Capstone Project Report

NLP Sentiment Analysis on IMDB Data Set

By Sairaman Ramasubramanian,
08-Oct-2021

Contents

Problem statement

A Leading AI & ML Organization is expanding its footprint in NLP with specific focus on Sentiment analysis and has an aggressive roadmap of 3 years to capture a significant market share.

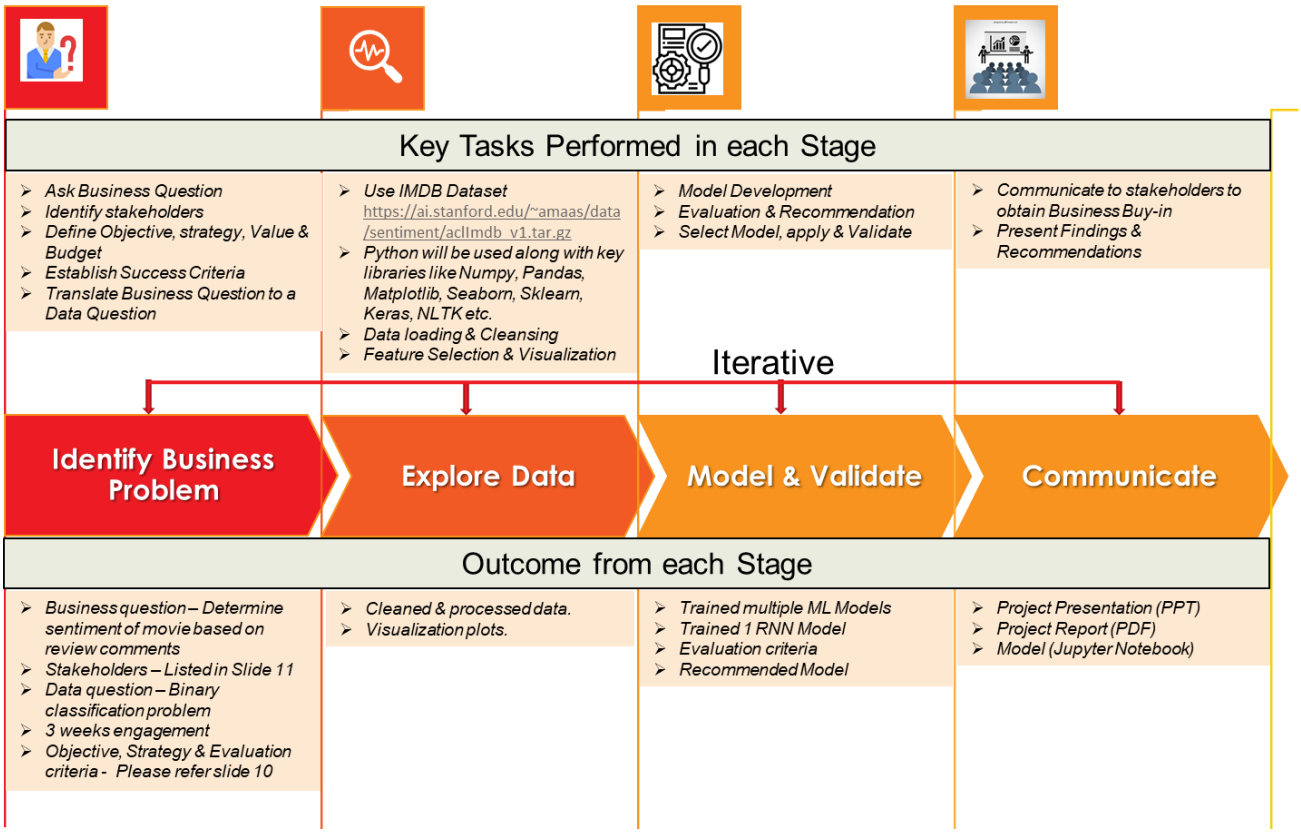
As part of the pilot phase, they need to device a mechanism to read reviews of movies and classify it as positive or negative

Industry/ domain background

NLP is an emerging area in AI & ML industry with a huge focus. The buzz of NLP in the market is growing in an exponential manner which is expected to touch the **mark of $ 16 billion by 2021** with the compound growth rate of **16 % annually**.

