CS 410: Text Information Systems
University of Illinois at Urbana-Champaign
Fall 2022

Final Project Submission

Submitted by: Samuel Kirby
slkirby2@illinois.edu

November 26, 2022

Project Track: Free Topics

Project Title: Use of sentiment analysis on mutual fund quarterly commentaries

Team Members: Sam Kirby (slkirby2) – Solo project

Note: italicized text refers to elements of project submission rubrik.

Project Description:

(1) An overview of the function of the code (i.e., what it does and what it can be used for)

There are thousands of mutual funds available to investors, managed by hundreds of asset management firms. Each quarter, many (if not most) of these managers produce quarterly commentary documents, most of which follow a consistent format across the industry, providing information about how each of their funds performed, what contributed or detracted from performance, and their future outlook.

The analysis of these documents represents a tremendous opportunity to gauge, or “crowdsource,” the overall degree of market sentiment amongst a large, diversified group of professional investors. This is also task that would be impossible to perform manually, making it an ideal task for machine learning and sentiment analysis techniques.

This project represents a proof-of-concept that shows promise in using sentiment analysis to aggregate sentiment across hundreds of asset managers. There are several use-cases for this approach, including:

- Quickly gauge prevailing sentiment among a large set of professional investors across asset classes through a “wisdom of the crowd” approach
- Predictive modelling of how individual managers could perform given a range of market scenarios

- Time-series analysis of sentiment scores for individual managers. If a manager tends to be highly optimistic in most periods but shows a negative sentiment, this may represent more of a signal than a manager who tends to be more neutral.

Implementation Details:

(2) Documentation of how the software is implemented with sufficient detail so that others can have a basic understanding of your code for future extension or any further improvement

This project explored the use of sentiment analysis to assess the individual and aggregate market outlook among a large cross-section of asset managers, and then compare their sentiment with actual market conditions and fund results. I used R within the RStudio environment for this project, and the RMarkdown file with embedded R code and documentation, as well as a summary PDF generated from this file are included with this submission. Please refer this the '*CS410_FinalProject_Fall2022_slkirby2_RMarkdown.pdf*' document for full implementation and documentation.

My approach was to:

1. Construct a simplistic crawler to locate and download a suitable sample of quarterly commentary PDF files for the quarter ending 2022-09-30. Results obtained by the crawler were augmented manually for firms whose commentary files are difficult to locate via Google search results. Finally, I manually screened out non-relevant documents, such as annual or semi-annual reports.

2. I employed the 'pdftools' library within R to extract text from the PDFs. The resulting text was highly unstructured, but I used straightforward techniques to narrow-in on the text related for forward outlook.

3. I performed sentiment analysis on the documents using the 'SentimentAnalysis' library to determine a binary sentiment (positive/negative), as well as a continuous sentiment score. Multiple dictionaries were evaluated.

4. For performance evaluation, I obtained the prospectus benchmark for each fund using the 'rvest' library and Ycharts web site. I also obtained performance information for funds and benchmarks for the period 2022-10-03 to 2022-11-23 using the Yahoo! Finance API.

5. The final step was an analysis of model performance. For this step, I evaluated model performance in three ways:

    a. Binary sentiment assessment
    Did the binary (positive or negative) sentiment of the fund manager match the actual binary (positive or negative) benchmark results over the following eight weeks? This step was performed using a Confusion Matrix and accompanying Sensitivity and Specificity calculations.

b. Degree of benchmark-relative out(under)performance
How well did the managers whose outlook was correct perform, relative to those whose outlook was incorrect? This was performed using mean Upside Capture ratio values for the two cohorts.

c. Explanatory power of continuous sentiment scores
How well do the continuous values of sentiment scores explain the degree of benchmark-relative performance? For this step I employed linear regression analysis.

Software and Documentation

(3) Documentation of the usage of the software including either documentation of usages of APIs or detailed instructions on how to install and run a software, whichever is applicable

Please refer to the enclosed ''*CS410_FinalProject_Fall2022_slkirby2_RMarkdown.pdf*' document for code and detailed documentation. The RMarkdown file and a zip archive with source files and intermediary steps are also included for replication of results.

Team Contributions

(4) Brief description of contribution of each team member in case of a multi-person team.

This was a solo project.

