MSIN0032

Management Science Dissertation

How can BlackRock leverage automated news analysis about securities to simplify short-term investment opportunity research for analysts?

Matic Potocnik

2022

Word Count: 11991 of 15002 words

Executive Summary

This report focuses on improving the workflow of security analysts by providing a clear overview of BlackRock's operations, goals, and explores different types of security analysis which can be grouped into 3 categories: risk analysis, fundamental analysis, and news analysis.

It than narrows down the scope and focuses on the most underdeveloped part of security analysis process which is financial news analysis. It then breaks down the current universe of applications used in financial news analysis and finds the most impactful ones to be sentiment analysis and topic modeling.

Next the report further breaks down each of those categories presenting different models and finding FinBERT to be best performing for sentiment analysis, LDA most applicable to topic modeling, and CNN-LSTM most accurate when it comes to market return predictions.

The report then presents an article recommendation solution based on predictions from those models. It specifically considers the relationships between the predictions and carefully determines the recommendation score.

Furthermore, it presents the development process of the solution, its limitations, and explains the exact functionalities of the application.

It finishes by considering the costs and viability of integration and analyzes the impact the solution could have on the security analysts' workflow.

The report concludes that the implementation of state-of-the-art news analyzing techniques is to be explored further and presents a great opportunity for BlackRock to get ahead of the industry.

Contents

Executive Summary 2
Glossary 4
Abbreviation Explanations 4
Developed Solution 4
BlackRock 5
I. Operations 5
II. Target Group 7
III. Breakdown of used applications 12
Problem Statement 17
Solution Research 18
I. Academic grounds 18
II. Available Technology 23
III. Available Data 29
Solution Development 34
I. Solution Overview 34
II. Solution Justification 35
III. Detailed Breakdown of Solution 36
IV. Functionality Showcase 45
V. Developed Solution Limitations 48
Solution Integration and Impact 49
I. Integration 49
II. Impact 50
Conclusion 51
Appendix 52
I. Computational Requirements: 52
II. Code Used for Solution Development 52
Bibliography 53
I. Interview 53
II. References 54

Glossary

Abbreviation Explanations

HFT = High-Frequency Trading

LDA = Latent Dirichlet Allocation

ETF = Exchange Traded Funds

ESG = Environmental Social Governance

VADER = Valance Aware Dictionary and sEntiment Reasoner (sentiment lexicon)

BERT = Bidirectional Encoder Representations from Transformers (Google's NLP model)

ML = Machine learning

Developed Solution

Part of this dissertation was developing the proposed solution. The solution's dashboard is published on free Heroku webserver dynos. Because it's published on free dynos the app goes to sleep after 30 min of inactivity, this means when starting the app, it can take from 30-60 seconds to start. Another limitation of the free service is memory, meaning not many instances of the app can be opened at once, in this case if the app returns an error, please contact me to restart the server.

The application is used as a supplementary to the report and can be accessed via:
https://msin0032-matic-potocnik.herokuapp.com/

BlackRock

BlackRock is the world's largest asset management firm with more than $10 trillion assets under management.[1] It has more than 18,000 employees and operates in 36 countries[2].

I. Operations

BlackRock divides its operations into 5 revenue segments[3]:

1. Investment advisory, administration fees and securities lending revenue

This is BlackRock's main revenue stream representing 78% of their revenue. The main operational part of this segment is investment advisory which can be further broken down by types of services and types of clients.

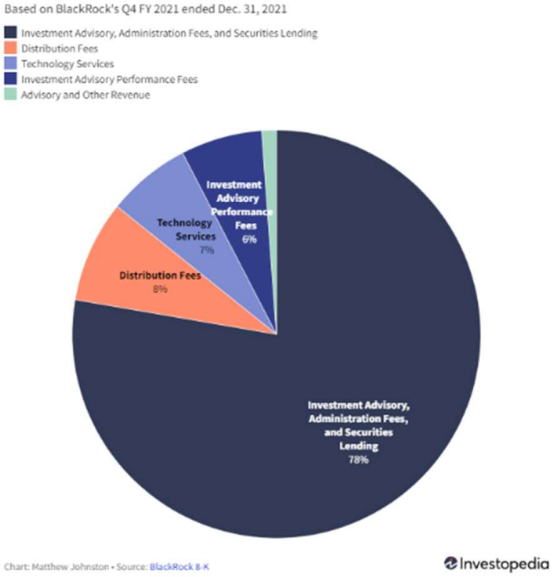


Figure 1 Breakdown of BlackRock revenue streams

Services Breakdown:

Investment Funds, BlackRock provides 3550 different investment funds, 2349 of them being active funds and 1201 being index funds.[4] 565 of those funds are ETF's provided by iShares. The major success of BlackRock can be vastly contributed to the structure of those funds and BlackRock's ability to analyze, create, and manage them.

