Project Mid-Point Submission and Presentation (20%)

1. Pick a real-life dataset of size 100+ MB
2. Should contain a mix of slangs, abbreviations, missing/incomplete text, spelling errors and potentially other categorical and numerical features with missing values etc.
3. Understand the business domain, problem and the need for a custom/differentiated solution, plan the project and build a project board (share its link) (30 pts)
4. Perform data pre-processing, feature engineering and visualize separability/predictability of the data for different preprocessed featuresets (40 pts)
5. Build, evaluate and analyze the 1st model (30 pts)
6. Presentation will be 10 mins + 5 mins for questions (only 1 member from group should present but everyone can answer questions)

Final Project Submission and Presentation (40%)

1. Due a data analysis and a model system built on a real-world text data using the NLP workflow + modelling techniques we have gone over in class, with a deployment and demo component and a well-maintained git repository (that has at least 1 commit/week from every group member for at least 5 weeks).
2. Presentations: Done by any 1 member of the group, 10 mins of presentation time, 5 mins for demo and 5 mins for questions