Appendix

Additional Discussion: Sentiment Analysis Approach

As discussed in Week 11, the objective of sentiment analysis is to mine and analyze opinion buried within text. Opinions reflect what a person thinks or believes, and depends heavily upon the context and domain of the text being analyzed.

The library used for this project employs dictionary-based sentiment analysis, and includes three base dictionaries:

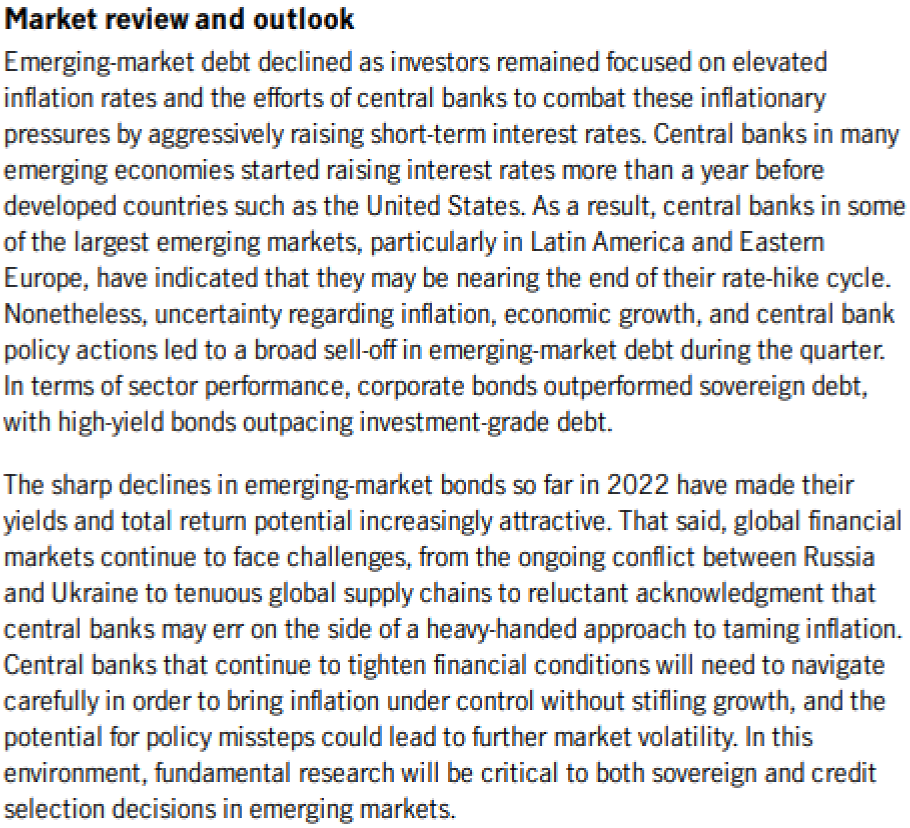
- **DictionaryGI** – a general-purpose dictionary with opinionated words from the psychological Harvard-IV dictionary