Financial Market Advisory, on top of providing different funds to investors, BlackRock also provides an array of tailored financial advisory services: Portfolio Construction, Climate Risk Advisory, Transaction Support, Financial Modeling, Risk & Regulatory Advisory.[5]

Client Breakdown:

Retail investors, representing around 11.2% of BlackRock's assets under management, are composed of individuals who invest in BlackRock's investment funds and portfolios to increase the value of their savings.[6]

Institutional Investors, representing 53.5% of BlackRock's assets under management, are composed of organizations that put their assets in BlackRock's management. Those are often: pension funds, investment banks, hedge-funds, mutual funds, credit unions, insurance companies, etc.[7]

[1] BlackRock Surges past $10tn in AUM
[2] BlackRock – About us
[3] BlackRock Form 8-K
[4] BlackRock – Investment Funds
[5] BlackRock FMA – Our Services
[6] BlackRock – About us
[7] Institutional Investor

ETF's, representing 35.3% or BlackRock's assets under management, composed of mainly iShares – a subsidiary of BlackRock which offers exchange-traded funds (ETF's)[8] helping investors an easy low-cost entry into different markets.

2. Distribution Fees

Representing 8% of their revenue and includes fees BlackRock collects by distributing their products.

3. Technology Services

Representing 7% of BlackRock's revenue, its main component is BlackRock's portfolio management system Aladdin.

Aladdin is an all-round investment management platform with key features such as: risk analytics, regulatory compliance analytics, and trading, which enables over 200 institutions composed of banks, asset management firms, insurance companies, pension funds, to make informed decisions about trades, asset management and risk diversification.[9]

This is BlackRock's main platform used by internal analysts to derive insights and evaluate investment opportunities but does not provide any news analytics or coverage about securities and funds.

4. Investment Advisory Performance Fees

Representing 6% of their revenue and includes performance fees investors pay on some funds and financial advisory.[10]

5. Advisory and Other revenue

Other revenue streams representing <1% BlackRock's revenue.

[8] iShares
[9] Aladdin FAQs
[10] BlackRock Fund Charges

II. Target Group

The group of people this dissertation will focus on are financial analysts at BlackRock.

When talking about financial analysts and what they do its important to understand under which BlackRock's operations category they fall to narrow the scope and impact. The focus of this dissertation is to simplify the workflow of security analysts meaning it falls under BlackRock's most valuable revenue stream *Investment advisory, administration fees and securities lending revenue*, specifically under the fund analysis subcategory.

1. Goals of the group

Goal 1:

The main goal of security analysts is to find securities such as stocks, bonds, options, etc. with return and risk characteristics that suit the fund.

At the highest-level the characteristics and goals all funds share is that they're trying to provide highest returns with the least amount of risk. The Modern Portfolio Theory[11] says there are two types of risk:

Market Risk, risk that the entire market faces and is based on business cycles, interest rates, state of the economy, etc. – this risk is undiversifiable and is present in every risky portfolio/fund.

Specific risk, risk that a specific security faces – this risk is unique to each security and can be eliminated/reduced by diversification of portfolios/funds.

This process of diversification of portfolios requires the analysts to understand the risk and return profiles of the specific securities to determine their optimal weight in the final fund.

Take for example MADVX Equity Dividend fund made by BlackRock.

Holdings

Top

as of Apr 29, 2022

Name	Weight (%)	Name	Weight (%)
ANTHEM INC	3.19	ENTERPRISE PRODUCTS PARTNERS L.P.	2.33
WELLS FARGO & COMPANY	2.96	ASTRAZENECA PLC	2.28
CITIGROUP INC	2.46	CISCO SYSTEMS INC	2.28
AMERICAN INTERNATIONAL GROUP INC	2.37	BP PLC	2.23
SANOFI SA	2.34	HUMANA INC	2.01

Figure 2 MADVX Equity Dividend Fund, equity weights

The fund is diversified to the point where the most contribution a single company's stock has on the fund is 3.19%.

