GRADUATE CERTIFICATE IN PRACTICAL LANGUAGE PROCESSING

SINGAPORE COMMUNITY HELP SOCIAL NETWORK

PROJECT PROPOSAL PRESENTATION

Siva Krishna THOTA

About The Project

Community help is a social network website based on Singapore to connect group of people and share useful information like other social network people can post tweets, likes and share text, images & videos through this website. Unlike other social networks, there is special feature which focus more on community interaction to help people and news sharing.

Features :

- User is allowed to find near by shops/food courts in and around the community and can also explore on more offers.
- Allow users to find friends in near communities.
- Merchants will receive the notifications if a user request for particular item from near by communities
- Seller/Buyer can post and get the news or offers from near by community.
- Based on history and interest user will get notifications on mobile app by tracking user location.
- Similar like whatsup groups can be created in this application.

Business problem statement

Problem statement # 1

When multiple users posted multiple tweets which leads to mass data could create nonuser friendly application. As a result, Posted data need to be segregated into categories

Problem statement # 2

Sentiment analysis is perfect for customer insights because you can identify sentiments in social mentions, hashtags, posts, comments, and videos. With sentiment analysis on community help , you can discover and attract new target audiences, and improve your brand presence . .

Problem statement # 2

translation is necessary because it can significantly and positively impact its benefactors On the other hand, social media translation can also be used as outputs to improve the popularity and accessibility of the business. Instead of broadening the market holistically, you can also choose specific audiences to engage in your posts. Translating and speaking using their everyday vernacular is one way to spark their interest.

Technical Problem Statement

Problem statement # 1

Classification and detection problem statement

The aim of this business use case is to develop a model that uses multi-label classification techniques by NLP

This project focuses on Category Classification detection via text language using different kinds of machine learning and deep learning classification techniques

Our goal is to build a high-accuracy model to distinguish text language into below categories.The input of our model is text language which is tweeted by end users

We first preprocess these text to get input vectors, then use different classifiers to predicted type of category

	Category
1	BUSINESS
2	ENTERTAINMENT
3	POLITICS
4	SCIENCE & TECHNOLOGY
5	HEALTH
6	SPORTS
7	WORLD NEWS
8	ARTS & CULTURE

	Category
9	TASTE
10	TRAVEL
11	WOMEN
12	GREEN
13	RELIGION
14	STYLE
15	CRIME
16	EDUCATION

Problem statement # 2

Sentiment analysis

This business use case derives to develop a model that uses NLP techniques to accurately detect emotions in text data. The model can be used for sentiment analysis, customer feedback analysis, and social media monitoring. The model is trained on a dataset of text data that has been labeled with the corresponding emotions expressed in it.

❖ **Positive : Happy, Love**

❖ **Negative : Sad, Anger**

❖ **Neutral**

Technical Problem Statement

Problem statement # 3

The Good

translation is necessary because it can significantly and positively impact its benefactors

Obtaining Information By Translating Other Sources

Social media translation can be used as inputs to aid the building of a brand and ensure relevant decision-making.

Benefits of social media translation for businesses

On the other hand, social media translation can also be used as outputs to improve the popularity and accessibility of the business.

Reach More People

Translating the original social media posts into ones of different languages is incredibly beneficial to an enterprise. This is because the translated work can now be read by many more people who may not be fluent in the original language.

Cater to Specific Audiences

Instead of broadening the market holistically, you can also choose specific audiences to engage in your posts. Translating and speaking using their everyday vernacular is one way to spark their interest.

What happens to brands not utilizing social media translation?

However, as much as it can give benefits to those that do it, it can also be a downfall for those that choose not to translate their social media posts. This is especially true for businesses that are just starting out and need exposure to fuel them.

Machine-Translation-English-to-Hindi

The model translates English text to Hindi text with the help of LSTM. The project was implemented in Keras Framework on TensorFlow. An encoder was used to convert the English phrases to feature vectors that can be trained upon and a decoder converts the output vector back to normal Hindi text (utf-8).

Encoder - Model

Encoder takes the English data as input and converts it into vectors that is passed to an LSTM model for training. We discard the encoder output and only keep the states.

Decoder - Model

The decoder takes in Input the states of the encoder and the Hindi data points corresponding to the English input of Encoder. It trains an LSTM to produce the translated phrase in output. The decoder used SoftMax layer.

About Dataset – Category Classification

Context

This dataset contains more than 800k article from 2012 to 2022 from This is one of the biggest news datasets and can serve as a benchmark for a variety of computational linguistic tasks

Content

Each record in the dataset consists of the following attributes

- ❖ Category : category in which the article was published.
- ❖ Headline : the headline of the article.
- ❖ Authors : list of authors who contributed to the article.
- ❖ Link : link to the original news article.
- ❖ short_description: Abstract of the article.
- ❖ date: publication date of the article.

There are a total of 42 categories in the dataset.