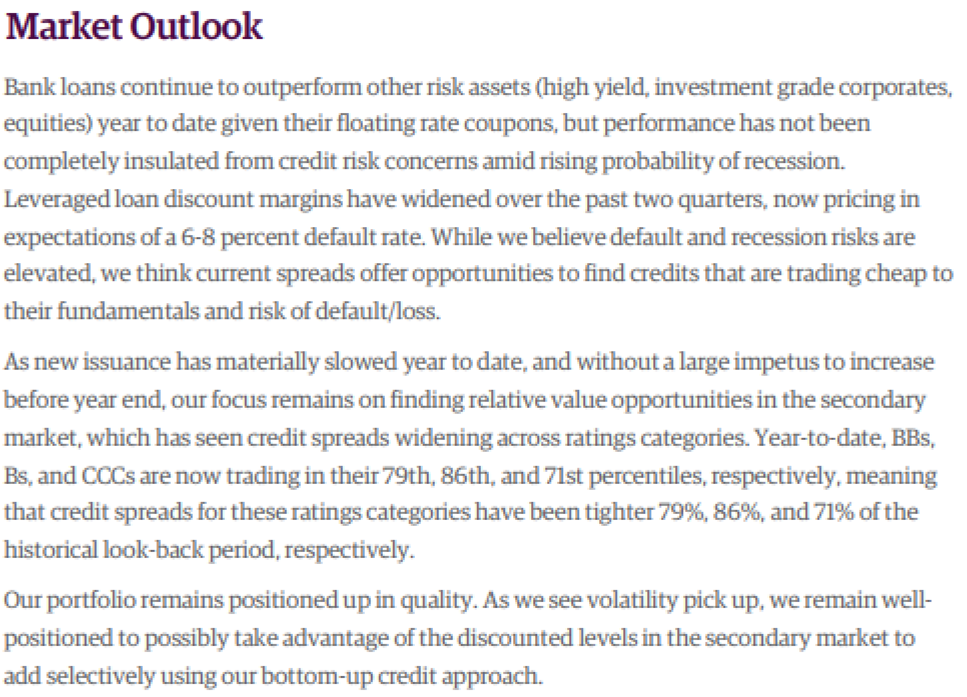
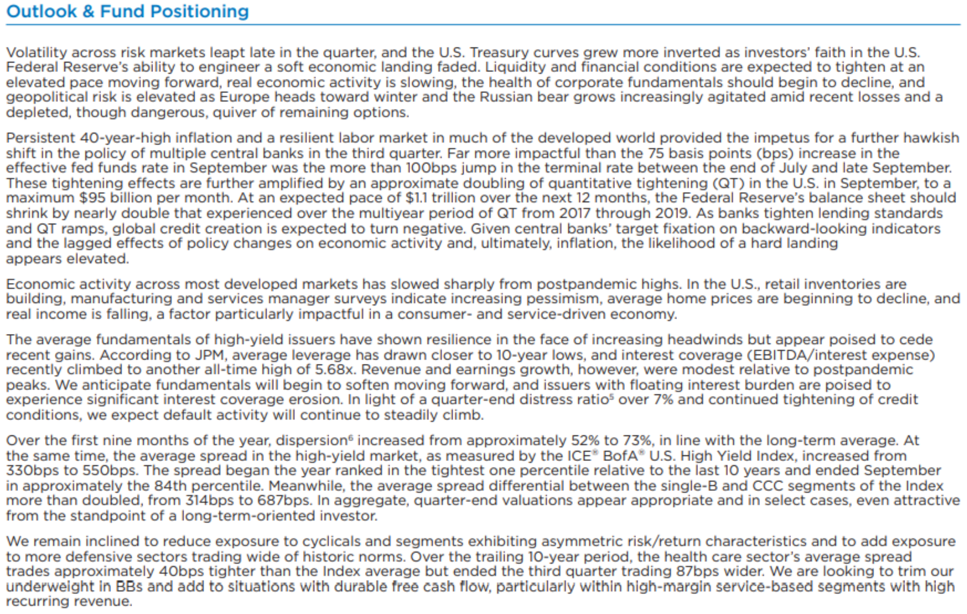
- **DictionaryHE –** a dictionary with a list of positive and negative words according to the Henry's finance-specific dictionary. This dictionary was used in some of the earliest projects on the use of text analysis in the finance discipline

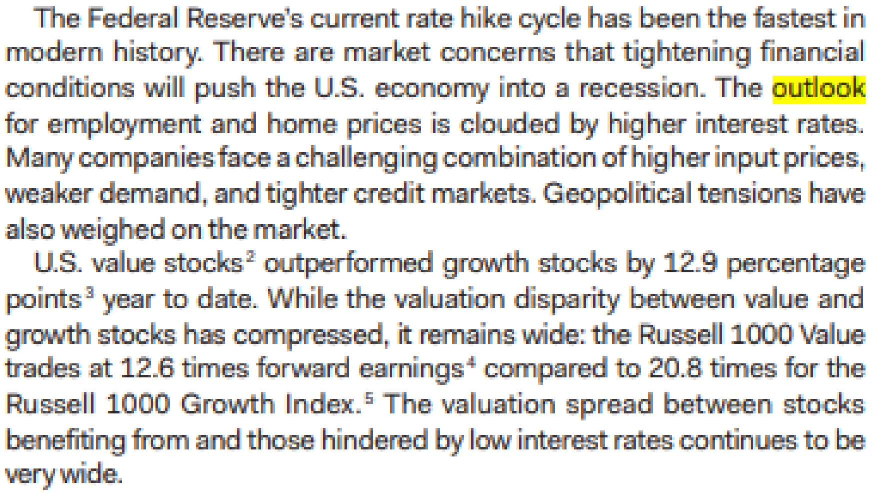
- **DictionaryLM –** a dictionary with a list of positive, negative and uncertainty words according to the Loughran-McDonald finance-specific dictionary and is widely used in the finance domain.

