Group Proposal

Submitted by Team 5:
JiWoo Suh
Sanjana Godolkar
Upmanyu Singh

What problem did you select and why did you select it?

We've chosen to predict corporate bankruptcy by analyzing the Management Discussion and Analysis sections of financial filings. This choice is based on the assumption that the text within these sections contains vital insights regarding a company's financial health. The objective is to develop a model that can effectively predict whether a company will go bankrupt based on the sentiments or indications gleaned from this textual data.

What database/dataset will you use?

Our datasets comprise two distinct groups: companies that faced bankruptcy between 2018 and 2022, totaling 138 companies, and a collection of S&P 500 companies. The data is extracted from the 'Management Discussion and Analysis' (MD&A) sections of financial filings, with a primary focus on the 10K and 10Q filings. The 10K reports are annual submissions, presenting audited financial statements, while the 10Q reports are filed quarterly, providing unaudited financial statements. Our analysis centers on comprehending the language, tone, and contents of these financial documents.

What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

The primary NLP methods we aim to employ are text classification and sentiment analysis to determine bankruptcy indicators. Initially, we plan to utilize classical models. However, we remain open to customization as we progress through the project.

What packages are you planning to use? Why?

Our toolkit comprises NLTK, Scikit-Learn, Pandas, Numpy, and TensorFlow, among others. Each of these packages offers a range of functionalities. NLTK and Scikit-Learn will support natural language processing tasks, while Pandas and Numpy will aid in data manipulation and visualization. TensorFlow may be utilized for any custom model development.

What NLP tasks will you work on?

The primary NLP tasks include text classification, sentiment analysis, and the extraction of predictive indicators from the 'Management Discussion and Analysis' section in financial filings.

How will you judge the performance of the model? What metrics will you use?

The current consideration for judging the model's performance is to employ Recall as a metric. This metric will provide an understanding of the model's ability to correctly identify bankrupt companies among the actual bankrupt entities in the dataset.

Provide a rough schedule for completing the project.

The project is expected to take approximately 4-5 weeks to complete, including data preprocessing, model training and evaluation, and presentation of results.

The schedule for completing the project is as follows:

Week 1-2: Data pre-processing and exploration

Week 3-4: Training and evaluation of Model

Week 5: Presentation of results and conclusion.