About Dataset – Category Classification

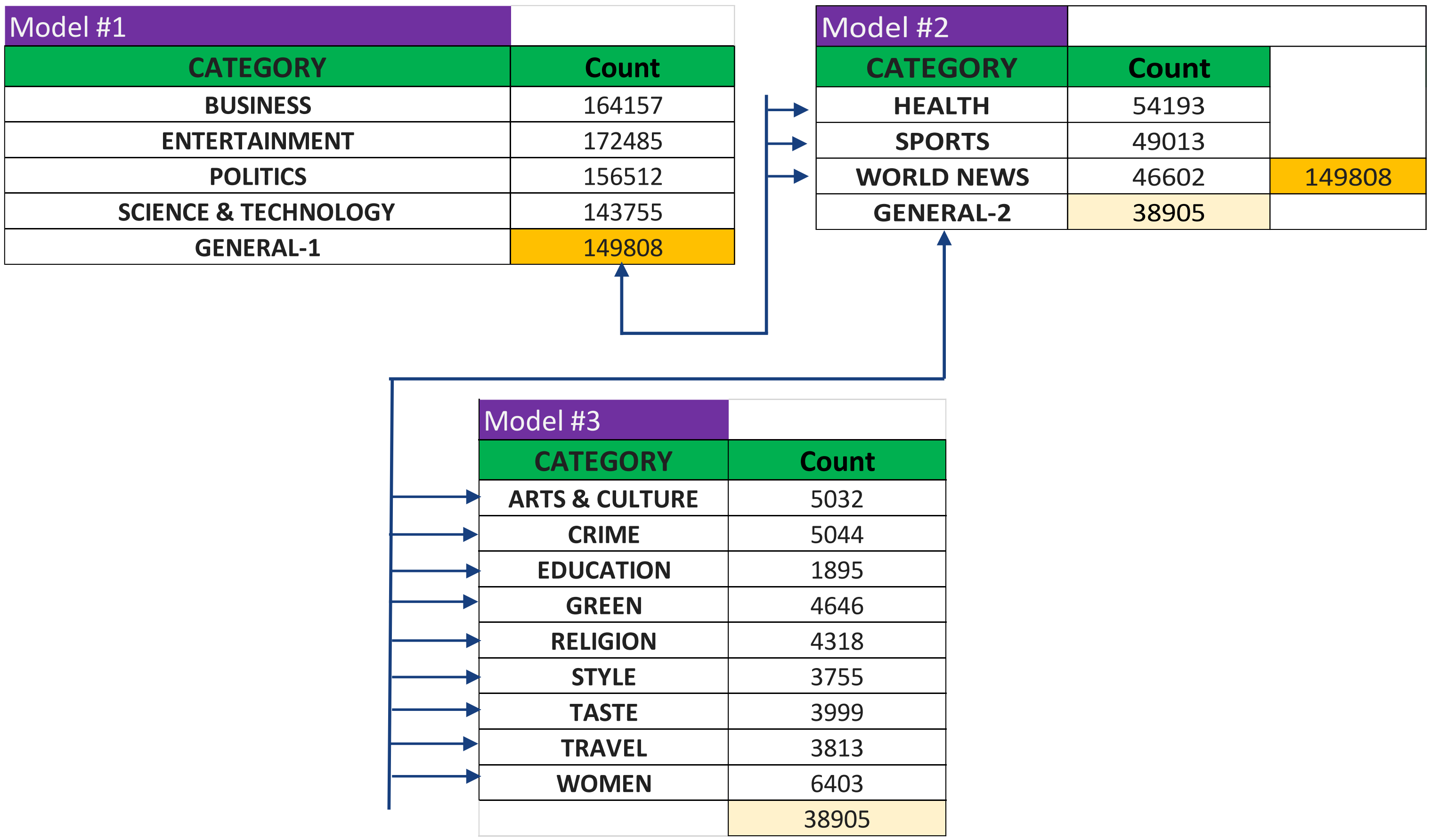
The top-16 categories and corresponding article counts are as follows:

Category	Count
BUSINESS	164157
ENTERTAINMENT	172485
POLITICS	156512
SCIENCE & TECHNOLOGY	143755
HEALTH	54193
SPORTS	49013
WORLD NEWS	46602
ARTS & CULTURE	5032
CRIME	5044
EDUCATION	1895
GREEN	4646
RELIGION	4318
STYLE	3755
TASTE	3999
TRAVEL	3813
WOMEN	6403

Class imbalance is a common problem in machine learning that occurs when the distribution of examples within a dataset is skewed or biased. This can lead to a bias in the trained model, which can negatively impact its performance

