CSE 523 : Machine Learning

Group 17 - Hardly Humans
Weekly Project Report - 1

Quora Insincere Questions Classification

Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

1) Tasks Performed in the week.

- Understanding the problem statement.
- Finding dataset.
- Importing important modules and libraries.
- Finding research papers.
- Analyzing the data.

2) Outcomes of the tasks performed.

- Dataset:
  https://drive.google.com/file/d/1zaoYf1kR_umxyuQs5hA5YkmaF4-x4R8K/view?usp=sharing

- Research Papers:
    1. http://ceur-ws.org/Vol-2517/T5-3.pdf
    2. http://ceur-ws.org/Vol-2517/T5-1.pdf
    3. https://www.researchgate.net/publication/334549103_Quora_Insincere_Questions_Classification
- Data analysis:
    1. Q ID : Unique question ID assigned to each question.
    2. Question Text : The actual question.
    3. Target : Either 1 or 0. 1 denotes that the question is   insincere while 0 denotes that it is sincere. Counted the number of sincere and insincere questions in the dataset: Sincere- 1225312, Insincere- 80810.

3) Tasks to be performed in the upcoming week.

- Identifying existing literature on similar classifying tasks (BBC Document classification or Turkish News text).
- Data understanding: What are the parameters that are used to classify a particular question as insincere - and try to identify any patterns?
- Data cleaning.
- Data pre-processing.
- Searching for the applicable machine learning algorithms for the model.