SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING

TEAM 2 MEMBERS:

- **Anthony Ngatia**
- **Elsie Nduta**
- **Jessyca Aperi**
- **Joy Kipkemboi**
- **Naomi Rotich**

SUBMISSION DATE: 29 / 07/ 2023

BUSINESS UNDERSTANDING

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.

BUSINESS OBJECTIVES

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

Specific Objectives

- Develop a model to classify text data into different sentiment categories, such as positive, negative, neutral, or mixed sentiments.
- To establish the most appropriate model amongst the ones chosen, to accurately predict sentiments.
- 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.

ASSESSING THE SITUATION

Resource Inventory

Datasets

- Google AI GoEmotions data - https://www.kaggle.com/datasets/shivamb/go-emotions-google-emotions-dataset/download?datasetVersionNumber=1

Software Used

- Google Colab
- Pandas
- Numpy
- GitHub
- Streamlit
- Spyder

ASSUMPTIONS

The data provided is correct and up to date.

CONSTRAINTS

The texts contained multiple emotions presented in single text data, which resulted in a multi-classification problem.

The texts were many in number, resulting in a significant increase in model run times.

TERMINOLOGIES

- **CNN** - Convolution Neural Networks
- **RNN** - Recurrent Neural Network
- **SVM** - Support Vector Machines

PROJECT PLAN

A team of five members conducted the project. The project was divided into major sections i.e. Data preprocessing, modeling, evaluation, deployment, non-technical presentation, and reporting. The teammates were allocated tasks and expected to deliver within the timelines specified.

DATA MINING GOALS

Our data mining goals for this project are as follows;

- To import and download the dataset from Google AI GoEmotions dataset: - The datasets were already provided and hence we did not perform data scrapping.

DATA MODELLING SUCCESS CRITERIA

Our success will be measured by the following criteria.

- Performing modelling and evaluation of the text data and obtaining an accuracy result of 70% and above.
- Deployment of the model using Streamlit to predict unseen data.

DATA UNDERSTANDING

DATA DESCRIPTION

Columns:

The dataset contains categories of emotions identified by Google together with psychologists and includes:

12 positive emotions (*admiration, amusement, approval, caring, curiosity, excitement, gratitude, joy, love, optimism, relief, surprise*)

11 negative emotions (*sadness, pride, fear, embarrassment, disapproval, disappointment, confusion, annoyance, anger, nervousness, desire*)

4 ambiguous emotions (*remorse, realization. Grief, disgust*)

1 neutral emotion (*neutral*)

Which makes the dataset suitable for solving tasks that require subtle differentiation between different emotions.

Shape

The dataset contained 211225 rows and 31 columns.

Data types

The dataset contained three data types namely:

- Boolean
- Integers
- Strings

DATA PREPARATION

In this phase, we conducted data munging which involved preparing the final dataset for modeling.

Data Selection - The dataset was provided on Kaggle.

Data Cleaning - The dataset was corrected by removing the rows which has no associated labeled emotions. Upon performing EDA, there was no collinearity observed between the different emotions. The 28 different emotions were then grouped into four categories namely; positive, negative, neutral, and ambiguous to facilitate modeling.

Dealing with duplicates - the dataset contains duplicates which were represented as similar text, but different emotions. We performed deduplication of the dataset so as to prepare the data for modelling.

DATA MODELLING

The tasks involved in this phase were as follows:

- Selecting modeling techniques
- Performing Train Test splits
- Building the model
- Model Evaluation and HyperParameter Tuning.
- Deployment

We used four modeling techniques namely CNN, RNN, SVM, and Transformers. Upon performing evaluation and parameter tuning, we chose the model with the best performance metrics. The models and respective results are shown below:

CNN Modelling - *val_loss: 0.9967 - val_accuracy: 0.5742*

RNN Modelling - *Accuracy: 0.50 Precision: 0.50 Recall: 0.50 F1-Score: 0.49*

SVM Modelling - *Accuracy: 0.97 Precision: 0.97 Recall: 0.97 F1-Score: 0.97*

We proceeded to further perform deduplication upon realization that there were text duplicates depicting different emotions. This further improved our models' performance as shown below:

CNN Modelling - *Precision: 0.6656 Recall: 0.6102 F1 Score: 0.6377*

RNN Modelling - *Precision: 0.52 Recall: 0.52 F1 Score: 0.52*

SVM Modelling - *Precions: 0.97 Recall: 0.96: F1 Score: 0.96*

For the transformer model, modelling was performed on text without preprocessing and after preprocessing and the results were obtained as below:

Before preprocessing: -

Transformers - *val_loss: 0.8464 -val_accuracy: 0.6264*

After preprocessing**: -**

Transformers - *val_loss: 0.8787 -val_accuracy: 0.5976*

From the results obtained, we chose SVM Modelling as our preferred model for deployment.

DEPLOYMENT

For the deployment, we used the streamlit extension in Anaconda and upon loading the saved model, defined the user input parameters and used the function defined in the application to predict the sentiments.

CONCLUSION

CNN: -

The performance of the model attained an accuracy of 0.5790 translating to an accuracy score of ~ 58%. This indicated that approximately 58% of all text sentiments were predicted correctly using the CNN model.

RNN: -

The performance of the model attained an accuracy of 0.52 translating to an accuracy score of ~ 52%. This indicated that approximately 52% of all text sentiments were predicted correctly using the RNN model.

SVM: -

This metric indicates that the model is able to predict the correct sentiment for approximately 96% of the examples in the dataset. This was the highest among the models evaluated.

Transformers: -

For this model, the data was modelled before and after preprocessing. It was noted that the model performed better on uncleaned data and hence providing an accuracy score of 0.6264 which translates to ~63% accuracy on predicted sentiments on the text data.

REFERENCES:

1. Kaggle Dataset - https://www.kaggle.com/datasets/shivamb/go-emotions-google-emotions-dataset/download?datasetVersionNumber=1

2. Webpages:
   A. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 Accessed on 10 June 2023 at 1700hrs.

   B. https://www.tensorflow.org/tutorials/images/cnn Accessed on 12 June 2023 at 1000hrs

   C. https://www.simplilearn.com/tutorials/deep-learning-tutorial/ Accessed on 20th June at 2100hrs

3. Canvas Content.