[11] Learned in Finance II.

Goal 2:

The second goal of security analysts is to understand what the changes in the market mean for the risk/reward profile of specific securities, how the updated profile impacts the entire portfolio, and how should the weighting of the particular security be adjusted.

However, only understanding how to respond isn't enough, as the prices of securities adjust quickly after major news (such as earnings reports) come out, meaning analysts must constantly monitor the security space evaluate the impact of all news and price changes that happen, and respond to them before the market to generate abnormal returns.

Goal 3:

In the time where HFT strategies and quantitative funds have algorithms in place which respond to notable changes almost instantaneously adjusting the security price minimize the analyst's potential for gains. So, analysts utilize another approach to generating above market returns which is to extract additional information from less obvious sources.[12] This usually generates returns for smaller neglected companies which aren't analyzed by a lot of investors or for specific hard to analyze and access data sources.[13]

However, finding such "undiscovered" information about securities can often feel like wandering in the dark and its discovery more like a lucky event rather than result of structured analysis. Therefore, another goal analysts strive for is discovering these types of information and framing it in a replicable manner.

Summary of goals

The three main goals security analysts have are: construction of a comprehensive security profile used in the diversification of the funds, the development of a framework letting them respond to changes before the market to generate above market returns, and creating a work process that allows them to discover new information consistently.

[12] Video: Interview with Justin Sheetz
[13] Finance II

2. Current Work process

BlackRock approaches understanding risks and returns of the securities with two primary types of strategies which complement each other:[14],[15]

1. **Systematic strategies:**

Teams of portfolio analysts, managers and data scientists take a quantitative approach to finding securities with required risk/reward profile. They examine the data on thousands of securities using computer-driven approaches in search to understand their risk profiles and find the best performing ones.

Analysts use vast amounts of different data to create, statistical and ML models which capture the searched aspects of securities. Some of these metrics are widely used in the industry, some of them are specific to BlackRock's understanding of the market and what drives it, and some of them are specific to the analyst or analyst team. Unfortunately, BlackRock does not disclose any of those metrics in detail as they are an industry secret that gives them an edge over the competition. However, we know which general types of data BlackRock uses in their analysis as explained by Justin Sheets, a former BlackRock VP[16]:

Market Data:

Is the data retrieved from stock exchanges and data about the entire market performance:

Price and volume data, which are the trading prices and number of trades for a selected security in a selected interval.
Index data, prices data of different indexes (such as S&P 500)
Technical data, data from various indicators (usually relating to price momentum)
Economic performance data, data on the performance of the economy (GDP, Inflation)
Short data, data on shorted securities
Insider trading data, data on trading activity of politically and business impactful people to monitor insider trading activity

Financial Data:

Is the data about companies:

Company financial statements
Company earnings announcements
Annual reports
Analyst report estimates
Share ownership data
Project evaluations

[14] Video: Interview with Justin Sheetz
[15] BlackRock Active Equities
[16] Video: Interview with Justin Sheetz

Other:

Various other data sources used by analysts:

Risk data, metrics measuring a securities risk profile (P/E ratio, SMB, HML)[17]

ESG Reports and standards, lately especially important data stream as integrating ESG impact is one of BlackRock's biggest current considerations[18]

News, from both financial news (CNBC, Yahoo Finance, Bloomberg) streams for identifying market specifics and traditional news sources (NY Times, Reuters, BBC) for identifying economy wide specifics

Social media, monitoring twitter and reddit activity has become another analyst tool after GameStop and AMC incident[19]

Patent filings data

Specific data:

- Shipping container location data
- Real estate
- Oil tanker location data
- Airplane flight data
- Purchase data (in retail)
- Pharmaceutical testing results data

Other

By analyzing vast amounts of data shown above analysts can capture the security's performance in comparison to other securities and the market. This allows them to develop ranking systems based on security modeled performance which in turn helps them narrow the array of viable securities.

As systematic strategies take a wide range of data and perform market wide security analysis the provided depth of understanding of each security will be shallower than if they were to focus on each security separately and in depth like fundamental strategists.

