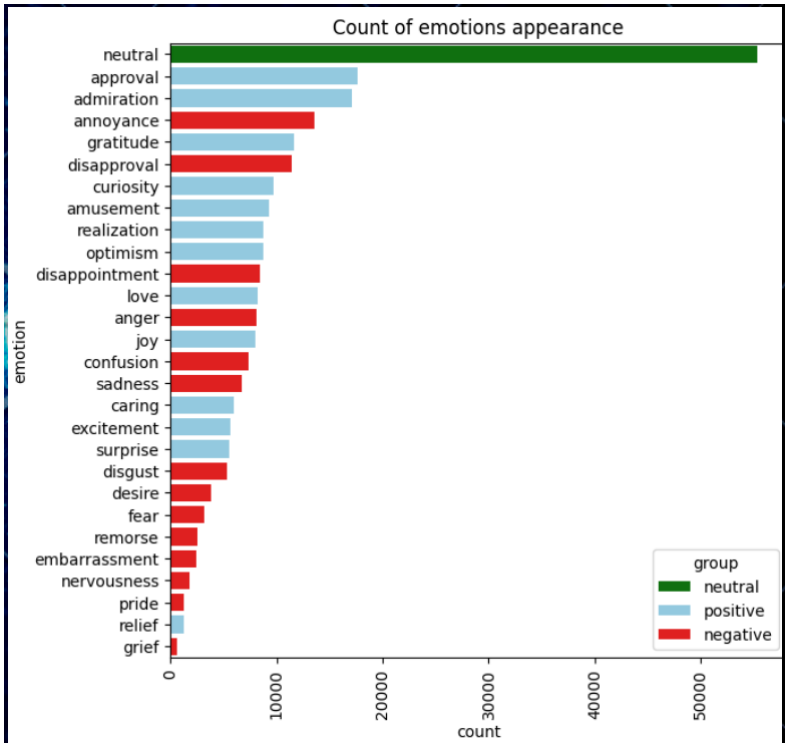
3) DATA UNDERSTANDING

The dataset was sourced from Kaggle, [hover above to go to the dataset.].The dataset has 31 columns, the *id,text,example_very_unclear and various emotions*.The Google AI GoEmotions dataset contains labeled comments from Reddit users expressing diverse emotions. This dataset is suitable for training neural networks to analyze text tonality, as it provides a comprehensive emotional spectrum and allows for subtle differentiation among various emotions. The dataset includes detailed emotional annotations and descriptive statistics, facilitating the analysis of emotions in text. While there may be limitations such as potential biases and subjective categorization, the GoEmotions dataset remains valuable for enhancing chatbot sensitivity, detecting online hazards, and improving customer support through the analysis of diverse emotions.

4) DATA EXPLORATION.

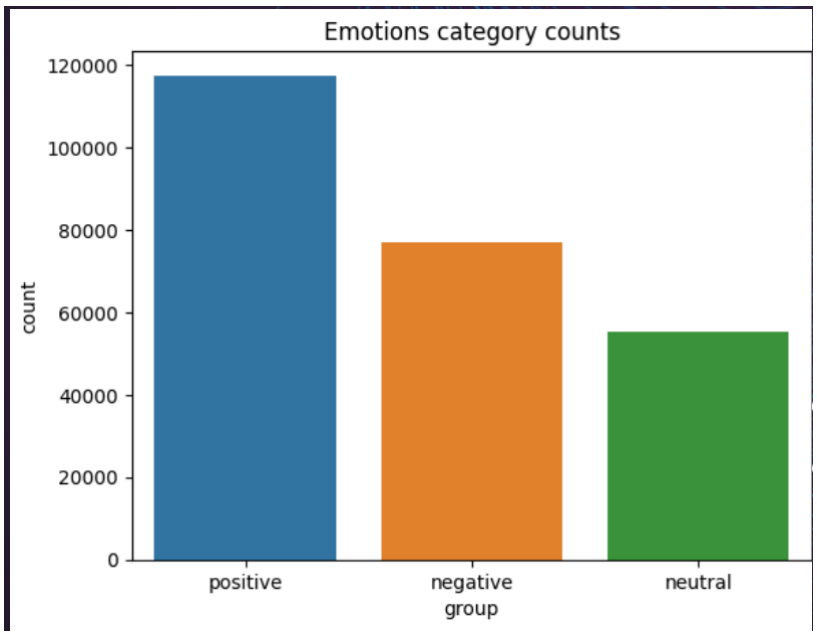
We did various analysis for our EDA , below are some of the visualizations from our analysis.

a)A visualizations of the emotions and the counts of how they appear.

