INTERNSHIP: PROJECT REPORT

Internship Project Title	TCSiON Remote internship RIO 45
Project Title	RIO 45- Automate detection of different sentiments from textual comments and feedback
Name of the Company	TCSiON
Name of the Industry Mentor	Harish Kumar
Name of the Institute	Prepinsta

Start Date	End Date	Total Effort (hrs.)	Project Environment	Tools used
07/02/2023	16/04/2023	50	Google Collab	Python, Eclipse, Virtual box

PROJECT SYNOPSIS:

In Sentiment analysis, the data mainly focuses on whether the feedback to a product is :

Negative (less then 5)
Positive (Greater then 6)
Neutral (between 5 and 6)
This provides a top level review of the product and its acceptance in the market.

Emotion Analysis provides a deeper insight of consumer emotions.
Analyzing textual feedback to the level of reading between lines.
Categorizing feedback and analyzing its emotion by picking up words, contexts, patterns, behaviors. This can be even taken to the level of individual's expressive capability of a particular situation.

SOLUTION APPROACH:

- **Rule Based Systems:**

1. Rule-based approaches classify text into organized groups by using a set of linguistic rules.
2. Each rule comprises of a pattern based on semantics and its predicted category.

- **Machine Learning based Systems:**

1. Text classification based on past observations.
2. By using training data, the algorithm can learn the different associations between pieces of text and that a particular output (i.e. tags) is expected for a particular input (i.e. text).

3. Feature extraction: Transforms each text into a numerical representation in the form of a vector. E.g. bag of words [a vector represents the frequency in a predefined dictionary of words]
4. The algorithm is fed with training data consisting of feature sets.
5. Once trained with enough training samples, the machine learning model can begin to make accurate predictions on unseen text with similar feature sets.

ASSUMPTIONS:

- The collected data is representative of the entire population: We assume that the textual comments and feedback collected from various sources are representative of the opinions and sentiments of the entire population.

- The labeled data is accurate and reliable: We assume that the labeled data used to train the sentiment classification model is accurate and reliable. Inaccurate or biased labeling can affect the performance of the model.

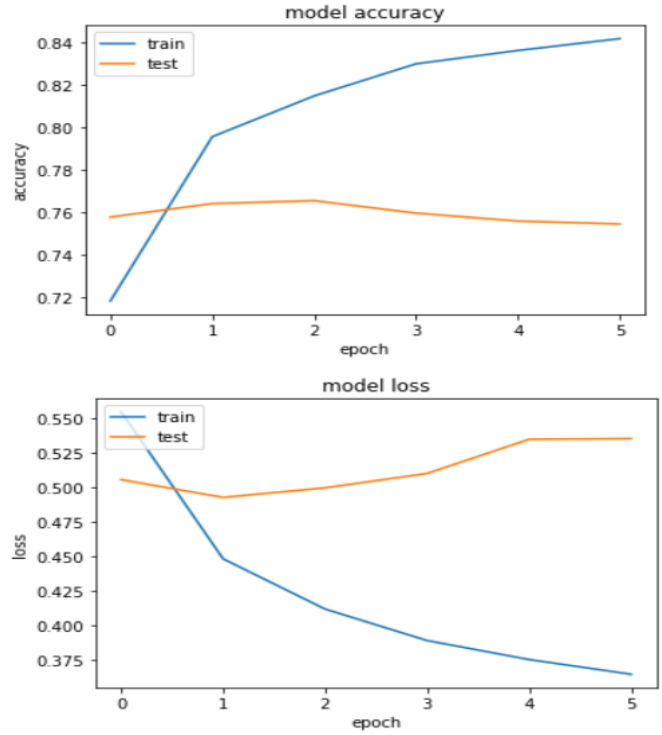
- The language used in the textual comments and feedback is consistent: We assume that the language used in the comments and feedback is consistent across all sources and does not have any significant variations that could affect the performance of the sentiment classification model.

- The sentiment categories used in the model are appropriate: We assume that the sentiment categories used in the model, such as positive, negative, and neutral, are appropriate for the specific business domain and the context in which the data is collected.

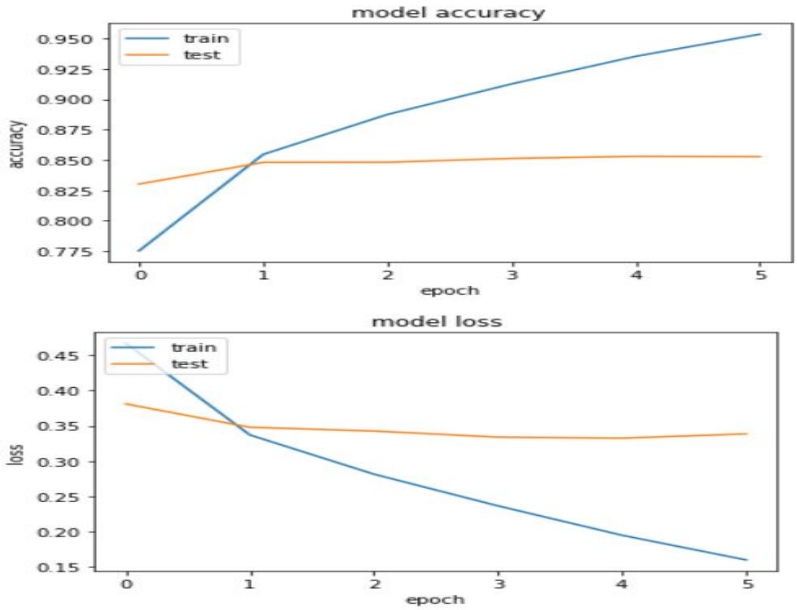
- The model is generalizable: We assume that the sentiment classification model developed on one dataset can be generalized to other datasets with similar characteristics. However, the performance of the model may vary depending on the characteristics of the new dataset.