Therefore, the narrowed subset can then be passed to Fundamental strategists which are able to conduct detailed specific analysis on each of the securities.[20] By narrowing down the universe of securities fundamental analysts should analyze with quantitative methods systematic strategies minimize the amount of research targeted at less impactful securities and therefore minimize the opportunity costs that would arise in otherwise researching security based on only intuition.[21]

[17] Fama-French 3 factor model discussed in Finance II.
[18] BlackRock ESG
[19] AMC Vs. Game Stop
[20] BlackRock Active Equities
[21] Video: Interview with Justin Sheetz

2. Fundamental strategies:

Teams of portfolio managers and analysts focus on a small subset of securities, often ranging from 50-100.[22] They carefully examine the:

Industry: they're in, their position in it, and their risk/return profile compared to the industry.

Company performance: annual reports, financial reports, ESG reports, comparison of company results to security substitutes

News: financial surrounding the security, and general news with potential impact on the security

Company specific data: any highly specific data types unique to the security (e.g. shipping container location for shipping/production companies[23], drug trial results for pharmaceutical companies[24], customer lifetime value for e-commerce[25])

Analysis of the data above helps the analysts to get an in-depth insight into each individual security's prospects.[26]

The main advantage of this approach in contrast to systematic strategies is a deep understanding the analyst gains for each of the 50-100 companies they're researching, and the confidence to report on how the company should perform on the financial markets based on conditions gathered from the current data.[27] However, diving that deep into a few securities presents an inherent disadvantage of scope, as the analyst specializes only in those few securities, this is offset by providing them with a smaller space of viable securities created by performing systematic strategies.

[22] Video: Interview with Justin Sheetz
[23] Modeling Stock Returns and Risk Management in the Shipping Industry
[24] Stock Market Returns and Clinical Trial Results of Investigational Compounds: An Event Study Analysis of Large Biopharmaceutical Companies
[25] Customer-Based Corporate Valuation for Publicly Traded Non-Contractual Firms
[26] BlackRock Active Equities
[27] Video: Interview with Justin Sheetz

III. Breakdown of used applications

Above we explored how the two prevailing analysis strategies are used to generate insights in BlackRock. In this section we will explore the concrete solutions BlackRock and industry analysts use in their work-process and explain which parts of it they improve and how do they help them reached the outlined goals.

1. Aladdin

The first solution we're exploring is BlackRock's proprietary investment management platform Aladdin. The main value creation of Aladdin for analysts is all in one system which includes detailed breakdowns of portfolios/funds, the industry leading risk analysis, and newest market and financials data.

Aladdin is used as the resource analyst resort to when analyzing a new security as it provides a clear breakdown of securities returns over time and risk profile compared to the security's industry sector (e.g. comparing Apple to tech industry, or Pfizer to healthcare), its issuer (e.g. different bonds and options issued by the same entity), country of origin (e.g. comparing a German company to other German companies), industry analytics ratings (e.g. Morningstar ratings), etc.[28] It also provides the analyst a comprehensive security specific risk profile with over 2000 analytics metrics[29], BlackRock derived forecasts on trends and future cashflows, and access to BlackRock's statistical and analytical models.

Furthermore, Aladdin is the primary resource used in evaluating how a security would fit in a portfolio. Here the analyst can use tools to understand the portfolio's exposure and how the addition of a security will cover/expose it, to perform scenario analysis that tests the portfolio in different market conditions and how a security performs under them, a comprehensive value-at-risk measure displaying how exposed a security is and how does that change the value-at-risk of the portfolio, an efficient frontier optimization and analysis which helps the analyst determine the optimal weights in the portfolio, and a display of potential compliance violations.[30],[31]

Finally, Aladdin is used to evaluate the performance of a portfolio with intraday P&L reports providing rapid feedback on the quality of the decisions made in restructuring, the risk and return profile of the portfolio and how it compares to other portfolios, the results your portfolio achieves in widely adopted benchmarks.[32]

Aladdin's comprehensive portfolio analytics help analysts achieve goal 1 in its entirety and its extensive scenario analysis and up to date market and financials data a large part of goal 2. However, the lack of news and company specific data integration forces analysts to resort to other services if they wish to fully achieve goal 2 as news plays an important factor in rapid market changes[33] and achieve goal 3 at all as information edge is usually generated with obscure data sources[34].

[28] Aladdin Risk
[29] Aladdin's Benefits to Risk Managers
[30] Aladdin Risk
[31] Aladdin's Benefits to Portfolio Managers
[32] Aladdin's Benefits to Portfolio Managers
[33] Financial news predicts stock market volatility better than close price
[34] Video: Interview with Justin Sheetz

2. Bloomberg Terminal

A resource used by most financial institutions that provides an extensive security analysis tool lacking in comprehensive portfolio management and risk analysis which makes it a perfect complement to Aladdin.

