

**Project:** Text Classification Competition

**Team:** Subhasish Bose (sbose4) and Soumya Kanti Dutta (skdutta2)

This report contains information on project progress as of 11/25/2020.

**Progress:**

We have completed the following steps:

- 1) Get train and test datasets from GitHub Repo.
- 2) Inspect the datasets to understand the format and relation.
- 3) Performed data cleaning steps on the dataset.
- 4) Performed train-test split on training dataset to test accuracy.
- 5) Tested the dataset with the following algorithms:
  - a) Linear SVC
  - b) Naïve Bayes
  - c) Logistic Regression
  - d) Random Forest
  - e) Neural network – BERT

**Remaining Tasks:**

The following tasks are pending:

- 1) Run the above-mentioned algorithms on Test Dataset.
- 2) Our software will go with the majority decision found from these 5 algorithms mentioned above. For example, if any 3/4/5 of these algorithms find the tweet as 'SARCASM' – our code will tag it as 'SARCASM'.
- 3) Create output in desired format.
- 4) Create project Report.
- 5) Create Software Usage tutorial presentation.
- 6) Code, Project Report and Presentation submission.

**Issues/ Challenges:**

As the test dataset is not labeled, it is difficult for us to judge accuracy of our code on it.