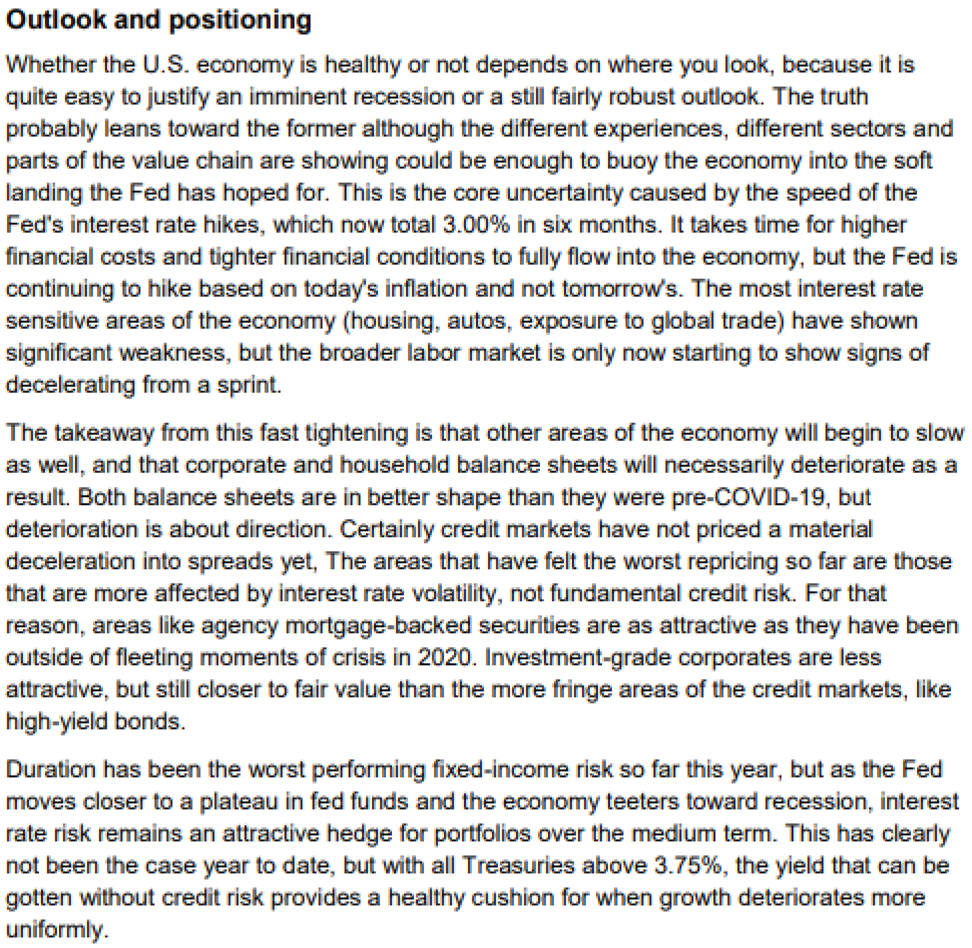
For this project, I used a combination of the balance of positive/negative scores, as well as non-scientific spot-checking to determine the best dictionary. Given a larger sample size and evaluation time horizon, a quantitative approach would be more appropriate.