Solve Imbalanced Class Problem

To solve imbalance class problem I divided dataset in to 3 blocks with 3 models

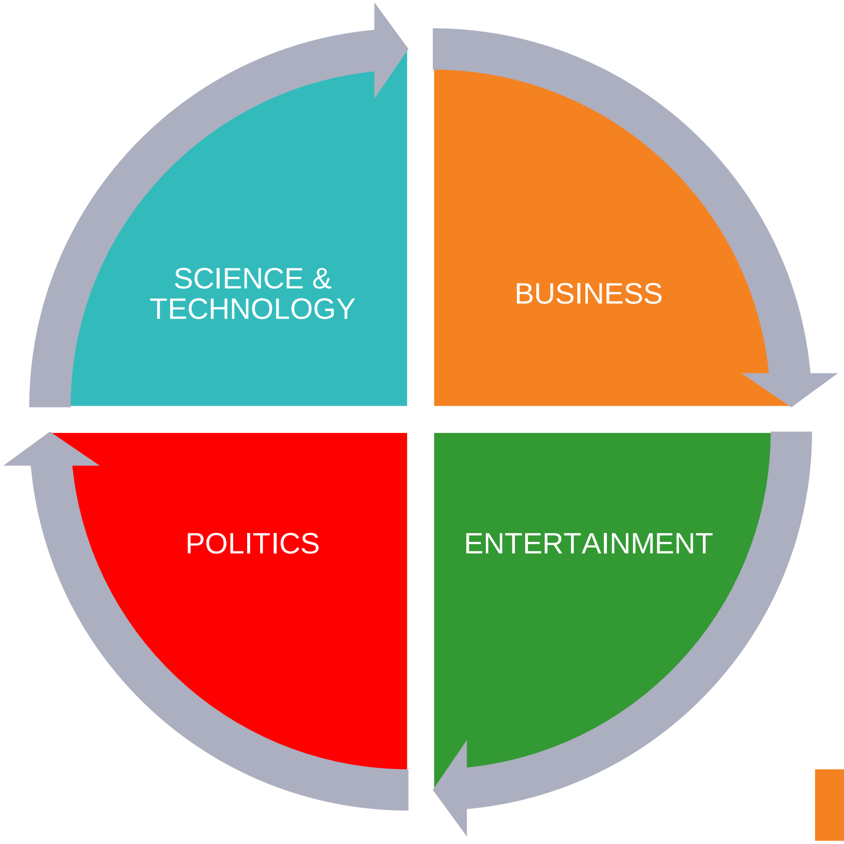


Model #1	
CATEGORY	Count
BUSINESS	164157
ENTERTAINMENT	172485
POLITICS	156512
SCIENCE & TECHNOLOGY	143755
GENERAL-1	149808

Model #2		
CATEGORY	Count	
HEALTH	54193	
SPORTS	49013	
WORLD NEWS	46602	149808
GENERAL-2	38905	

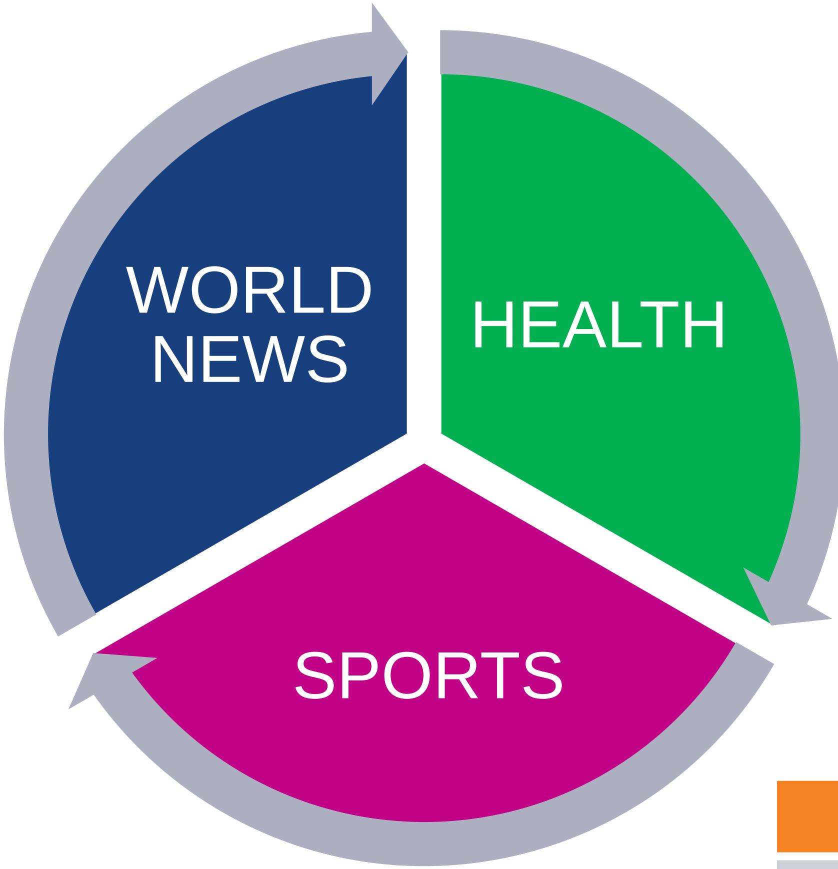
Model #3	
CATEGORY	Count
ARTS & CULTURE	5032
CRIME	5044
EDUCATION	1895
GREEN	4646
RELIGION	4318
STYLE	3755
TASTE	3999
TRAVEL	3813
WOMEN	6403
	38905

