Checkpoint A: Topic Choice, Initial Schema Definition, and Likely Data Sources

Jake Remsza

Sean Padgett

MSDS 459: Knowledge Engineering

Abstract

The purpose of this paper is to introduce and outline plans for a knowledge base concerning public companies in the healthcare sector. This outline is provided in two main sections. These are a literature review and a section entitled methods. In the literature review the researchers detail related published works. The methods section presents the intended project approach for the researchers covering the five questions presented in the assignment posting in addition to the eight steps outlined by Chakrabarti et al. (1999).

Introduction

Traditional methods of evaluating the future performance potential of stocks were mainly computational. Over time various other methods have been proposed including analyzing sentiment of publicly available information. One of the earliest and most influential papers challenging the status quo was that by Cootner (1962). He challenged the prevailing theory of the trajectory of stock prices being purely a random walk forecastable by precise calculations. Cootner (1962), instead proposed that this random walk wavered between less precise price barriers. This wavering was driven by professional traders not utilizing the same methodology, and non-professional traders assessing value with less precise measures (Cootner, 1962).

The work by Cootner (1962) laid the foundation that has since been expanded on. This began as early as the 1970's with work by Zweig (1973) to analyze closed-end funds as an indicator of investor sentiment. In this work, the researchers present a plan to expand upon this area with the intent of analyzing sentiment related to specific stocks present in various focused blogs to forecast whether a particular stock would make a good investment. This

begins with a literature review of some of the more recent efforts of other researchers in this area followed by a presentation of the planned methods.

Literature Review

Various other researchers have investigated the merits of utilizing sentiment found in online text documents for forecasting stock performance. Some of these studies such as that by Feuerriegel and Gordon (2018) have utilized company and regulatory reports. Additionally, the works presented by Schumaker et al. (2012), Wüthrich et al. (1998), and Wu et al. (2021) have utilized news articles. Finally, some such studies have also utilized social media and comment boards. Examples of this final type are works by Deng, Sinha, and Zhao (2017), Li and Li (2013), and Bollen, Mao, and Zeng (2011). Following is a review of one study from each of these three approaches.

The research by Feuerriegel and Gordon (2018) analyzed sentiment in the text of 75,927 ad hoc regulatory announcements in both English and German. Researchers in this study utilized several different machine learning techniques to analyze this text data including “lasso, ridge regression, elastic net, gradient boosting, principal component regression, and random forest” (Feuerriegel and Gordon, 2018). They then used the results of these analyses to predict the performance of the German stock indexes DAX, CDAX, and STOXX Europe 600. The results of utilizing sentiment analysis were compared to results from a linear autoregressive model utilizing historic time series price data (Feuerriegel and Gordon, 2018). Feuerriegel and Gordon (2018) found their methods utilizing sentiment analysis were able to outperform established linear models when forecasting future stock prices.

Schumaker et al. (2012) chose to analyze sentiment in news articles for their research. The researchers in this article utilized the existing stock prediction tool AZFinText and augmented the results from that system utilizing the OpinionFinder tool. The combined results were then used as instructions for a trading engine to make a trade 20 minutes after each prediction was made. The researcher's indicated that their approach netted returns from those trades between 2.4 and 3.3 percent (Schumaker et al., 2012).

In their research Bollen, Mao, and Zeng (2011), utilized the tools OpinionFinder and Google-Profile of Mood States to measure public reactions around the 2008 election. The researchers then compared the results of this analysis to changes in the value of the Dow Jones Industrial Average (DJIA). The researchers found that simply comparing the sentiment results to trends in the DJIA did not seem to reveal much correlation. However, they then trained a Fuzzy Neural Network with their results organized in a time series along with historical DJIA values. They claimed the results of the neural network showed a higher degree of accuracy compared to their initial attempts (Bollen, Mao, and Zeng 2011).