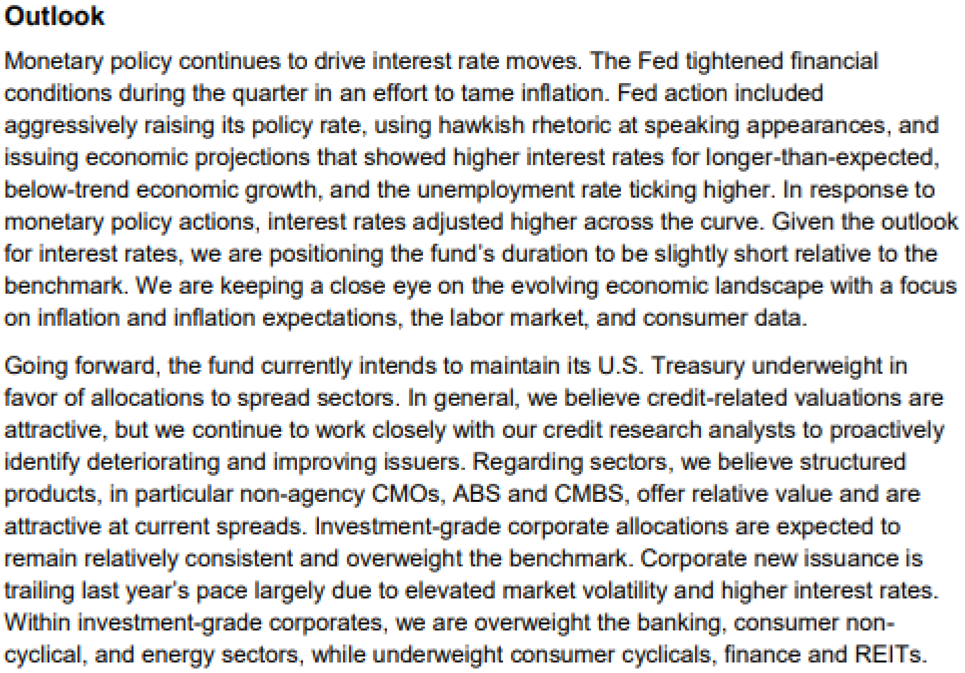
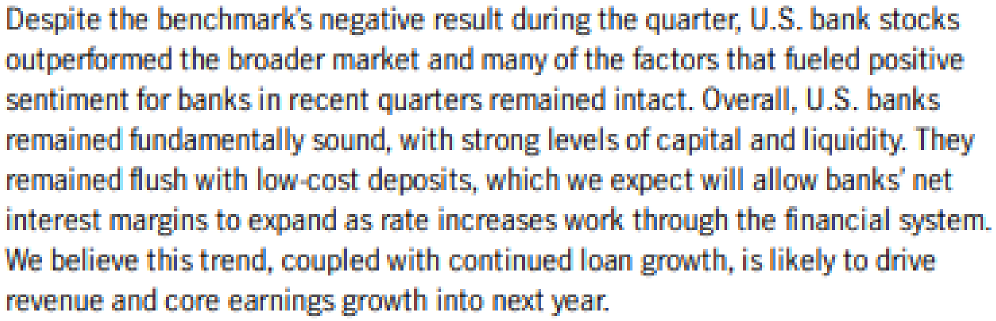
I initially eliminated the GI dictionary given its 100% positive sentiment score, as well as the general-purpose nature of the dictionary. Below are some examples from the LM and HE dictionary which led me to determine that the HE dictionary was most effective for this project.