Data Set #1 Category Classification



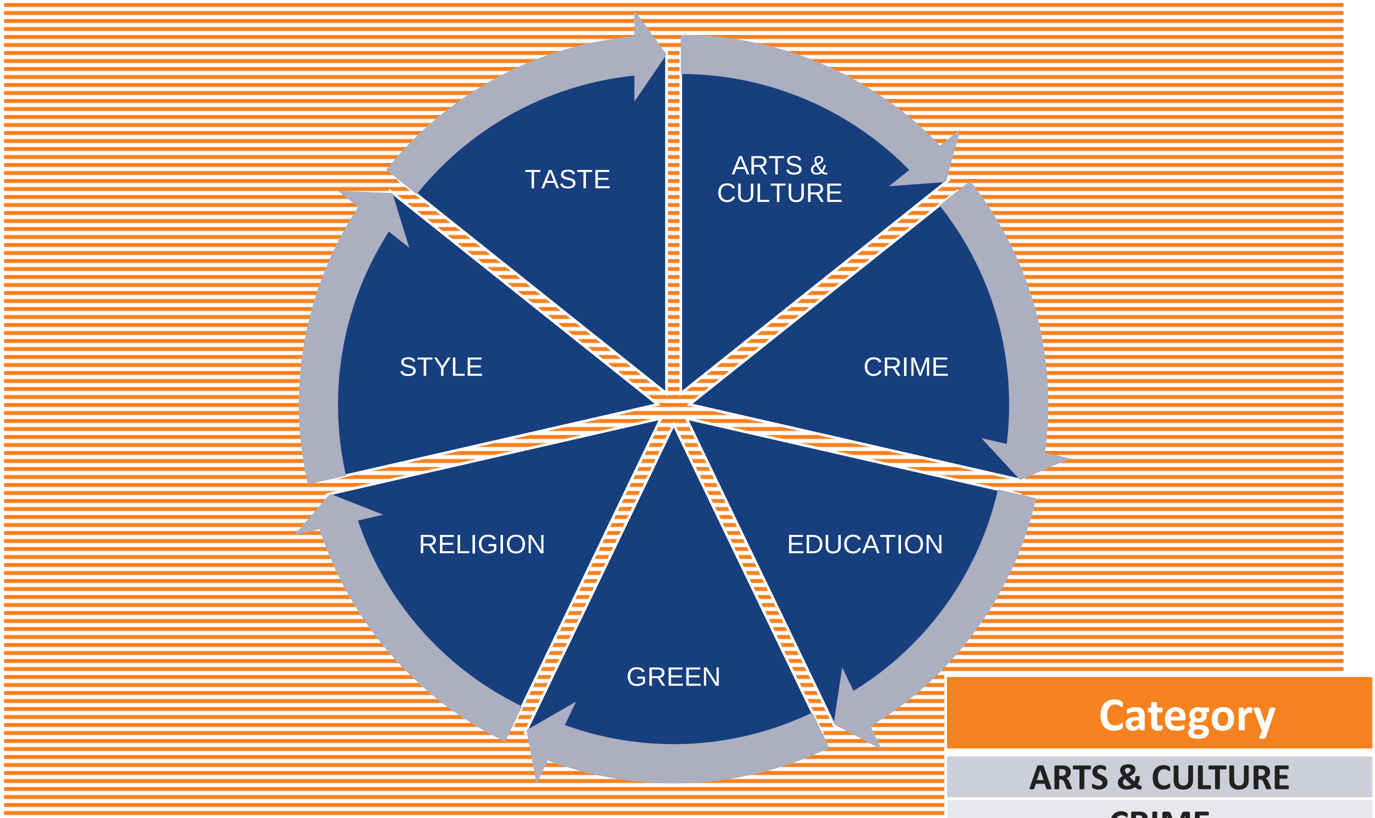
Category	Count
BUSINESS	164157
ENTERTAINMENT	172485
POLITICS	156512
SCIENCE & TECHNOLOGY	143755

Data Set #2 Category Classification



Category	Count
HEALTH	54193
SPORTS	49013
WORLD NEWS	46602

Data Set #3 Category Classification

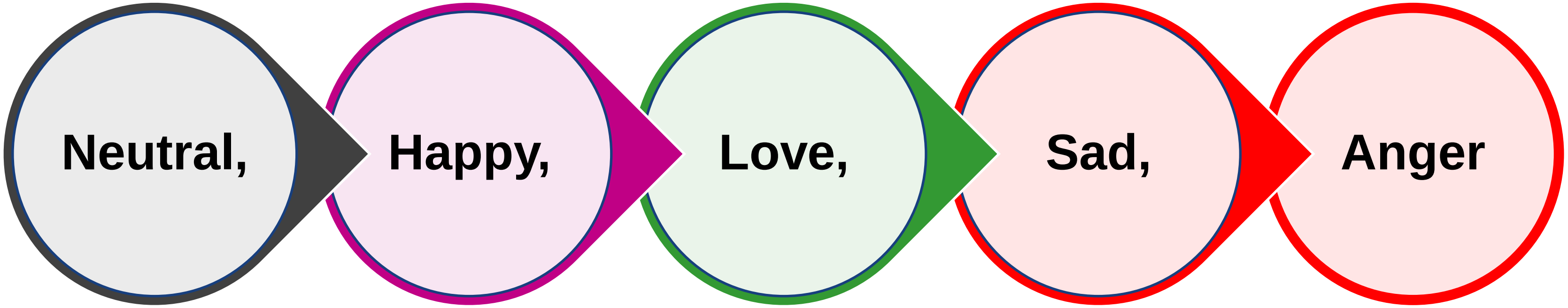


Category	Count
ARTS & CULTURE	5032
CRIME	5044
EDUCATION	1895
GREEN	4646
RELIGION	4318
STYLE	3755
TASTE	3999
TRAVEL	3813
WOMEN	6403

Dataset - Sentiment analysis

Multi-Class Text Emotion Analysis

After deep pre-processing of tweets in done (lemmatization, removal of stop words, etc.)
This dataset is comprised of 62,015 tweets from Twitter with labelled emotions of five classes



Neutral, Happy, Love, Sad, Anger

Dataset - Sentiment analysis

Category	Count
HAPPY	157317
LOVE	56203
ANGER	62560
SAD	138750
NEUTRAL	75886

Class imbalance is a common problem in machine learning that occurs when the distribution of examples within a dataset is skewed or biased. This can lead to a bias in the trained model, which can negatively impact its performance