Methods

The healthcare sector is incredibly dynamic and is responsible for human progress through the development of cutting-edge technologies. This sector is also distinctive in that it is less sensitive to broader economic fluctuations compared to most industries (Correa 2024). Given these characteristics, the healthcare sector presents a unique opportunity for focused analysis of its leading companies in the market.

The aim of this study is to build a knowledge base that is used for market sentiment analysis within the healthcare sector. The intended user of this tool is likely a financial

professional such as a fund manager or financial analyst. It is not limited to only professionals, but anyone looking to invest, such as retail investors. Successful implementation of the knowledge base relies on prerequisite steps that include the identification of appropriate web sources, the development of a web crawler using the Python programming language to perform web scraping with the scrapy package, and the development of a database schema where the information will be housed and built with the graph-relational database Gel.

The knowledge domain must encapsulate the healthcare sector for effective sentiment analysis. The healthcare sector is vast, not only in the range of company sizes, but also in the diversity of its subsectors. For this reason, the top ten healthcare companies by market capitalization are the focus of the crawler, as they are the primary drivers of the market (Kavout 2025). Web sources chosen for the analysis are centered around these ten companies.

To gain insight into company profiles, Wikipedia was chosen for the initial web scrape. The goal is to obtain background information on the ten companies such as name, ticker symbol, market capitalization, location and description. Another source of information was found on Yahoo Finance, specifically within the 'Community' section of each company's page. In this section, users can post comments that often reflect sentiment concerning the company's performance or stock movements. Other sources that provide a similar comment style approach are the social media sites BlueSky and Reddit. The same approach will be applied with those sources where comments posted about the top ten companies will be extracted and added to the knowledge base. Lastly, both Yahoo Finance and CNBC have healthcare sector pages that provide a continuous news feed. The crawler will scrape relevant news articles from these sources as well, with the goal of obtaining opinions and sentiment from journalists and experts.

To ensure the crawler's effectiveness, an interactive exploration of the web would be conducted as outlined by Chakrabarti et al. (1999). This involves manually inspecting sample web pages from each targeted source. The goal is to verify data availability, assess page structure, and identify potential challenges, such as dynamic content in Yahoo Finance comments or potential pay wall issues. This exploration will clarify the design of scraping logic, ensuring the crawler can extract relevant data efficiently. After the initial exploration activities are completed, iterative cycles of user evaluation and feedback will also be conducted for crawler effectiveness.

Chakrabarti et al. (1999) discusses the importance of incorporating a distiller and classifier into a full-blown focused crawler. These tasks will be omitted for this initial stage of the experiment due to limited resources.

The process of defining a schema for the database is necessary for laying a foundation for data storage once text data is obtained through the crawler activities. The type of database used for this study is a graph-relational database built using Gel. This schema combines relational tables (objects) with graph-like relationships (links), making it ideal for modeling our healthcare sector effectively.

Once the data is properly parsed and cleaned it will be fed into the database. Nodes that will be used include company, sentiment source, and financial metric. Properties stored within the company nodes include name, ticker, market cap, and description. This will serve as the central node, connecting to sentiment sources, financial metrics, and sectors for sentiment and financial analysis. Sentiment source nodes will have properties such as source type, content, publish date, sentiment score and url. These nodes will connect companies to capture

sentiment data, assisting with the analysis of public perceptions. Financial metric nodes will have properties such as metric type, metric value and date. These nodes will link to companies to track financial trends over time.

Edges will link the relationships between nodes that will resemble: Company -to- Sentiment Source (Multi-Link), Sentiment Source -to- Company (Multi-Link, Inverse), Company -to- Financial Metrics (Multi-Link), Financial Metrics -to- Company (Single Link, Inverse).

With this initial structure in place, the final application will serve to build up a knowledge base to expose users to the sentiment within the healthcare sector. The intended purpose of this application is for information retrieval within the knowledge base. Additionally, this market sentiment application aids in the decision-making process for trading stocks by offering the user recommendations of buy or sell points in the market. The completion of these methods outlined will ensure a complete foundation for the subsequent analysis, enabling an informed interpretation of the sector's underlying forces and market drivers.