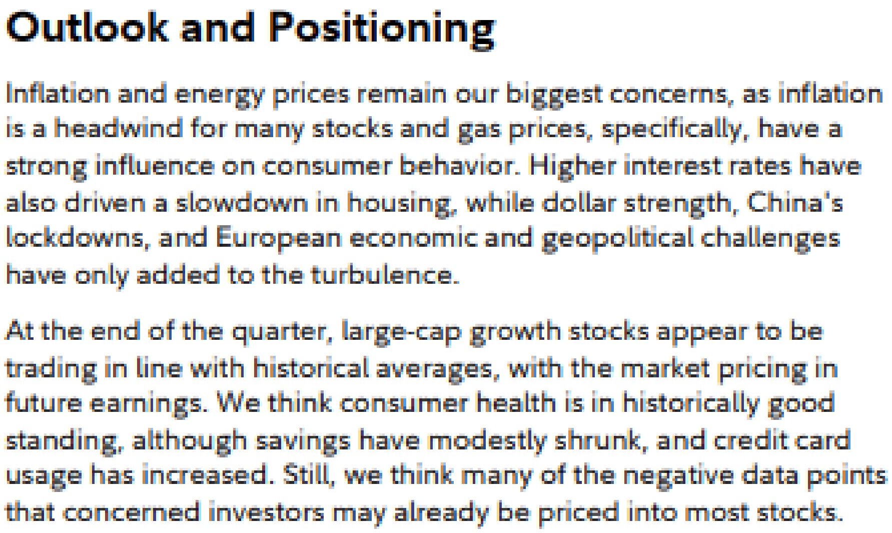
Examples: Loughran-McDonald Dictionary (LM)	
LM Dictionary – worst score JMKAX -0.036 This text does seem to reflect a negative, or at least cautious outlook.	 **Market review and outlook** Emerging-market debt declined as investors remained focused on elevated inflation rates and the efforts of central banks to combat these inflationary pressures by aggressively raising short-term interest rates. Central banks in many emerging economies started raising interest rates more than a year before developed countries such as the United States. As a result, central banks in some of the largest emerging markets, particularly in Latin America and Eastern Europe, have indicated that they may be nearing the end of their rate-hike cycle. Nonetheless, uncertainty regarding inflation, economic growth, and central bank policy actions led to a broad sell-off in emerging-market debt during the quarter. In terms of sector performance, corporate bonds outperformed sovereign debt, with high-yield bonds outpacing investment-grade debt. The sharp declines in emerging-market bonds so far in 2022 have made their yields and total return potential increasingly attractive. That said, global financial markets continue to face challenges, from the ongoing conflict between Russia and Ukraine to tenuous global supply chains to reluctant acknowledgment that central banks may err on the side of a heavy-handed approach to taming inflation. Central banks that continue to tighten financial conditions will need to navigate carefully in order to bring inflation under control without stifling growth, and the potential for policy missteps could lead to further market volatility. In this environment, fundamental research will be critical to both sovereign and credit selection decisions in emerging markets.

LM Dictionary – second worst score GIFIX -0.035 This score also appears accurate; the outlook is indeed cautious.	 **Market Outlook** Bank loans continue to outperform other risk assets (high yield, investment grade corporates, equities) year to date given their floating rate coupons, but performance has not been completely insulated from credit risk concerns amid rising probability of recession. Leveraged loan discount margins have widened over the past two quarters, now pricing in expectations of a 6-8 percent default rate. While we believe default and recession risks are elevated, we think current spreads offer opportunities to find credits that are trading cheap to their fundamentals and risk of default/loss. As new issuance has materially slowed year to date, and without a large impetus to increase before year end, our focus remains on finding relative value opportunities in the secondary market, which has seen credit spreads widening across ratings categories. Year-to-date, BBs, Bs, and CCCs are now trading in their 79th, 86th, and 71st percentiles, respectively, meaning that credit spreads for these ratings categories have been tighter 79%, 86%, and 71% of the historical look-back period, respectively. Our portfolio remains positioned up in quality. As we see volatility pick up, we remain well-positioned to possibly take advantage of the discounted levels in the secondary market to add selectively using our bottom-up credit approach.
LM Dictionary – best score EVIBX 0.042 This appears inaccurate; the text does not convey a positive sentiment.	 **Outlook & Fund Positioning** Volatility across risk markets leapt late in the quarter, and the U.S. Treasury curves grew more inverted as investors' faith in the U.S. Federal Reserve's ability to engineer a soft economic landing faded. Liquidity and financial conditions are expected to tighten at an elevated pace moving forward, real economic activity is slowing, the health of corporate fundamentals should begin to decline, and geopolitical risk is elevated as Europe heads toward winter and the Russian bear grows increasingly agitated amid recent losses and a depleted, though dangerous, quiver of remaining options. Persistent 40-year-high inflation and a resilient labor market in much of the developed world provided the impetus for a further hawkish shift in the policy of multiple central banks in the third quarter. Far more impactful than the 75 basis points (bps) increase in the effective fed funds rate in September was the more than 100bps jump in the terminal rate between the end of July and late September. These tightening effects are further amplified by an approximate doubling of quantitative tightening (QT) in the U.S. in September, to a maximum $95 billion per month. At an expected pace of $1.1 trillion over the next 12 months, the Federal Reserve's balance sheet should shrink by nearly double that experienced over the multiyear period of QT from 2017 through 2019. As banks tighten lending standards and QT ramps, global credit creation is expected to turn negative. Given central banks' target fixation on backward-looking indicators and the lagged effects of policy changes on economic activity and, ultimately, inflation, the likelihood of a hard landing appears elevated. Economic activity across most developed markets has slowed sharply from postpandemic highs. In the U.S., retail inventories are building, manufacturing and services manager surveys indicate increasing pessimism, average home prices are beginning to decline, and real income is falling, a factor particularly impactful in a consumer- and service-driven economy. The average fundamentals of high-yield issuers have shown resilience in the face of increasing headwinds but appear poised to cede recent gains. According to JPM, average leverage has drawn closer to 10-year lows, and interest coverage (EBITDA/interest expense) recently climbed to another all-time high of 5.68x. Revenue and earnings growth, however, were modest relative to postpandemic peaks. We anticipate fundamentals will begin to soften moving forward, and issuers with floating interest burden are poised to experience significant interest coverage erosion. In light of a quarter-end distress ratio[5] over 7% and continued tightening of credit conditions, we expect default activity will continue to steadily climb. Over the first nine months of the year, dispersion[6] increased from approximately 52% to 73%, in line with the long-term average. At the same time, the average spread in the high-yield market, as measured by the ICE® BofA® U.S. High Yield Index, increased from 330bps to 550bps. The spread began the year ranked in the tightest one percentile relative to the last 10 years and ended September in approximately the 84th percentile. Meanwhile, the average spread differential between the single-B and CCC segments of the Index more than doubled, from 314bps to 687bps. In aggregate, quarter-end valuations appear appropriate and in select cases, even attractive from the standpoint of a long-term-oriented investor. We remain inclined to reduce exposure to cyclicals and segments exhibiting asymmetric risk/return characteristics and to add exposure to more defensive sectors trading wide of historic norms. Over the trailing 10-year period, the health care sector's average spread trades approximately 40bps tighter than the Index average but ended the third quarter trading 87bps wider. We are looking to trim our underweight in BBs and add to situations with durable free cash flow, particularly within high-margin service-based segments with high recurring revenue.