Data science process

Below is an illustrated flow of the Data science process in the context of the current implementation



As depicted in the above diagram,

The process will comprise of the 4 pillars or stages

Pillar1 - Identify Business Problem:

This pillar consists of the following activities:

1) Ask Business questions:

This refers to discussions and questions with the different stakeholders and formulating the business problem which in this case is summarized below:
The Organization needs a robust Sentiment Analysis Model to be built as part of the pilot phase based on the IMDB Movies Data set provided.
The organization has engaged the services of Sai Science Pte Ltd to develop the same.

2) Identify Key Stakeholders:

In this activity, we identify the key stakeholders that will be part of the project.

3) Below is a summary of the Key stakeholders

Key Client Stakeholders	Vendor Stakeholders
Client Engagement Director	Engagement Director
Client Project Manager	Project Manager
Client BA / SME	Lead Data Scientist
Business Sponsor – Head of New Business	Data Architect
Technology Sponsor – Head of Technology	Developers

- The Business Sponsor has clearly mentioned this project as one of their High visible & High importance projects for the year as it directly impacts their growth plan for the next 3 years.
- For the same reasons, the Technology sponsor has mobilized a team to provide a fully representative data and all the support required to answer any queries on the data.

4) Corresponding Data Problem/Question:

This activity consists of translating the Business question identified to a corresponding Data question/Problem. Below is the data question for this project.

- A Binary classification problem to classify the sentiment as positive or negative based on the reviews using a robust Model.

5) Summary of activities:

The summary of activities to be performed as part of the engagement is

- Data Loading
- NLP pre-processing to clean the data & visualization
- Vectorization, Feature selection & visualization of top frequency words
- Build word cloud for positive & negative words pre and post feature selection
- Train & Evaluate multiple ML classifier models based on different parameters
- Train one DNN LSTM Model & check performance.
- Recommend the best model& test couple of reviews.

6) Objective:

- To recommend a model with the best overall score.

7) Strategy & Evaluation Criteria:

- To Evaluate multiple models using the evaluation parameters identified and recommend the one with the best overall score.
- Accuracy, Precision, Recall and ROC_AUC scores will be the evaluation scores used for the various ML models that will be trained and tested.

8) Establish Success Criteria:

- To validate the scores of the different models trained and tested.
- The recommended model should have an overall score of at least 85%.

9) Key Assumptions

- The client team would make themselves available to clarify any questions on the data set. (2 sessions of 2 hours each have been planned to tackle such questions)
- The scope of the project covers only supervised data.
- As discussed, and agreed upfront, this is a 3-weeks engagement
- The data set is complete and a significant representation.
- The output will be the Machine learning Model (Jupyter Notebook), Power point presentation and a project report in MS word.

Pillar2 – Explore Data:

This pillar consists of the following activities

1) Identify dataset:

For the pilot phase, we will use the below dataset

- **Data Source (IMDB Data) –** https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz This dataset contains highly polar movie reviews split in to
- **Training dataset** – Stored as individual files in training folder
- **Testing dataset** – Stored as individual files in testing folder
- **# of records**
    - Training Data set **– 25000**
    - Testing Data set **- 25000**
- **Type of Data – Movie Sentiment Analysis –** To predict whether the sentiment is positive or negative based on Movie review.

2) Deliverables & Tools used:

The model will be built using Python. As part of the engagement, the following will be delivered.

- **Project Presentation** – MS Power Point
- **Project Report** – MS Word
- **Codebase** – Jupyter Notebook. 3 Notebooks will be created
    1) Data loading, Cleansing & Feature Selection
    2) ML Models including ensemble techniques
    3) RNN Model

Key Libraries Used – Numpy, Pandas, Seaborn, Matplotlib, NLTK, Keras, sklearn, pickle.

3) Data Loading & Cleansing

Data Loading & cleansing using Python will be done in this stage. The details are captured in the solutioning section.

4) Feature selection & Visualization

Vectorization, Feature selection & visualization using Python will be done in this stage. The details are captured in the solutioning section.

The output from this stage will be the cleaned & processed data.

Pillar3 – Model & Validate:

This pillar consists of the following activities

- Training multiple models
- Testing the models based on the evaluation parameters identified.
- Recommend the best performing model

The details are captured in the solution section.