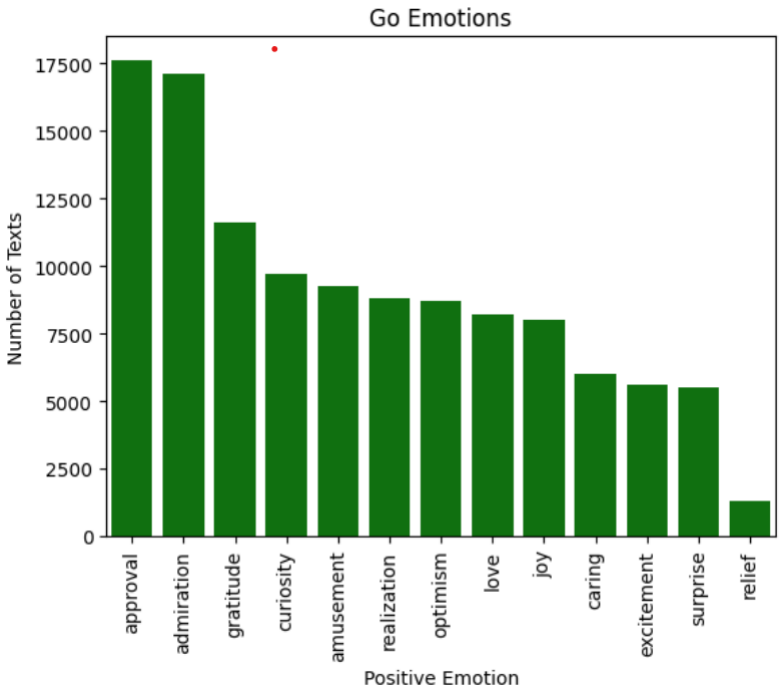


b) A visualization of the emotions categories and their counts.

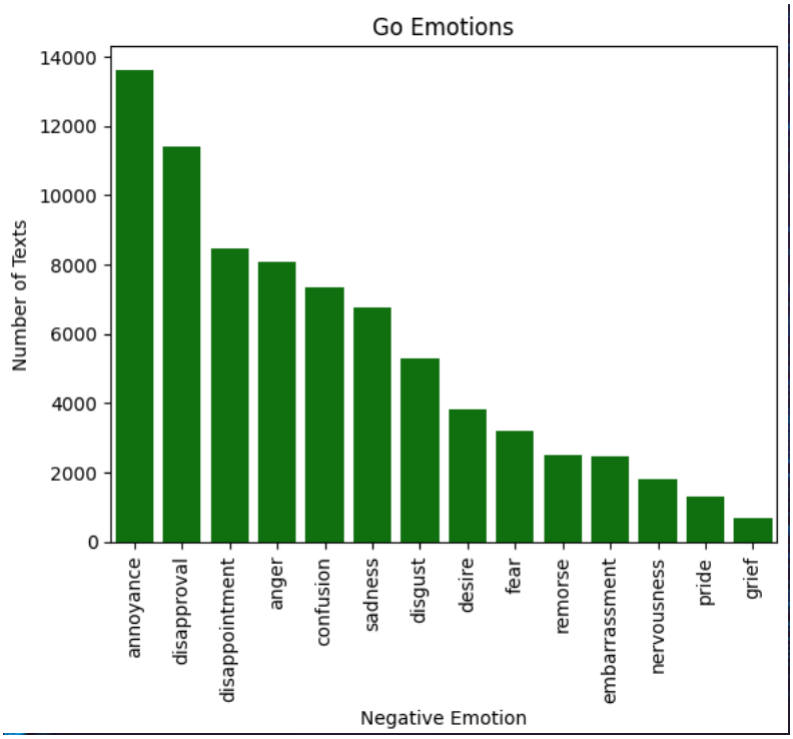
So we have three emotion categories, namely: 1) Positive emotions. 2) Negative emotions 3) Neutral emotions.