Dataset - Machine-Translation-English-to-Hindi

Context

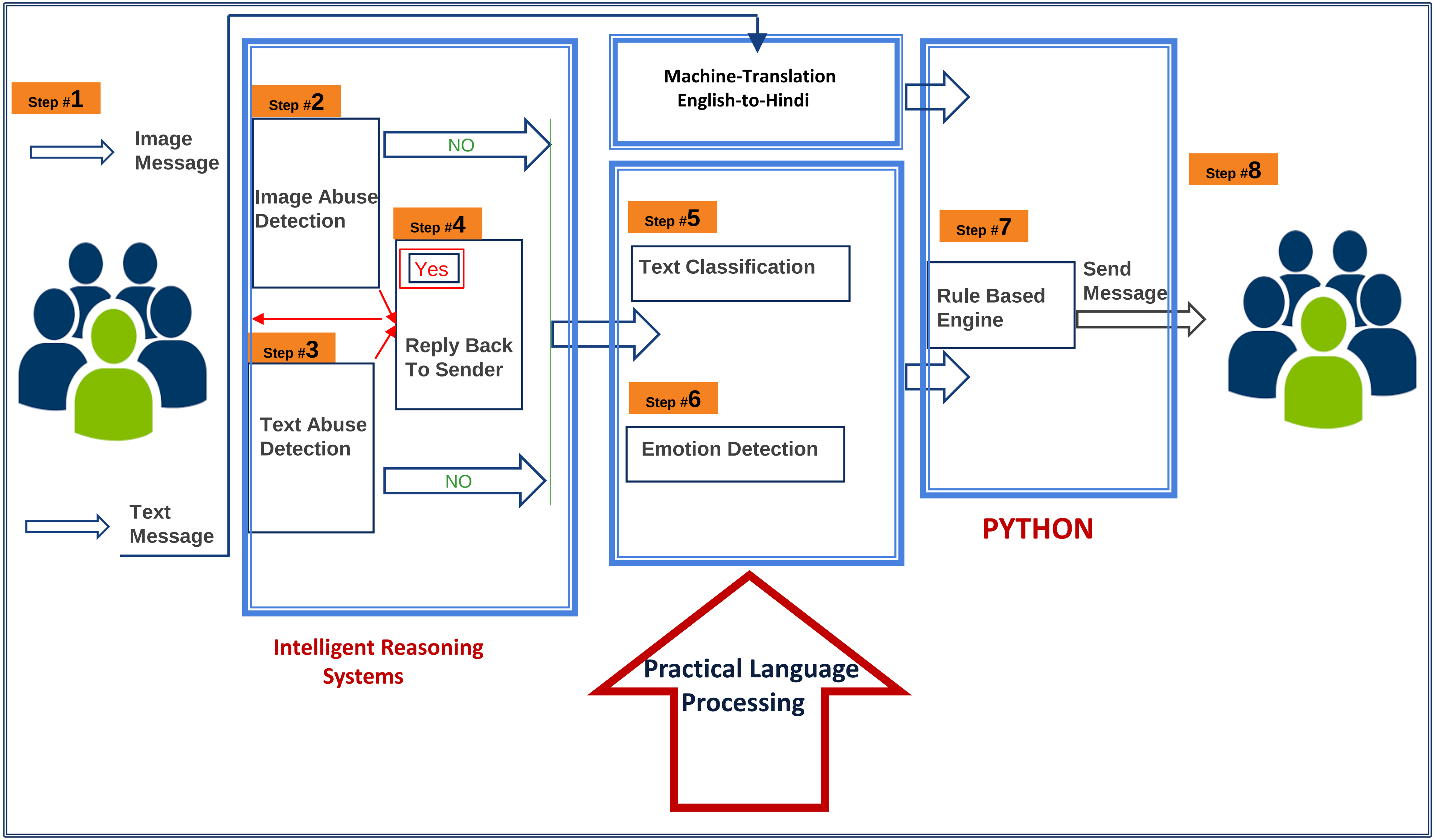
- The dataset consist of 1000000 English phrases along with their Hindi translations. The data is given in utf-8 format.

Content

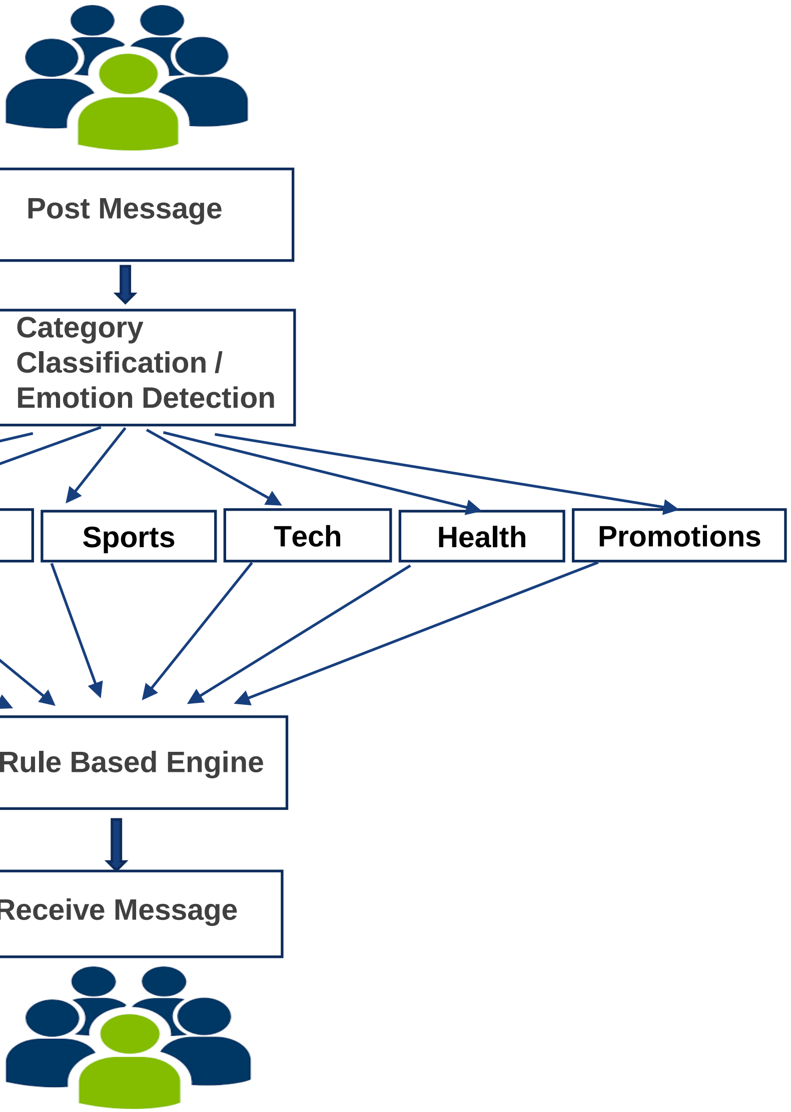
Each record in the dataset consists of the following attributes

