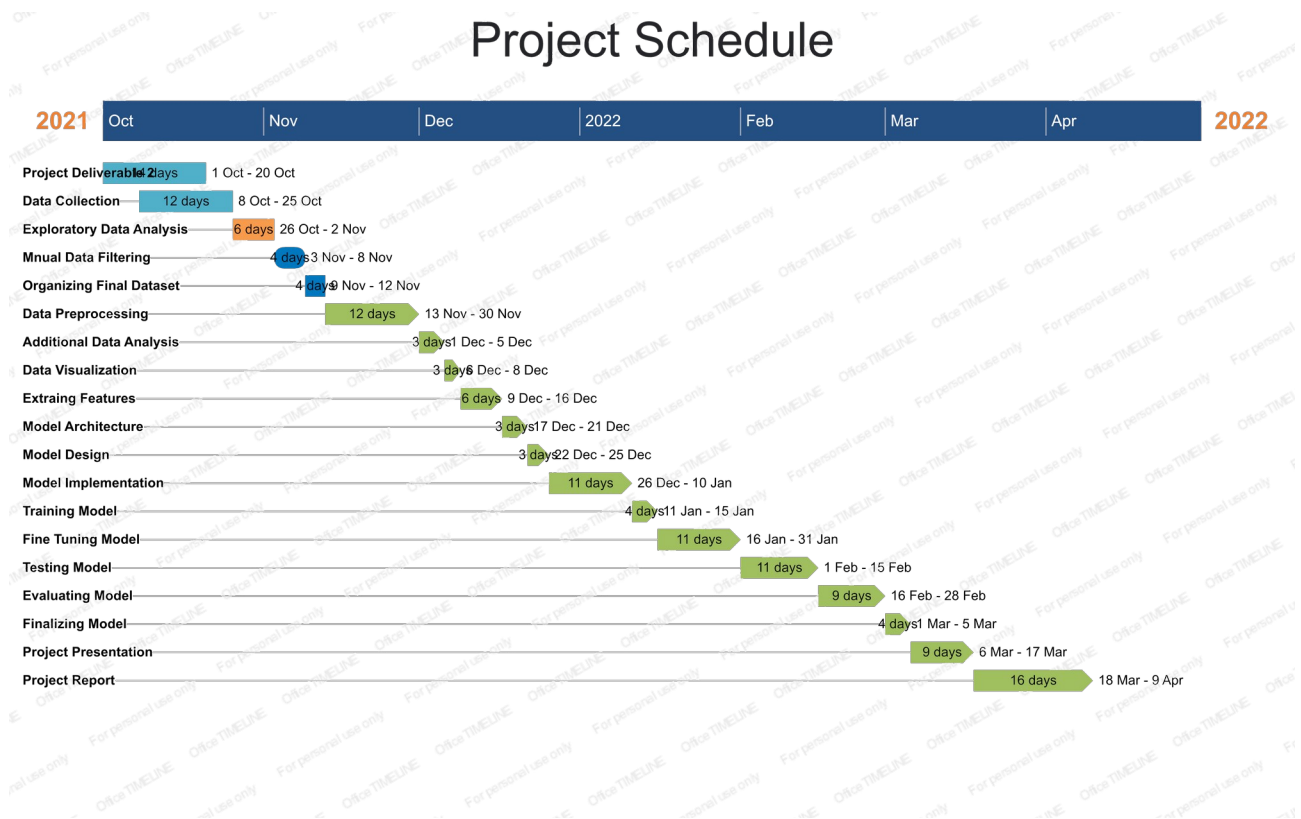


Hate Speech Detection

Yidong Huang (1116935) and Devansh Mody (1130532)

Gantt Chart:

Our gantt chart contains estimated sprints, tasks, timeline, responsibilities and hours.



We both have divided the different stages of development and are actively working on various stages together.

Outcomes of different stages for our project development are given below:

- Project Deliverable: To design our project deliverable and define the exact goals that we want to achieve
- Data collection: Collecting data from multiple sources and creating a final huge dataset.
- Exploratory Data Analysis: Further analysis on data, finding similar patterns and understanding the data organization in depth.
- Manual Data Filtering: Arranging the data in proper format and labeling the data for all the datasets collected.
- Organizing Final Dataset: Completing the final dataset with proper label information and merging huang and devansh datasets collected individually.
- Data Preprocessing:
 - Different pre processing steps are as follows:
 - Normalizing text converting everything to text that is GIF, Images, emoticons, numbers, emojis etc.
 - Converting to lower case

- reducing repeated words and punctuations
- Expanding contractions
- removing date time information
- correcting misspelled words based on text similarity using edit distance or cosine similarity
- tokenization
- standardize and remove @usermentions and URLs
- removing stopwords, accented characters
- grammar correction
- Additional Data Analysis: Performing additional analysis by plotting the data and generating our own data dictionary or vocabulary.
- Data Visualization: visualizing the data, numbers of values etc.
- Extracting Features: Conceptual number batch embeddings will be used for word embeddings and to handle out of vocabulary words characters based embeddings will be used.
- Model Architecture: Architecting our model deciding layers, inputs and outputs, encoding and decoding stage.
- Model Design: Finalizing the design of our model with layers
- Model Implementation: Implementing the model writing code for the implementation first based on our approach A we will develop the model and later on plan B approach will be used.
- Training Model: Training the model on the train data.
- Fine Tuning Model: Improving the accuracy or output of the model
- Testing Model: Testing the model on test data
- Evaluating Model: evaluating the model performance using F1, Precision, Recall and LIME methods
- Finalizing Model: Finalizing the planned approach and incorporating speech to text model for finding hate speech on audio data.
- Project Presentation: Making presentation of the report
- Project Report: Final project report

Resources required for our project are 10GB to 50GB of storage for data and our model, 12GB of GPU and 12 GB of RAM is enough for sample data set size for our project.