**Topic Submission Form**

This form should be submitted by the mentioned deadline.

Name: Hari Thapliyal

Student Number:

Course:     M.Sc (Data Science)

**Fill your topic/s below**

Project Title/Area 1: **Impact of Quality of Hindi Sentences on the Performance of Sentiment Analysis Models**

**Dataset**

<https://www.kaggle.com/disisbig/hindi-movie-reviews-dataset>

<https://github.com/sid573/Hindi_Sentiment_Analysis/blob/master/hi_3500.ods>

**Description**: In our conversation in any language we communicate our intent using sentences or dialogues. The intent of each sentence or dialogue may be negative or positive or neutral. It is a particular very hot area when we are analyzing the review of a product (movie, book, electronics item, etc). Based on the spoken dialogue if we want to know what kind of role X actor has played in a movie then we can know that by performing SA on all his dialogues. In this project we are planning to do the Sentiment Analysis of Hindi dialogues. In this project for training purposes we are planning to take 2 kinds of datasets 1- Movie Review in Hindi (which are not be refined or structured) 2- Hindi Movie dialogues which are highly refined and optimized. We want to know what impact it has on the overall performance of the prediction model when we use two datasets of different quality. Null Hypothesis is there is no impact on the quality of model prediction results. The second major task here is trying different Sentiment based models and evaluate which model gives the best results. The third task is if we use state of art English language-based models like SemBERT, RoBERTa then how does our model perform in terms of speed and accuracy both.

Project Title/Area 2: **An Approach to Detect Sarcasm in Hindi Sentences**

**Dataset**

<https://github.com/rkp768/hindi-pos-tagger/tree/master/News%20and%20tweets>

**Description**:

A language of communication has many beauties and one of the beauty is you can say something which is true but you want to say that in a sarcastic sense. Humans can detect this in a familiar environment. But in an unfamiliar environment even humans fail to understand the intent of a statement. Creating an ML-based model that can detect whether a given statement is sarcastic or not is really challenging task. In this project we look into this problem and predict this. In this project we are ignoring the background in which a statement is made and we will look statements independently. Dataset size is small and I planning to expand this at least double the existing size. We are planning to do this project the Hindi language.

Project Title/Area 3: **Performance Evaluation of the NLU models developed between 2016 -20.**

**Dataset**

Microsoft: <https://www.microsoft.com/en-us/download/details.aspx?id=52398>

Quora: <https://www.kaggle.com/quora/question-pairs-dataset>

STS-B: <http://ixa2.si.ehu.es/stswiki/images/e/ee/Stscompanion.tar.gz>

**Description**:

The ultimate goal of speaking or writing is one should be able to comprehend, understand what is being said. It is one of the most challenging areas in AI. Many standard datasets are in place and many new algorithms are developed in the last 4-5 years. In the project we are trying to understand what are the different state of the art algorithms and for what NLU task they are most suitable for. We are considering following NLU task in this project Reading Comprehension, Semantic Similarity, Classification, Natural Language Inference.

Project Title/Area 4: **Sentiment Analysis of Hindi-English mixed Tweet**

**Dataset**

<https://github.com/Hari-thapliyal/atmanirbharbharat/blob/master/anb_23_05_20.txt>

**Description**:

Indians speak multiple languages. You should not get surprised when you find some uneducated person who is driving a taxi or auto can speaking 3-4 languages. A middle-class average educated youth of India seemly type or speak his emotions in 2-3 languages without even realizing that he or she has switched his language. When we share our opinion on tweeter or product review page that may be written in Roman or Devganari or Kannada script but the thought expressed may be coming from Sanskrit, Bangala, or Hindi. In this project, we are planning to take the two most popular langauge of India Hindi and English. We will develop a model that can do a sentiment analysis of tweets that are written in either of these two languages. But a particular tweet need to be either of the language and not both. If time or resources permit then we can expand our project to analyze a tweet written in two languages.

Project Title/Area 5: **Sentiment Analysis of a writer from the Chapters of a Hindi Book**

Dataset

<https://www.kaggle.com/kapilverma/hindi-bible>

**Description**:

Any story based book which has different authors or different characters in different chapters can be analyzed based on the Character or Author of that particular chapter. In this project we are planning to use different modeling techniques from NLP and general ML and predicting what kind of character or author a person is. Using techniques of unsupervised learning we will create 3 groups for Negative, Neutral, or Positive.

Project Title/Area 6: **Sentiment Analysis of English Language Tweets - Predictivity sentiment of the next tweet.**

Dataset

<https://www.kaggle.com/kazanova/sentiment140?select=training.1600000.processed.noemoticon.csv>

**Description**:

A tweet is 140 characters in length. It contains hashtags, emoticons, url, and normal text which does not respect English spelling and grammatical structure. In this project we want to do sentiment analysis of tweeting accounts. We will predict whether the next tweet from that account will be positive or negative.

Project Title/Area 7: **Labelling Human Emotions of Hindi & English Conversation**

**Dataset**

<https://www.kaggle.com/disisbig/hindi-movie-reviews-dataset>

**Description**:

Human has 10 fundamental emotions Joy, Excitement, Surprise, Sadness, Anger, Disgust, Contempt, Fear, Shame, Guilt. Knowing the emotion of a person is extremely important before you make the next statement. At this time we don’t have any dataset which has labeled these 10 emotions. Therefore in this project we will perform unsupervised learning and see how unsupervised learning can help in performing human emotion analysis.

Project Title/Area 8: **Sentiment Analysis of each Chapter of a given English Book**

**Dataset**

<https://www.kaggle.com/zusmani/the-holy-quran/data?select=English.csv>

**Description**:

Many times we argue over the overall message of a chapter or a book. In this project, we will do a sentiment analysis of each chapter. This is unsupervised learning and scale will vary from 1-10.

**Fill in this section if a member of staff has agreed to be your supervisor:**

Member of Staff:                                                                                   \_\_\_\_

If you have found a supervisor then you and the member of staff who agreed to supervise your project should sign below.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_                                                                        \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Signature                                                                         Supervisor Signature

\_\_\_\_\_\_\_\_\_\_\_\_\_                                                                            \_\_\_\_\_\_\_\_\_\_\_\_

Date                                                                                               Date