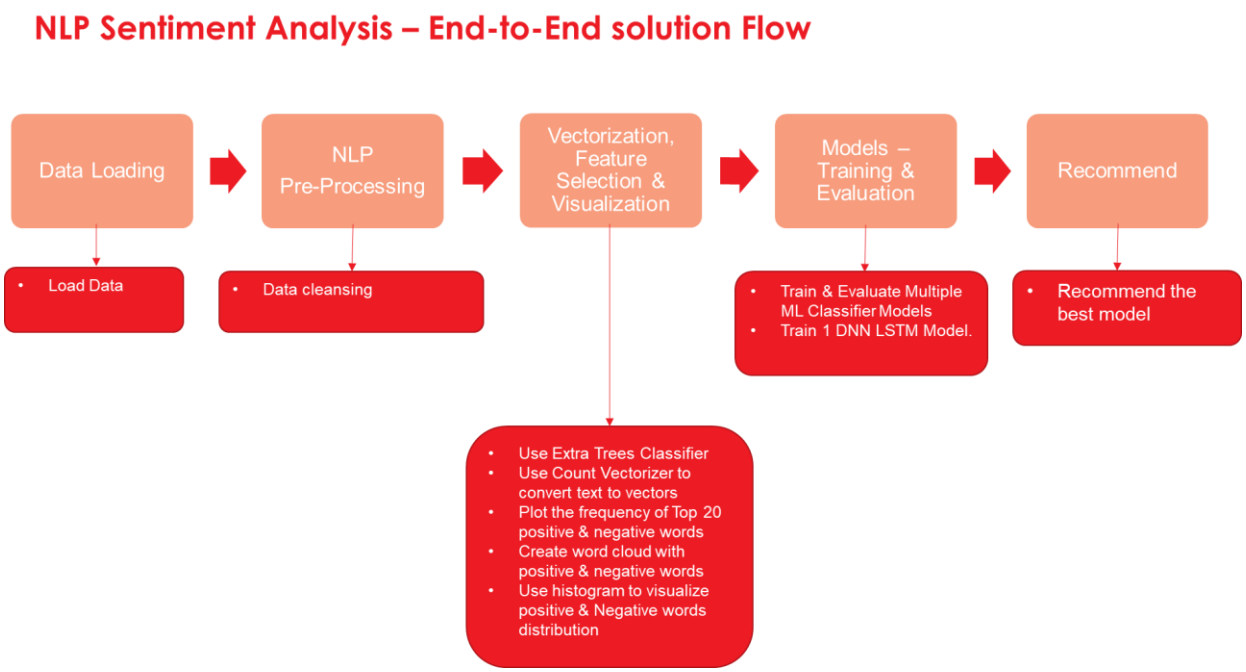
Pillar4 – Communicate:

This pillar consists of the following activities

- Present the Solution details & recommendations to the stakeholders
- Submit the recommended model along with the other deliverables.
- Obtain the sign-off and buy-in for the deployment
- Discuss next steps

End-to-end solution

The below diagram depicts the end-to-end solution flow



As depicted in the diagram above, the solution consists of the following components

Data Loading

Data Loading involves loading the raw data from folders / files using python libraries to get it ready for exploration and cleansing.

For this project, the source data is organized in the below form

Training data

\train
\train\pos --- Positive review
\train\neg --- Negative reviews

Testing data

\test

\test\pos --- Positive review
\test\neg --- Negative reviews

Data is loaded from all the folders to a pandas data frame for processing.
Below is a sample code snippet of the method used for loading positive reviews from the \train\pos folder.

```
Import os
path="C:\\Users\\Sairaman\\Downloads\\datascience\\Project\\Capstone\\aclImdb\\train\\pos"

data=[]
files=[path+'/'+f for f in os.listdir(path) if os.path.isfile(path+'/'+f)]
for f in files:
    with open(f, 'r', encoding='utf8') as myfile:
        data.append(myfile.read()
```

NLP Pre Processing

This component involves cleansing the loaded data to

- Remove Stop Words
- Remove Punctuations
- Remove Tags
- Lemmatization
- Stemming

We have developed python functions to take care of each of the tasks mentioned above.
Below is a visual representation of the Data Loading & Pre-Processing steps

This component applies the functions to the review text and outputs a cleaned text which will be used by the subsequent stage.

Vectorization, Feature selection & Visualization

Vectorization:

It is the process of converting text to vector of numbers.

Feature selection

It is a process where we automatically select those features in the data that contribute most to the prediction variable or output of interest

Method Chosen - An extra-trees classifier.