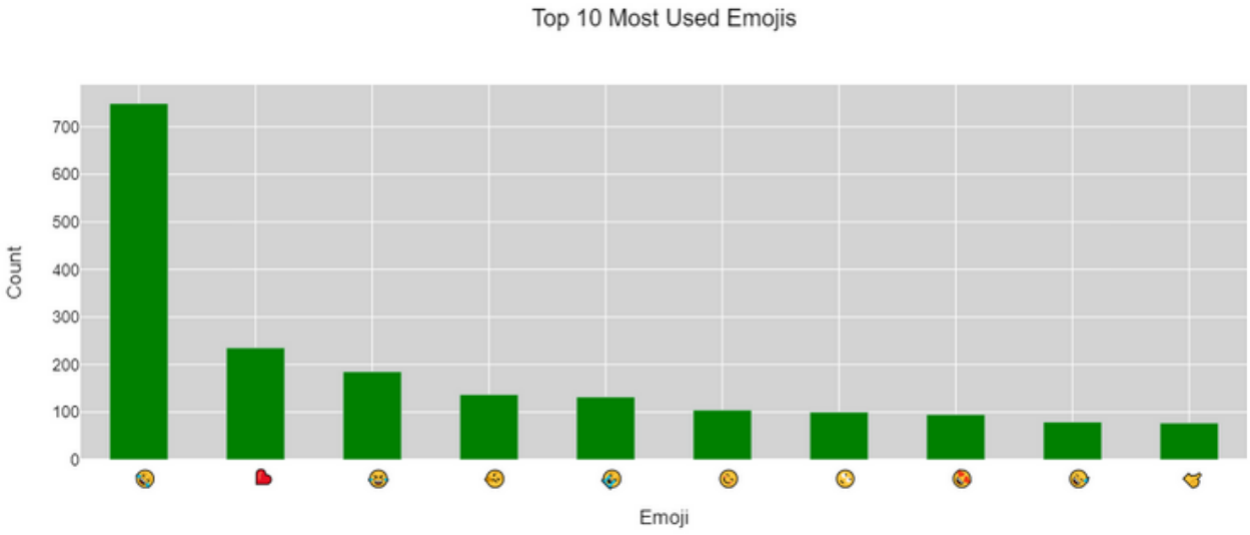
c) A visualisation of the positive emotions and their counts.



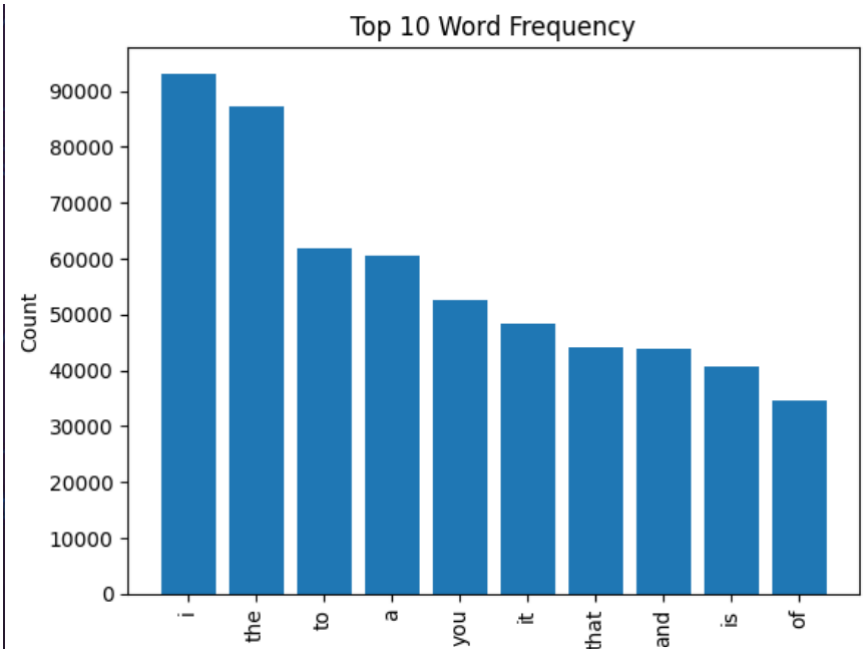
d) A Visualisation of the negative emotions and their counts.