- english_sent
- hindi_sent
- eng_sent_length
- hindi_sent_length

Application Flow Phase #1



Application Flowchart Phase #1



Post Message

Category Classification / Emotion Detection

News

Sell / Buy

Politics

Sports

Tech

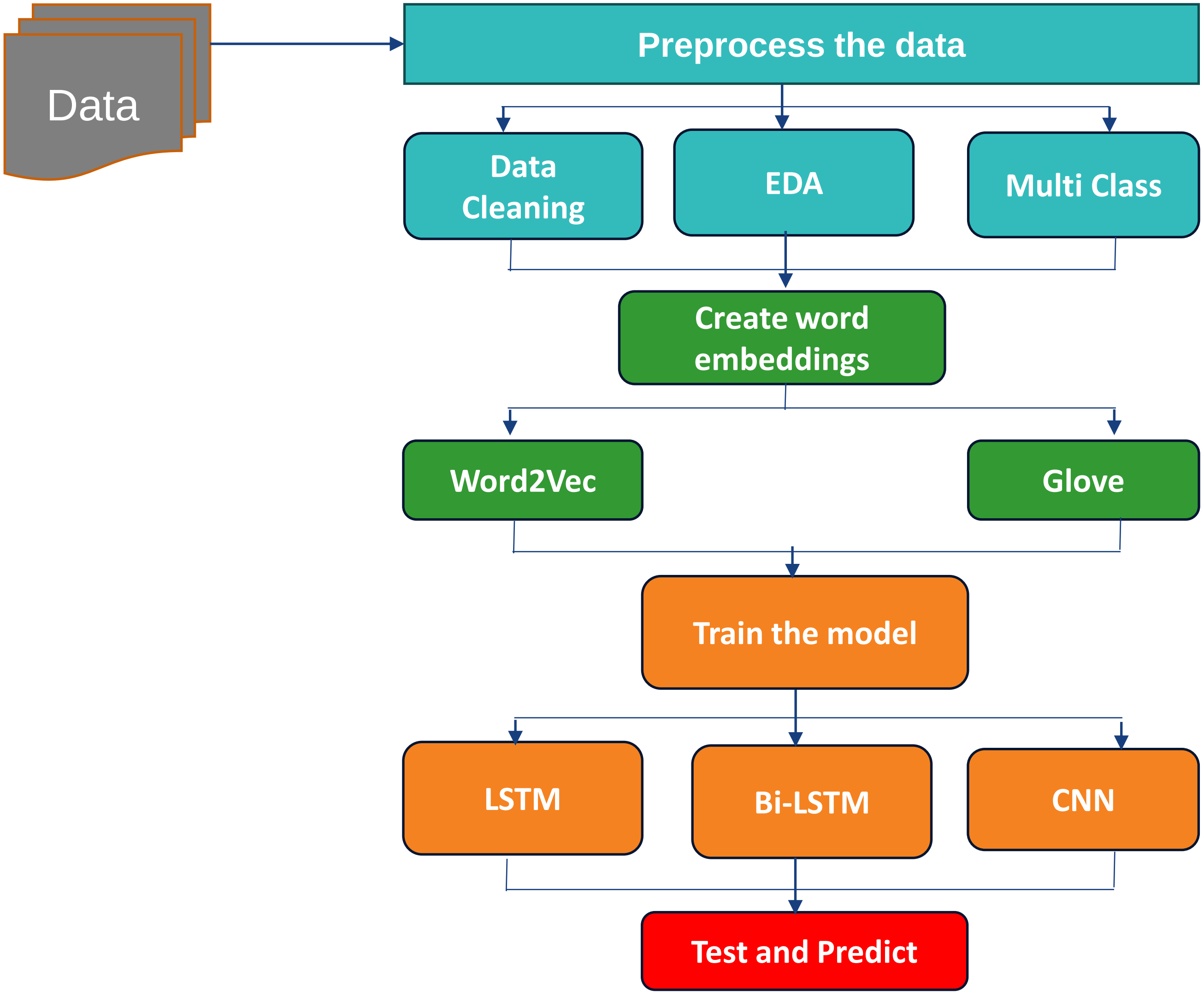
Health

Promotions

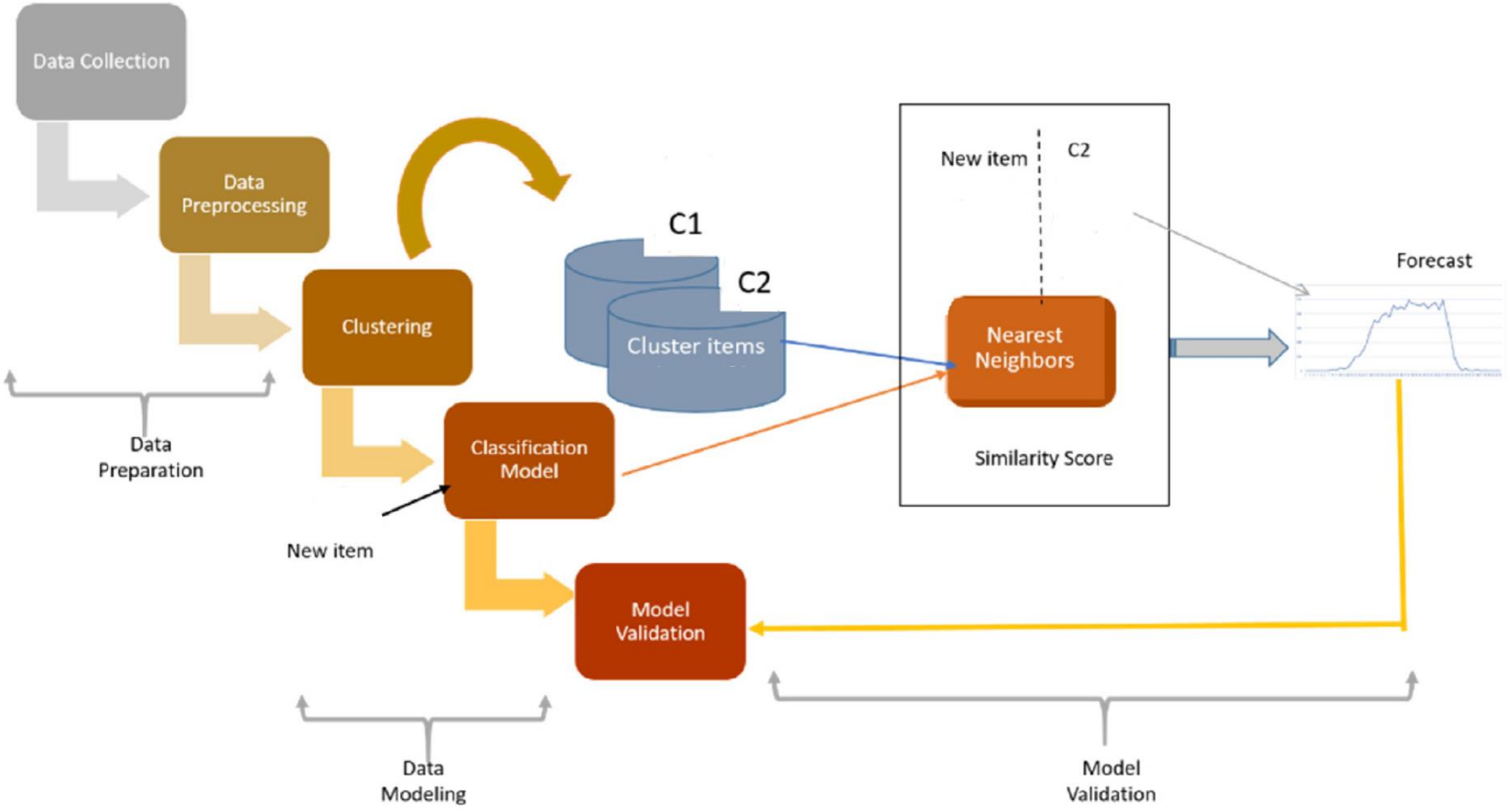
Rule Based Engine

Receive Message

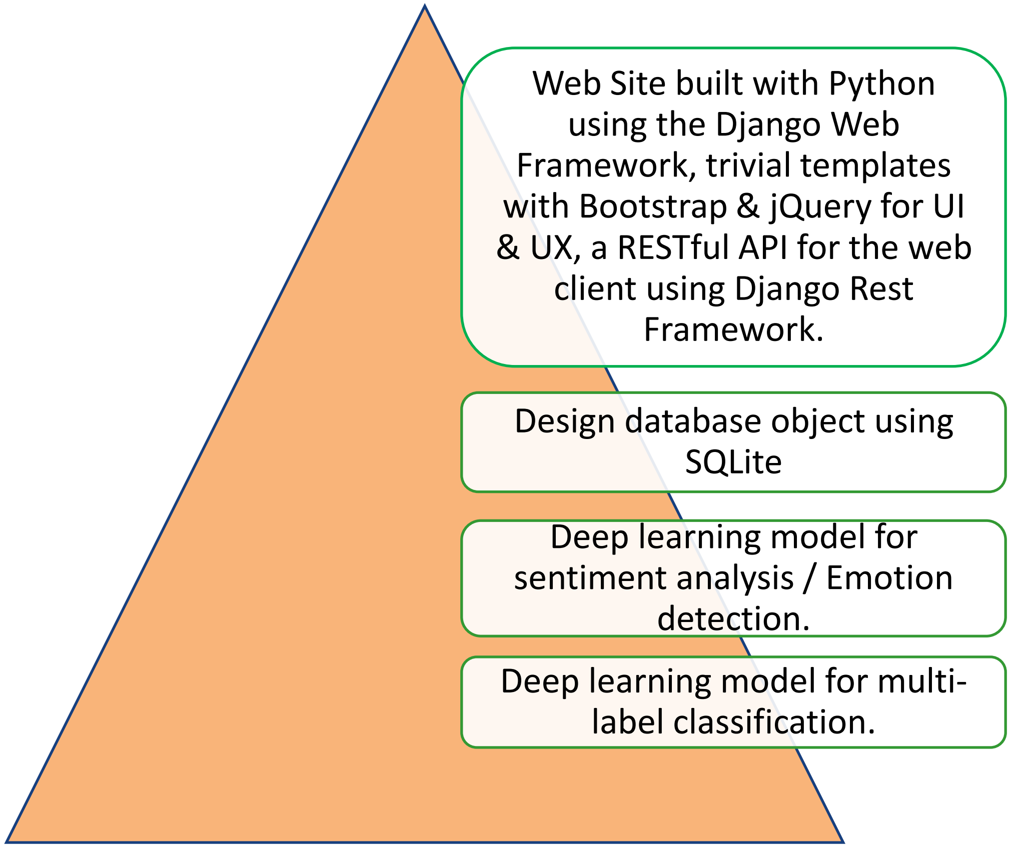
Modeling approach



Experimental design



Project deliverables With Effort Estimates



Web Site built with Python using the Django Web Framework, trivial templates with Bootstrap & jQuery for UI & UX, a RESTful API for the web client using Django Rest Framework.

Design database object using SQLite

Deep learning model for sentiment analysis / Emotion detection.

Deep learning model for multi-label classification.

Task	# Days
Design Web Project / DB Design	30 Days
Data Collection	10 Days
sentiment analysis / Emotion detection	10 Days
multi-label classification.	10 Days

Learning Objectives

- Get familiar with class imbalance through coding.
- Understand various techniques for handling imbalanced data, such as Random under-sampling, Random over-sampling, and Near Miss.
- Sentiment Analysis
- Apply the relevant models that need to be used for each task
- Apply the major guiding principles when choosing a model for a specific task within NLP
- Decide when to and when not to use neural network based or deep learning methods for a specific task within NLP
