DSO599 – Text Analytics & Natural Language Processing Project Brief

In groups of at least 4 and no more than 6 students,

1) **Pick any public dataset** to research and build a model off of.
    a. Must at **least 100,000 rows, or at least 20MB**, whichever is smaller.
    b. Must contain **text data** (obviously) as the primary feature.

2) **Identify a potential business problem that can be solved by utilizing the data in this dataset.** You may make any reasonable assumptions you'd like: for instance, if you have patient chat transcripts, you may assume that you are a healthcare provider that needs to improve customer satisfaction / success metrics.

3) **Build at least one machine learning model –** it may be supervised / unsupervised, regression or classification, using the concepts we have learned this course, that helps to solve this business problem. This model must at the very least perform better than baseline accuracy standards.

4) **Prepare a 9-minute business presentation, which will be followed with a 5 minute Q&A** to be delivered on April 30th detailing the business use case, the model methodology, implementation roadmap, and potential return on investment of your research/model. **You must assume that both technical staff (analytics and data science team members) as well as executives (the VP of Marketing / Operations / etc.) are in the audience.** In addition:

    a. **3 presenters** will be picked at random from your group to present. You will not know who the presenters will be until the presentation.
    b. The other **2-3 group members** will be responsible for the Q&A portion.
    c. Therefore, it is the responsibility of all group members to be prepared to speak regarding **any portion of the presentation**.
    d. The group will also be evaluated on the quality of the questions that are asked to other teams. At the very least, 5 questions must be asked (there are 9 groups).

5) **All source code (Jupyter notebooks, Python scripts, etc.) must be submitted to a Github repository** that your team will create. **All modified / cleaned datasets must be submitted to an AWS S3 bucket**. Instructions and walkthroughs for how to perform both tasks will be covered in class.

6) **A group 360 evaluation will also be conducted at the conclusion of the project to ensure that students who contributed significantly to a project are rewarded accordingly.** This evaluation will involve each member of the group evaluating the contributions and collaboration of other team members. A sample 360 evaluation will be discussed in class.

Timelines:

- Formal project kickoff (April 16th)
- Dataset submitted for approval by instructor (April 19th)
- **HW4** submitted as a group (April 23rd)
- **HW5** submitted as a group (April 30th)
- Final presentations (April 30th)
- **360 evaluations** due (May 1st)