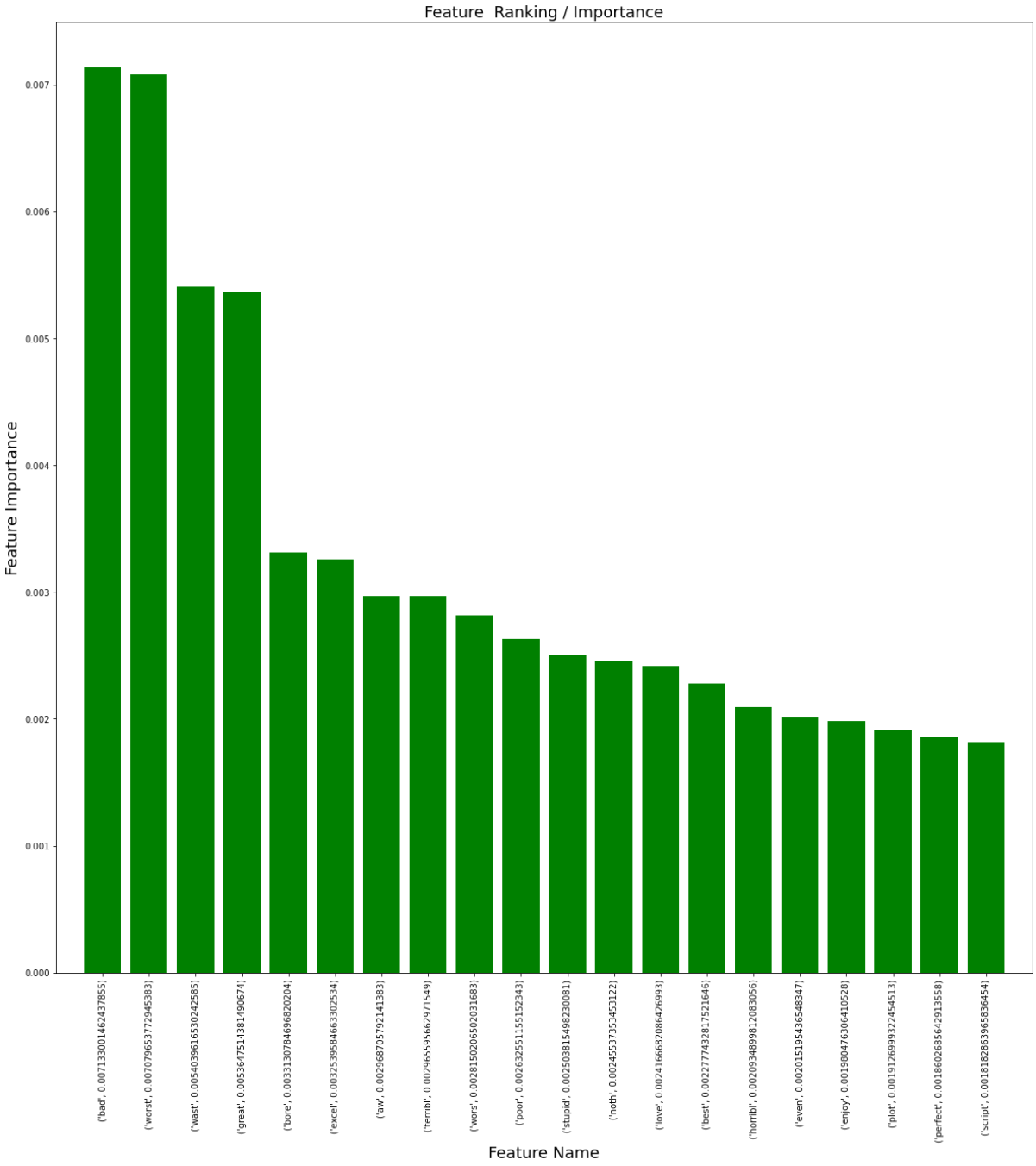
This class implements a meta estimator that fits several randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

The above is achieved by creating some python functions in this project
Get_Fetaures
Print_Features.

The above 2 functions do the vectorization and feature selection. The print_features function helps in visualizing the Top 20 (configurable) features

The chart below shows the visualization on training data (20 features)

```
Feature ranking:
feature 7908 : bad (0.007133)
feature 89968 : worst (0.007080)
feature 87677 : wast (0.005404)
feature 34573 : great (0.005365)
feature 11434 : bore (0.003313)
feature 27495 : excel (0.003254)
feature 7512 : aw (0.002969)
feature 80194 : terribl (0.002966)
feature 89939 : wors (0.002815)
feature 62639 : poor (0.002633)
feature 77664 : stupid (0.002504)
feature 56595 : noth (0.002455)
feature 47952 : love (0.002417)
feature 9717 : best (0.002278)
feature 38637 : horribl (0.002093)
feature 27193 : even (0.002015)
feature 26408 : enjoy (0.001980)
feature 62154 : plot (0.001913)
feature 60848 : perfect (0.001860)
feature 70725 : script (0.001818)
```