Project Diagrams:

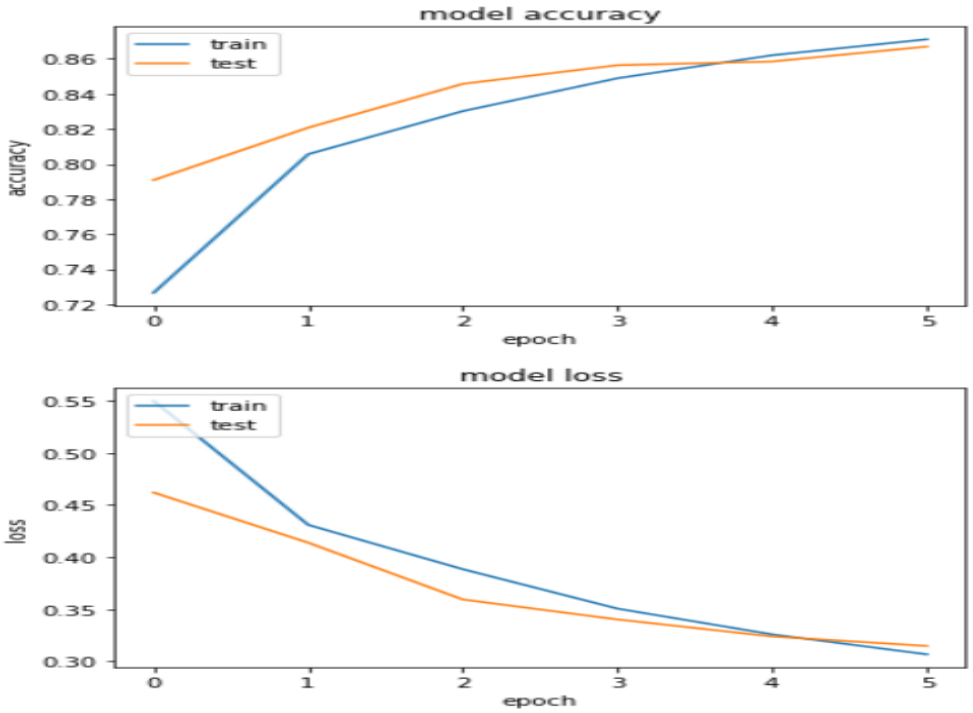
Simple neural network model graph-



Convolutional Neural Network model graph-



Recurrent Neural Network (LSTM) model graph-



OUTCOME:

After implementing different deep learning algorithms the following accuracy obtained:
Simple Neural Network:
Training accuracy- 84.12
Testing accuracy- 75.27

Model overfit as there is huge difference in training and testing accuracy

Convolutional Neural Network:
Training accuracy- 94.99
Testing accuracy- 85.40

Difference in training and testing accuracy is reduced still it is not acceptable

Recurrent Neural Network (LSTM):
Training accuracy- 87.37
Testing accuracy- 86.44

Difference in training and testing accuracy is minimum it can be accepted.

EXCEPTIONS CONSIDERED:

Several exceptions can occur during different stages of the project. Here are some of the common exceptions:

- Data Quality Issues: Data quality issues can arise during data collection or preprocessing, such as incomplete data, duplicate data, or inaccurate data. Such issues can affect the accuracy of the sentiment classification model and require additional effort to resolve.

- Overfitting: Overfitting occurs when the sentiment classification model is trained too well on the training data and becomes too specific to the data. This can result in poor performance on new, unseen data. To avoid overfitting, the model may need to be regularized or the training data may need to be augmented. Or different algorithms s can be used to analyse to reduce overfitting. Difference between training and testing accuracy should be minimal .

- NLP Limitations: NLP algorithms used for feature extraction may have limitations in processing specific types of text, such as informal or slang language, sarcasm, or irony. Such limitations can affect the performance of the sentiment classification model.

To address these exceptions, appropriate measures and strategies need to be taken during different stages of the project to ensure the accuracy and reliability of the sentiment classification model.

ENHANCEMENT SCOPE:

There is always a scope of improvement. Here are a few things which can be considered to improve.

Different classifier models can also be tested.
The remaining two models can be tuned for better results. For example, after plotting.
Try a different data set. Sometimes a data set plays a crucial role too.

Link of code:
https://colab.research.google.com/drive/1sms6BUPFAHafxZua_FEyMq2QglW8DhlI?usp=sharing

Link of csv file(data sheets used) :
https://drive.google.com/file/d/1Q3jArl3FvcxRVX13xUXpJrY-FTRYxgQ2/view?usp=sharing

https://drive.google.com/file/d/1He6pot7hsWfSURS0Kh-MlmDyQqfRUUdd/view?usp=sharing

CONCLUSION:

From observing the accuracy of the model and different deep learning algorithms we can conclude that **Recurrent Neural Network (LSTM)** is the most appropriate algorithm for our problem.