CS 410 Project Proposal

October 24, 2021

Team AHR

Members

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

Detailed Description

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. This would be a useful tool or addition to the task of stock screening, and could be implemented as an addition to a computational trading strategy. Secondly, we would 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. This dataset would potentially have many uses, including the possibility to aid in identifying securities that are more prone to volatility from individual (and potentially more naïve) investors. Our project would be considered a free topic relevant to the course that has a novel and useful purpose.

Expected Outcomes

- A usable metric that could be used to gauge the financial news sentiment within a specified time interval.
- Dataset generation for statistical analysis between news sentiment and asset prices within the specified time interval.

Toolsets

- Languages
    - Python
        - This is the most reasonable language to use for our project. It includes various libraries that simplify each major task. Additionally, as this class has been taught in Python, it would facilitate our work to reuse needed components from the coursework.
- Potential Python Libraries
    - Numpy
    - Beautifulsoup
    - requests
    - Metapy
    - Pandas
    - Tensorflow
    - tkinter
    - Huggingface
        - For prebuilt NLP models
- Potential News APIs
    - Google News
    - Polygon.io

Evaluation

Due to the ambiguity of textual information, much of the evaluation will have to be done by hand. This will be made possible by creating datasets containing source urls and records of corresponding 'sub-document' sections and the weights and sentiment scores used to calculate the total metric. The ability to go through the sources will allow us to debug, optimize and empirically judge the accuracy of the approach.

Planned Approach

- **Set Target**
- **Compile News Sources Centered on Target** (**est. 12 hours**)
    - Utilize news APIs to retrieve textual information
    - Due to simplicity of stock ticker symbols, need to have function to adjust query to include company name and potentially helpful key phrases
- **Analyze Sources** (approx. 45 hours)
    - Determining Relevance (**est. 30 hours**)
        - Problem: Most financial news articles are about more than one thing and many are irrelevant, can't gauge the relevance or sentiment on the whole article
        - 2 Rounds to gauge relevance before analyzing sentiment
            - Document Relevance
                - Established by query from API
            - Based on document relevance, find sections containing relevant information (having to do with target)
            - Compile relevant sections into smaller reviewable sub-document sets
                - Scanning Source Information
                    - Could use predefined list of 'Subjects' to determine change of focus
                        - Identify subject, until new subject detected, all information assumed to be relevant towards initial subject
                    - Could use an implementation of PLSA
                    - Could use an implementation of BM25
                        - loop through html sections, use BM25 to establish relevance to target
                - Weight metric
                    - Subject Count
                        - Weight lowered with additional subjects, highest weight if Target is only subject of document
                    - If Subject Count is >1
                        - Target Count in document (or document section)
                        - Target Placement in document
                            - Higher weight placed towards top
    - Analyze Sentiment of relevant set (**est. 15 hours**)
        - Utilize huggingface pre-trained models
        - Need to review the output of prebuilt models to see if they capture sentiment specific to financial lingo
            - May potentially re-train existing model to properly gauge sentiment
        - Output of the sentiment analysis should be directly applicable to the metric

        - 1 to -1
        - 100% positive to 100% negative sentiment
- **Combine Weighted Relevance and Sentiment to create Metric (est. 3 hours)**

Workload

This project will require each step in the Planned Approach section to be relatively reliable, and thus has potential to quickly reach the 20N (60) hours required per N members of the group. Due to the inconsistent nature of both textual data and the html layout of various news websites, creating a reliable pipeline may require extensive debugging to ensure the data used to calculate weights and sentiments is relatively uniform and reviewable. The estimated amount of hours are included at the major sections of the Planned Approach, based on our team of 3 members.

Our Project Repo

https://github.com/adp12/CourseProject.git

Cited Sources

https://github.com/huggingface/transformers
https://github.com/mkhorasani/Trading_Sentiment_Analyzer/blob/master/TSA%20v1.0.py
https://newsapi.org/s/google-news-api
https://docs.google.com/document/d/1ubzdWekH2WLzft-IaSnkYflKrR7DqW-X7WsWiG30loI/edit#