DSO560 – Text Analytics & Natural Language Processing Project Brief

In groups of at least 3 and no more than 5 students,

1) Pick any public dataset to research and build a model off from:
    a. Must at least 50,000 rows, or at least 20MB, whichever is smaller.
    b. Must contain text data (obviously) as the primary feature, although you are free to include other features as well.

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 slide deck / presentation 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.

All source code (Jupyter notebooks, Python scripts, etc.) must be submitted to a Github repository that your team will create. Instructions and walkthroughs for how to perform both tasks will be covered in class.

5) 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 – groups finalized (**Monday, April 25th**)
- Dataset submitted for approval by instructor (**Monday, May 2nd**)
- Homework 4 due (6:29pm, **Tuesday, May 3rd**)
- Final exam (7pm, **Tuesday, May 10th**)
- Final code and presentation slides submitted (11:59pm, **Wednesday**, **May 11th**)

Past Projects

- Sentiment Analysis of Restaurant Reviews (to identify drivers of satisfaction/dissatisfaction)
- Customer Segment Analysis of Hotel Reviews (analyzing which customer segments to allocate more marketing budget towards for major hotel chains)
- Summarizing top issues and public policy decisions from briefs of criminal court cases