- Design and implement fundamental algorithms used for various NLP tasks
- Analyze the time complexity involved for a specific NLP algorithm

Done So far

Problem statement # 1 : Category classification (Model #1,2,3)

Completed Task

1. **Data Clean & EDA.**
2. **Convert Text to vectors using different techniques like TFIDF ,Word2Vec**
3. **Trained Machine Learning and Deep Learning models.**
4. **Model Evaluation & Prediction using Test data**

Pending Tasks

1. **Model integration in community help website**
2. **Prediction using Real Time data**

Model #1	
Model	**Accuracy**
Support Vector Machine #2	**0.8923**
Logistic Regression #2	**0.884**
MLP Classifier	0.8413
Naive Bayes(Multinomial)	0.8382
LogisticRegression Classifier #1	0.7834
Support Vector Machine #1	0.7715
Randomforest #2	0.6603
Randomforest #1	0.6172
Model	**Accuracy**
LSTM LAYER1	**0.873**
GRU	**0.8605**

Model #2	
Model	**Accuracy**
MLP Classifier	**0.906**
LogisticRegression	**0.8613**
Support Vector Machine	0.8195
Randomforest	0.7182
Model	**Accuracy**
LSTM LAYER1	**0.8974**
LSTM LAYER2	**0.8921**

Model #3	
Model	**Accuracy**
MLP Classifier	**0.8011**
LogisticRegression Classifier	0.7026
Support Vector Machine	0.6653
Randomforest	0.5664

Done So far

Problem statement # 2 : Sentiment analysis

Completed Task

1. **Data Clean & EDA.**
2. **Convert Text to vectors using different techniques like TFIDF ,Word2Vec**
3. **Trained Machine Learning and Deep Learning models.**