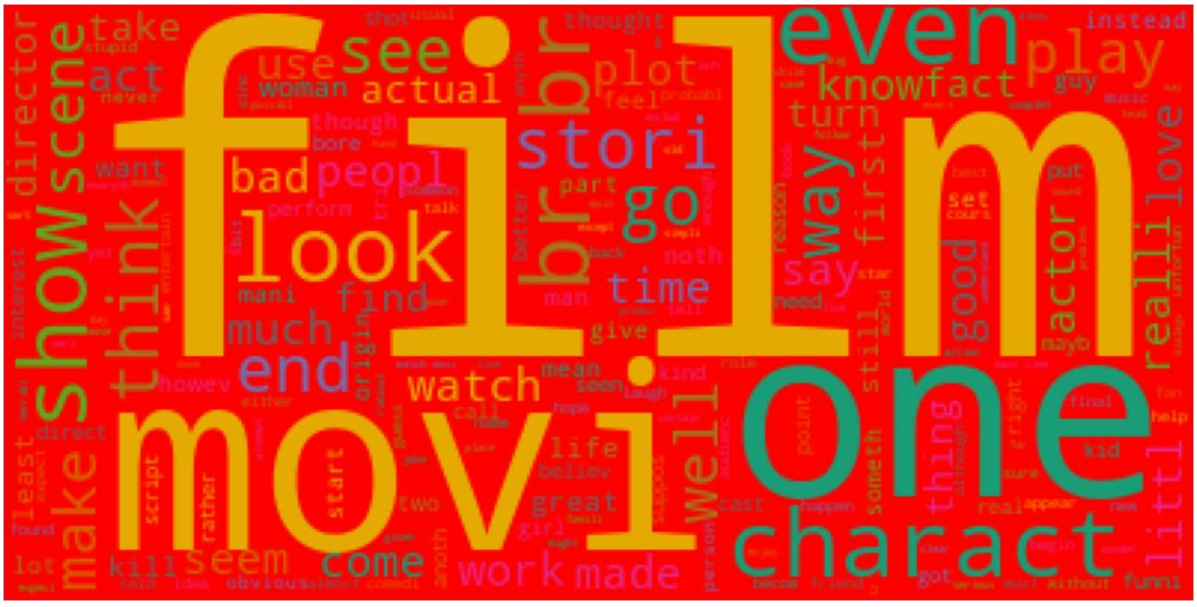


Similar approach is taken on the Test dataset as well.

Once the visualization is done, the training and testing data is stemmed together and a word cloud is built to show the frequency of the positive and negative words by size. **Word cloud** is a technique for visualising frequent words in a text where the size of the words represents their frequency.

Word Cloud of positive words:

Word cloud of negative words



The same details are captured again post feature selection i.e. initial word cloud is built based on the entire set of words in the text. Then post feature selection, the number of words is restricted to only those that are in the feature list and a second round of word cloud is published.
Similarly, the histogram distribution of positive and negative words is plotted pre and post feature selection.

Once this is done, the data is ready for modelling which is the next section.

Model – Training & Evaluation

This stage involves the training and testing of models based on the features and labels in the data.

We have used the pipeline approach in this project to do the modelling.

Pipeline is a utility that provides a way to automate a machine learning workflow. It works by allowing several transformers to be chained together.
One can also add an estimator at the end of the pipeline. Data flows from the start of the pipeline to its end, and each time it is transformed and fed to the next component.

A Pipeline object has two main methods:
fit_transform: this same method is called for each transformer and each time the result is fed into the next transformer
fit_predict: if your pipeline ends with an estimator, then as before the data is transformed until it arrives at the last step, where it is fed into the estimator and fit_predict is called on the estimator.