LM Dictionary - second best DODGX 0.032 This also does not appear accurate. The text seems negative.	 The Federal Reserve's current rate hike cycle has been the fastest in modern history. There are market concerns that tightening financial conditions will push the U.S. economy into a recession. The outlook for employment and home prices is clouded by higher interest rates. Many companies face a challenging combination of higher input prices, weaker demand, and tighter credit markets. Geopolitical tensions have also weighed on the market. U.S. value stocks[2] outperformed growth stocks by 12.9 percentage points[3] year to date. While the valuation disparity between value and growth stocks has compressed, it remains wide: the Russell 1000 Value trades at 12.6 times forward earnings[4] compared to 20.8 times for the Russell 1000 Growth Index.[5] The valuation spread between stocks benefiting from and those hindered by low interest rates continues to be very wide.

Examples: Henry's Finance-specific Dictionary (HE)	
HE Dictionary – worst score COSIX -0.022 This indeed seems pretty bearish.	 **Outlook and positioning** Whether the U.S. economy is healthy or not depends on where you look, because it is quite easy to justify an imminent recession or a still fairly robust outlook. The truth probably leans toward the former although the different experiences, different sectors and parts of the value chain are showing could be enough to buoy the economy into the soft landing the Fed has hoped for. This is the core uncertainty caused by the speed of the Fed's interest rate hikes, which now total 3.00% in six months. It takes time for higher financial costs and tighter financial conditions to fully flow into the economy, but the Fed is continuing to hike based on today's inflation and not tomorrow's. The most interest rate sensitive areas of the economy (housing, autos, exposure to global trade) have shown significant weakness, but the broader labor market is only now starting to show signs of decelerating from a sprint. The takeaway from this fast tightening is that other areas of the economy will begin to slow as well, and that corporate and household balance sheets will necessarily deteriorate as a result. Both balance sheets are in better shape than they were pre-COVID-19, but deterioration is about direction. Certainly credit markets have not priced a material deceleration into spreads yet, The areas that have felt the worst repricing so far are those that are more affected by interest rate volatility, not fundamental credit risk. For that reason, areas like agency mortgage-backed securities are as attractive as they have been outside of fleeting moments of crisis in 2020. Investment-grade corporates are less attractive, but still closer to fair value than the more fringe areas of the credit markets, like high-yield bonds. Duration has been the worst performing fixed-income risk so far this year, but as the Fed moves closer to a plateau in fed funds and the economy teeters toward recession, interest rate risk remains an attractive hedge for portfolios over the medium term. This has clearly not been the case year to date, but with all Treasuries above 3.75%, the yield that can be gotten without credit risk provides a healthy cushion for when growth deteriorates more uniformly.

