FE 690: Machine Learning in Finance
Text Analysis and Sentiment Analysis
Due Friday, October 30 5PM Eastern Time

Introduction

For your second assignment for the semester, you will work *individually* to use machine learning to build a tool to do text or sentiment analysis for financial problems. Broadly your projects will consist of three inter-related topics:

1. accessing and discussing applicable data;
2. selecting an appropriate machine learning method; and
3. results and analysis.

You will get the most out of the project if you interact with Professor Feinstein during this assignment, especially when planning a topic.

Project Components

Data Collection and Discussion

Download financial data from the database of your choice (e.g., Bloomberg, WRDS, *Yahoo Finance*, ...). Additionally, download textual data from the source of your choice (e.g., quarterly reports, Twitter, ...). You should spend time deciding on the appropriate data to analyze and whether it is sufficient for the method you will want to implement.

If you wish to consider sentiment analysis, consider finding a pre-analyzed training data set to assess positive and negative sentiments. For that purpose, you may wish to refer to sample scored documents (e.g., `https://www.cs.cornell.edu/people/pabo/movie-review-data/` or Kaggle) or prebuilt sentiment analyzers (e.g., VADER Sentiment Analysis or TextBlob).

Machine Learning Methods

Given the data stream you are analyzing, choose a machine learning method to analyze your text. Give serious thought to your proposed method as you will need to justify your choice.

Results and Analysis

Implement your methodology on your collected data in order to test the results. Depending on the question under consideration, you may want to compare your chosen methodology to a simple baseline in order to determine performance (e.g., analyzing the same data without any textual input). You should remark on whether your methodology appears suitable to answer the desired question; statistical analysis is *strongly* encouraged.

Report Details

You will submit your final write-up, which should include all of the information detailed below. This should be presented in roughly the order given, but your write-up need not have corresponding sections or bullet points. The write-up should be about 5-7 double-spaced pages, Times New Roman 12pt font. This does not include any appendices (of, e.g., your *Jupyter Notebook*) you may wish to include. Any external resources used should have clear citations and a reference page at the end of your work. This report **must** be submitted in pdf format; your code may be requested if not clear in the document so please keep that available.

1. **Overview** of the problem statement.
2. Detailed description of the **data collected** and why it is appropriate for the problem being considered. Mention any data cleaning if required.
3. Detailed description of the **machine learning method** and why it is appropriate for the problem being considered. If comparison to a *baseline* model is to be studied, provide the details of this methodology as well.
4. Describe the **results** obtained by your methodology on the data. Analyze these results to provide a recommendation.
5. **Next steps**: What do you recommend as a result of your analysis? Do you suggest attempting different algorithms or a larger test or more data? etc... What else could be done with the problem, but time did not permit?

Project Presentation

On Wednesday, October 28, eight-nine (8-9) of your classmates will have uploaded a 15-20 minute presentation to Canvas. These presentations are due by 6:30PM on Wednesday, October 28. The class time on October 28 will be left open for viewing these presentations. You should comment on each of these videos with *constructive* feedback and/or questions. Selection of presenters will be announced on Canvas on Monday, October 12. If you strongly prefer to present (or not present) in this second group, please contact me as soon as possible so that I can try to accomodate any request. If you have presented in the first group, then you do *not* need to present again. I cannot guarantee all requests will be granted. Beyond any such requests, selection will be made through a random number generator from the official course list. Note all students will present exactly once during the semester, those who do not present in either the first or second group will be required to present as a part of the third group.

Presentation Responses

On Wednesday, October 28, the class period should be used to watch the presentations of your classmates. Where appropriate, please provide feedback and questions for your classmates.