Sample code snippet using pipeline object from sklearn.

```
# Multinomial NB
pipelineNB = Pipeline([
    ('bow', CountVectorizer(analyzer=review_split)),  # strings to token integer counts
    ('tfidf', TfidfTransformer()),  # integer counts to weighted TF-IDF scores
    ('classifier', MultinomialNB()),  # train on TF-IDF vectors w/ Naive Bayes classifier
])
start_time = time.time()
pipelineNB.fit(msg_train,label_train)
cal_time1 = (time.time()- start_time)

predictionsNB = pipelineNB.predict(msg_test)
print(classification_report(predictionsNB,label_test),'time taken:',cal_time1)
```

ML Models Trained & Tested

- Naïve Bayes
- Decision Tree
- Random Forest
- Logistic Regression
- SGD Classifier
- ADA Boost

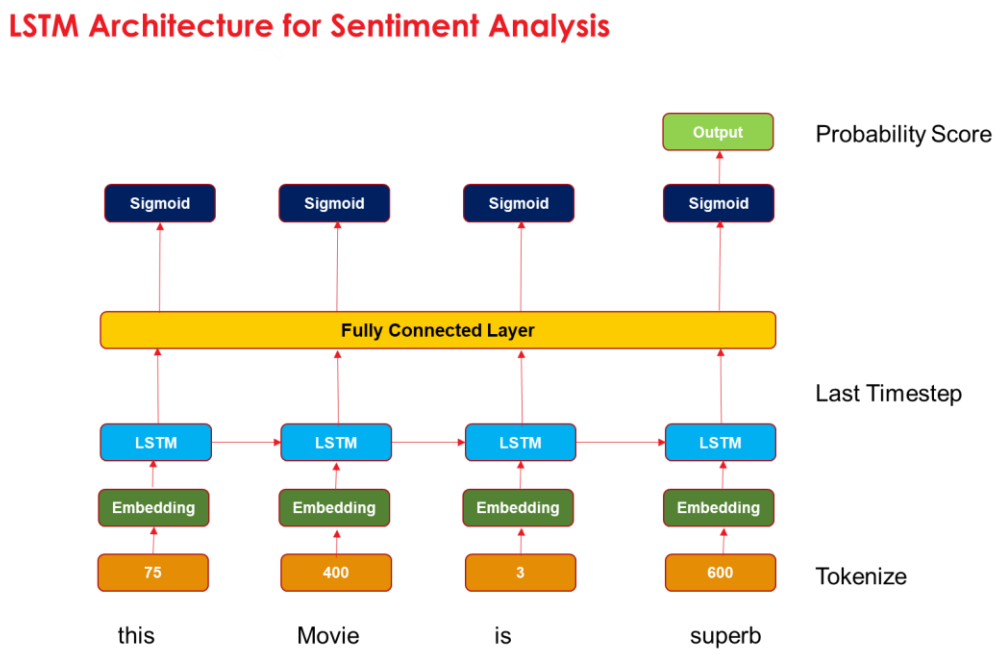
Evaluation Parameters Definition

- **Accuracy Score** – Model Accuracy
- **Precision** – represents the model's ability to correctly predict the positives out of all the positive prediction it made.
- **Recall** – quantifies the number of correct positive predictions made from all positive predictions that could have been made.
- **ROC_AUC** – AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

RNN Model

A RNN Model using LSTM model was also trained and tested.

LSTM Model Architecture & Details



LSTM Model Architecture – Current Project

```
EMBED_DIM = 32
model = Sequential()
model.add(Embedding(total_words, EMBED_DIM, input_length=max_length))
model.add((LSTM(32, return_sequences = True)))
model.add(Dropout(0.2))
model.add(Flatten())
model.add(Dense(250, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

LSTM Model - Params

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 106, 32)           96032
_________________________________________________________________
lstm (LSTM)                  (None, 106, 32)           8320
_________________________________________________________________
dropout_1 (Dropout)          (None, 106, 32)           0
_________________________________________________________________
flatten_1 (Flatten)          (None, 3392)              0
_________________________________________________________________
dense_2 (Dense)              (None, 250)               848250
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 251
=================================================================
Total params: 952,853
Trainable params: 952,853
Non-trainable params: 0
```

Model – Recommendation

	Model	Accuracy	Precision	Recall	ROC_AUC
0	Naive Bayes	0.84996	0.855795	0.84176	0.84996
1	Decision Tree	0.71724	0.719541	0.71200	0.71724
2	Random Forest	0.84296	0.849674	0.83336	0.84296
3	Logistic Regression	0.88196	0.877999	0.88720	0.88196
4	SGD Classifier	0.88060	0.876712	0.88576	0.88060
5	ADA Boost	0.81092	0.79136	0.84448	0.81092

LSTM Model produced an accuracy of 82% with 50 epochs

References

- Natural Language Processing (NLP) Simplified : A Step-by-step Guide (datascience.foundation)
- Sentiment Analysis — A how-to guide with movie reviews | by Shiao-li Green | Towards Data Science
- https://towardsdatascience.com/sentiment-analysis-a-how-to-guide-with-movie-reviews-9ae335e6bcb2