HE Dictionary - second worst CUSOX 0.0162 This appears somewhat neutral to somewhat negative (but contains a lot of asset class-specific jargon)	 **Outlook** Monetary policy continues to drive interest rate moves. The Fed tightened financial conditions during the quarter in an effort to tame inflation. Fed action included aggressively raising its policy rate, using hawkish rhetoric at speaking appearances, and issuing economic projections that showed higher interest rates for longer-than-expected, below-trend economic growth, and the unemployment rate ticking higher. In response to monetary policy actions, interest rates adjusted higher across the curve. Given the outlook for interest rates, we are positioning the fund's duration to be slightly short relative to the benchmark. We are keeping a close eye on the evolving economic landscape with a focus on inflation and inflation expectations, the labor market, and consumer data. Going forward, the fund currently intends to maintain its U.S. Treasury underweight in favor of allocations to spread sectors. In general, we believe credit-related valuations are attractive, but we continue to work closely with our credit research analysts to proactively identify deteriorating and improving issuers. Regarding sectors, we believe structured products, in particular non-agency CMOs, ABS and CMBS, offer relative value and are attractive at current spreads. Investment-grade corporate allocations are expected to remain relatively consistent and overweight the benchmark. Corporate new issuance is trailing last year's pace largely due to elevated market volatility and higher interest rates. Within investment-grade corporates, we are overweight the banking, consumer non-cyclical, and energy sectors, while underweight consumer cyclicals, finance and REITs.
HE Dictionary - best FRBAX 0.044 Although this text isn't particularly forward-looking (aside from the last sentence) the language and conclusion are indeed bullish.	 Despite the benchmark's negative result during the quarter, U.S. bank stocks outperformed the broader market and many of the factors that fueled positive sentiment for banks in recent quarters remained intact. Overall, U.S. banks remained fundamentally sound, with strong levels of capital and liquidity. They remained flush with low-cost deposits, which we expect will allow banks' net interest margins to expand as rate increases work through the financial system. We believe this trend, coupled with continued loan growth, is likely to drive revenue and core earnings growth into next year.

HE Dictionary - second best FBGRX 0.0355 This indeed seems moderately bullish.	 **Outlook and Positioning** Inflation and energy prices remain our biggest concerns, as inflation is a headwind for many stocks and gas prices, specifically, have a strong influence on consumer behavior. Higher interest rates have also driven a slowdown in housing, while dollar strength, China's lockdowns, and European economic and geopolitical challenges have only added to the turbulence. At the end of the quarter, large-cap growth stocks appear to be trading in line with historical averages, with the market pricing in future earnings. We think consumer health is in historically good standing, although savings have modestly shrunk, and credit card usage has increased. Still, we think many of the negative data points that concerned investors may already be priced into most stocks.

Further documentation for the *'SentimentAnalysis'* library can be found here: https://cran.r-project.org/web/packages/SentimentAnalysis/SentimentAnalysis.pdf

Ideas for Future Refinement/Extension of Project

There are many paths for improvement of this project, including:

- Expanding the sample-size of asset managers and mutual funds included in the analysis
- Extending the analysis from a single quarter, to a much longer period of time representing a variety of different market conditions
- This proof of concept does not reflect the lowest-cost share class of each fund, which means that fee differences can influence relative performance. Ideally, only the lowest-cost share class of each fund would be used
- Further isolation of the text specific to forward outlook
- Evaluation of different character n-gram tokenization approaches
- Creation of a customized library. The 'SentimentAnalysis' library includes capabilities for the creation of customized libraries using LASSO regularization to select relevant terms. Given the specialized lexicon used within these documents, I expect a custom dictionary would improve model performance considerably.

Software Usage Demonstration

Please refer to demonstration video located within my github repo: https://github.com/skirby1201/CourseProject

Acknowledgements

The following resources proved valuable for the completion of this project:

- The ‘SentimentAnalysis’ R library authored by Nicolas Proellochs and Stefan Feuerriegel. Project URL https://github.com/sfeuerriegel/SentimentAnalysis
- Google
- Yahoo Finance! API
- Ycharts.com (public web site and free trial)