4. **Model Evaluation & Prediction using Test data**

Pending Tasks

1. **Model integration in community help website**
2. **Prediction using Real Time data**

Model	Accuracy
MLP Classifier	0.8679
Logistic Regression	0.855
Support Vector Machine	0.855
Naive Bayes(Multinomial)	0.824
LogisticRegression Classifier	0.8057
GradientBoostingClassifier	0.8024
AdaBoost with Random Forest Classifier	0.769
Randomforest	0.732
Support Vector Machine	0.7039
Randomforest	0.67
Model	**Accuracy**
GRU	0.8106

Done So far

Problem statement # 3 : Translation

Completed Task

1. Data Clean & EDA.

3. Trained Machine Learning and Deep Learning models.

4. Model Evaluation & Prediction using Test data

Pending Tasks

1. Model integration in community help website

2. Prediction using Real Time data

Bleu Macro Score	Bleu Score	Meteor Score
0.04128	0.00139	0.0607

References

https://colab.research.google.com/github/d2l-ai/d2l-en-colab/blob/master/chapter_deep-learning-computation/use-gpu.ipynb#scrollTo=mwrEJrSCo-OR

https://github.com/opencv/opencv

https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

https://www.kaggle.com/code/jarvis11/text-emotions-detection

https://github.com/dair-ai/emotion_dataset

https://github.com/suvaansh/Machine-Translation-English-to-Hindi-/tree/master

https://www.kaggle.com/datasets/parvmodi/english-to-hindi-machine-translation-dataset?select=train.hi

https://www.kaggle.com/datasets/vaibhavkumar11/hindi-english-parallel-corpus