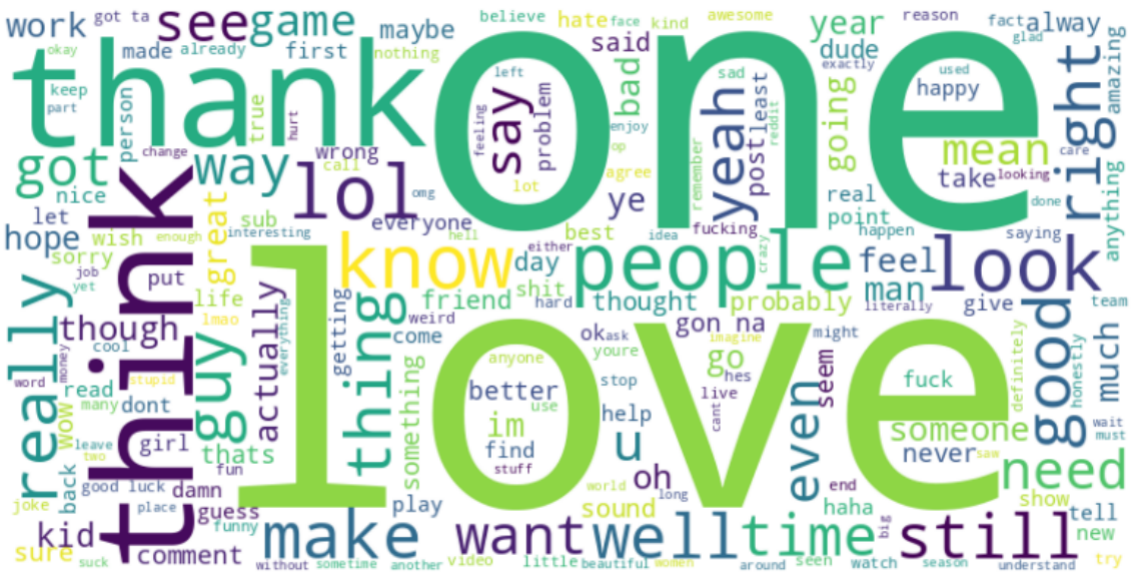
e) A visualization of the Top ten used emoji's.



f) A visualization of the word counts before the removal of stop words.



g) A visualization of a word cloud we created.



5) DATA PREPROCESSING.

In this stage, we seek to get the dataset in a format that is ready for the modeling , in our analysis , we did the following preprocessing steps. 1) Lowercasing the text column. 2) Removing punctuations. 3) Removing emoji's. 4) Performing word tokenization. 5) Removing stop words. 6) Lemmatization. 7) Emotions categorization. 8) Label encoding our emotions.

6) MODELLING AND EVALUATION.

In our sentiment analysis quest, we decided to model using four different models,these models include: 1) Support Vector Machine models(SVM). 2) Recurrent Neural Networks(RNN). 3) Convolutional Neural Networks (CNN). 4) Transformer Models.(Specifically the BERT model).

After performing our analysis , we were to choose the best performing model and proceed with it to the deployment of the model, in our case the SVM model was the best performing model, with an accuracy of 0.97. Below is how each model performed.

a) The SVM model.

Below is how our SVM model performed.

```
Accuracy: 0.97
Precision: 0.97
Recall: 0.97
F1-Score: 0.97
Classification Report:
              precision    recall  f1-score   support

    negative       0.91      1.00      0.95     18619
     neutral       0.98      0.88      0.93     16035
    positive       1.00      0.99      1.00     27455

    accuracy                           0.97     62109
   macro avg       0.96      0.96      0.96     62109
weighted avg       0.97      0.97      0.97     62109
```

b) The RNN model.