Conclusions

The main issues that may arise when defining the domain and schema are related to scope and relevance. The healthcare sector is vast, so it was decided to stick to the top 10 companies for the scope because it may be a good indicator of the sector but that is not guaranteed. If the market is shifting perhaps those companies are not relevant to that shift and create the urge to expand the scope.

The only special issue that may arise in implementing the knowledge base is the learning curve with Gel. This is a newer database that currently lies outside the expertise of the team.

This study is a good opportunity to learn and develop new skills but the ability to develop quickly and collaboratively within one database system is still not known at this time.

References

Bollen, Johan, Huina Mao, and Xiaojun Zeng. “Twitter Mood Predicts the Stock Market.” *Journal of Computational Science* 2, no. 1 (2011): 1–8. https://doi.org/10.1016/j.jocs.2010.12.007.

Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. 1999, May 17. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks: The International Journal of Computer and Telecommunications Networking,* 31(11-16): 1623–1640.

Cootner, Paul H. 1962. "Stock Prices: Random Vs. Systematic Changes." Industrial Management Review (Pre-1986) 3 (2) (Spring): 24. http://turing.library.northwestern.edu/login?url=https://www.proquest.com/scholarly-journals/stock-prices-ramdom-vs-systematic-changes/docview/214030918/se-2.

Correa, Monica L. 2024. "Health Care Is the Most Negatively Correlated Sector with Economic Cycles." *Seeking Alpha*, May 13, 2024. https://seekingalpha.com/news/4105399-health-care-is-the-most-negatively-correlated-sector-with-economic-cycles-bofa.

Deng, Shuyuan, Atish P Sinha, and Huimin Zhao. “Adapting Sentiment Lexicons to Domain-Specific Social Media Texts.” Decision Support Systems 94 (2017): 65–76. https://doi.org/10.1016/j.dss.2016.11.001.

Feuerriegel, Stefan, and Julius Gordon. “Long-Term Stock Index Forecasting Based on Text Mining of Regulatory Disclosures.” *Decision Support Systems* 112 (2018): 88–97. https://doi.org/10.1016/j.dss.2018.06.008.

Kavout. 2025. "Healthcare Stocks on the Rise: Top Picks for Smart Investors in 2025." *Sector Insights*. March 7, 2025. https://www.kavout.com/market-lens/healthcare-stocks-on-the-rise-top-picks-for-smart-investors-in-2025.

Li, Yung-Ming, and Tsung-Ying Li. “Deriving Market Intelligence from Microblogs.” Decision Support Systems 55, no. 1 (2013): 206–17. https://doi.org/10.1016/j.dss.2013.01.023.

Schumaker, Robert P, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. “Evaluating Sentiment in Financial News Articles.” Decision Support Systems 53, no. 3 (2012): 458–64. https://doi.org/10.1016/j.dss.2012.03.001.

Wu, Jheng-Long, Min-Tzu Huang, Chi-Sheng Yang, and Kai-Hsuan Liu. “Sentiment Analysis of Stock Markets Using a Novel Dimensional Valence–Arousal Approach.” Soft Computing (Berlin, Germany) 25, no. 6 (2021): 4433–50. https://doi.org/10.1007/s00500-020-05454-x.

wüthrich, B, D Permunetilleke, Steven Leung, W Lam, Vincent Cho, and J Zhang. “Daily Prediction of Major Stock Indices from Textual WWW Data.” Transactions (Hong Kong Institution of Engineers) 5, no. 3 (1998): 151–56. https://doi.org/10.1080/1023697X.1998.10667783.

Zweig, Martin E. “AN INVESTOR EXPECTATIONS STOCK PRICE PREDICTIVE MODEL USING CLOSED-END FUND PREMIUMS.” The Journal of Finance (New York) 28, no. 1 (1973): 67–78. https://doi.org/10.1111/j.1540-6261.1973.tb01346.x