CS 410 Progress Report

November 15, 2021

Team AHR

- **Anthony Petrotte (adp12@illinois.edu)**
- Hrishikesh Deshmukh (hcd3@illinois.edu)
- Rahul Jonnalagadda (rjonna2@illinois.edu)

Report

To reiterate, the goal of our project is firstly, to analyze the financial news cycle in different time intervals to create a time interval sentiment metric on particular global securities. Secondly, we will compile the intermediate sentiment results during the metric calculation into a time series dataset that can be compared to price movement in the underlying security. Throughout the process of completing the project, our group has collaborated to set up regular meeting times, which has led to timely organization. We have also identified and distributed key tasks during those meetings, which has tremendously increased our efficiency as a group.

Our team has made consequential headway on our final project and the deliverables we had outlined in the project proposal. We began by setting up our group repository and creating the skeleton structure of our project files. The first portion of the project we have completed has been the utilization of the News API Client and the Google API to retrieve textual information. We have successfully compiled multiple news sources and retrieved information about stock tickers, stock prices, company name, and more using user-generated queries. The next component of our project that we have made progress towards has been determining the relevance of the textual information we retrieve. We wanted to create relevancy scores for our retrieved documents using a Python library with an implementation of BM25. However, we faced compatibility issues due to conflicts between libraries and the specific Python version installed. We resolved this issue by successfully implementing a modified version of BM25 within our own codebase, which we use by looping through the HTML sections and then using BM25 to establish relevance to the target. As a result, our group has also been able to classify retrieved documents and give them a relevancy score.

The upcoming tasks we have to complete is to subsect relevant articles into their relevant and irrelevant parts, which will lead to smaller, reviewable sub-document sets. We have completed a rough implementation of this task, which includes a model, tokenizer, and classifier. Next, our team will develop a weight metric centered around the subject count, where the assigned weight is lowered with additional subjects, while a higher weight is assigned if the user

target is the only subject of the document. Finally, we will use huggingface pre-trained models to analyze the sentiment of the relevant subset. The challenge with using these pre-trained models will be the need to review the output of the prebuilt models to see if they capture sentiment specific to the financial terminology our project is focused on. Overall, we are on track to complete the final project and its key deliverables on time.