Below is how the RNN model performed.

	precision	recall	f1-score	support
0	0.49	0.32	0.39	6270
1	0.41	0.46	0.43	6041
2	0.61	0.69	0.64	10037
accuracy			0.52	22348
macro avg	0.50	0.49	0.49	22348
weighted avg	0.52	0.52	0.52	22348

c) The CNN model.

Below is how the CNN model worked.

```
Epoch 1/10
315/315 [==============================] - 46s 140ms/step - loss: 0.9559 - accuracy: 0.5348 - val_loss: 0.8865 - val_accuracy: 0.5899
Epoch 2/10
315/315 [==============================] - 42s 135ms/step - loss: 0.8293 - accuracy: 0.6272 - val_loss: 0.9053 - val_accuracy: 0.5832
Epoch 3/10
315/315 [==============================] - 44s 140ms/step - loss: 0.6961 - accuracy: 0.7077 - val_loss: 0.9937 - val_accuracy: 0.5613
Epoch 4/10
315/315 [==============================] - 44s 138ms/step - loss: 0.4858 - accuracy: 0.8123 - val_loss: 1.1969 - val_accuracy: 0.5481
Epoch 5/10
315/315 [==============================] - 43s 138ms/step - loss: 0.2905 - accuracy: 0.8926 - val_loss: 1.5867 - val_accuracy: 0.5398
Epoch 6/10
315/315 [==============================] - 43s 137ms/step - loss: 0.1794 - accuracy: 0.9358 - val_loss: 2.0172 - val_accuracy: 0.5239
Epoch 7/10
315/315 [==============================] - 42s 134ms/step - loss: 0.1232 - accuracy: 0.9560 - val_loss: 2.4043 - val_accuracy: 0.5268
Epoch 8/10
315/315 [==============================] - 44s 139ms/step - loss: 0.0889 - accuracy: 0.9671 - val_loss: 2.7913 - val_accuracy: 0.5157
Epoch 9/10
315/315 [==============================] - 44s 139ms/step - loss: 0.0756 - accuracy: 0.9715 - val_loss: 3.0677 - val_accuracy: 0.5166
Epoch 10/10
315/315 [==============================] - 48s 152ms/step - loss: 0.0717 - accuracy: 0.9730 - val_loss: 3.4910 - val_accuracy: 0.5172
350/350 [==============================] - 3s 9ms/step
Confusion Matrix:
[[1420  868  847]
 [ 833 1250  923]
 [ 965 1117 2951]]
```

d) The transfomer models(BERT model).

Below is how the bert model performed.

```
/usr/local/lib/python3.10/dist-packages/transformers/optimization.py:411: FutureWarning: This implementation of AdamW is deprecated and
  warnings.warn(
```

[998/1250 1:08:01 < 17:12, 0.24 it/s, Epoch 1.60/2]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.939100	0.843202	0.626800

[1250/1250 1:27:13, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.939100	0.843202	0.626800
2	0.785600	0.846391	0.626400

```
TrainOutput(global_step=1250, training_loss=0.862377880859375, metrics={'train_runtime': 5237.6736, 'train_samples_per_second': 3.818,
'train_steps_per_second': 0.239, 'total_flos': 328891772160000.0, 'train_loss': 0.862377880859375, 'epoch': 2.0})
```

From the above results, we can conclude that the best performing model was the SVM model hence this is what we shall use in the deployment stage.

7) MODEL DEPLOYMENT.

There is really not much to get into, we deployed our best model (SVM) in the streamlit app.

8) CONCLUSION.

Reddit data shows predominantly positive sentiments towards the company. The sentiment analysis model demonstrates high accuracy, precision, and recall for positive and neutral sentiments, making it reliable for sentiment analysis.

9) RECOMENDATIONS.

We would recomend that the from all the models the SVM model be used highly when performing this task of sentiment analysis as its results have proven to be quite high.