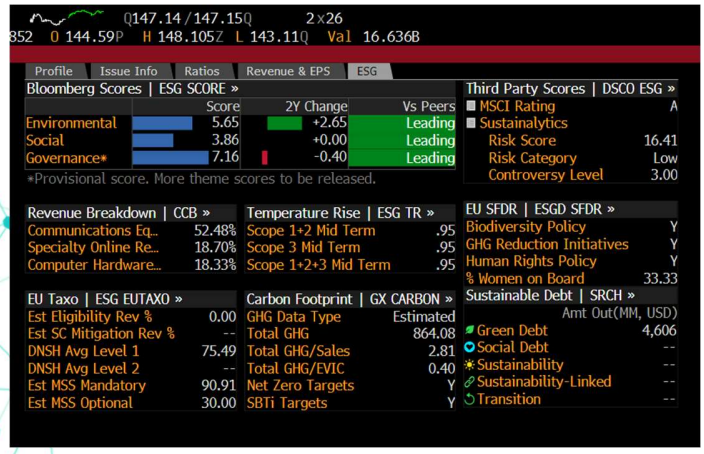
Some of the Bloomberg Terminal features that complement Aladdin are:

Aggregation of company financial filings:

This simplifies the data retrieval for the fundamental analysts but does not help quantitative analysts which still have to preprocess the text data into a usable data format.

ESG reporting:

With the rise of importance of Environmental Social Governance reporting and how companies perform on those fronts. Having a comprehensive breakdown of a companies impact on ESG metrics as well as a simple scoring system and peer comparison simplifies the research for fundamental analysts and the spreadsheet style of reporting an easily integratable solution for quants.



News research:

The news research breakdown in Bloomberg Terminal is constructed of three parts:

Aggregated news rankings:

Bloomberg uses a proprietary ranking algorithm that ranks news based on importance to a specific security[35]. While the specifics of Bloomberg's algorithm are unknown as they only promote the buzzword of being machine learning based the actual ranking system most likely work on the basis of word occurrences (e.g. number of times Apple is mentioned in an article) and number of views.[36],[37]

Recommending the most impactful articles helps the fundamental analysts save time and the Bloomberg API integration of this feature can help quantitative analysts working on models that include news analysis incorporate importance ratings.

News breakdown by themes

Bloomberg uses machine-generated themes and groups the news articles by them. This gives the analysts a great breakdown of news themes surrounding the security but has a few issues:

Articles have 5 topics per article on average[38] which can differentiate even if the articles are assigned into a specific theme

The lack of information about the importance of a theme (e.g. intuitively a theme about mergers or acquisitions will most likely be more impactful than product comparisons), so an analyst must still read through the themes to understand the impact the articles in them could have on the market

[35]Bloomberg - News Headlines Powered by AI.
[36] In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse
[37] Google News Ranking
[38] An algorithm for unsupervised topic discovery from broadcast news stories

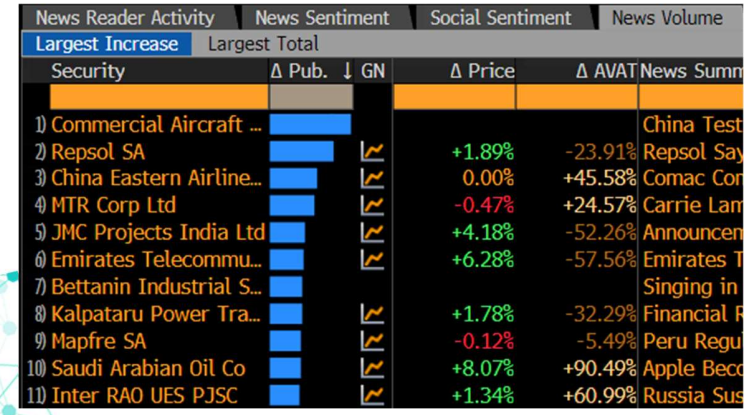
Aggregated news breakdown by sentiment, publication volume, and reader activity

A simple breakdown of news articles based on selected metric:

Sentiment – displays articles with highest positive/negative sentiment analysis score or the aggregated display of article volume and sentiment about security over time.

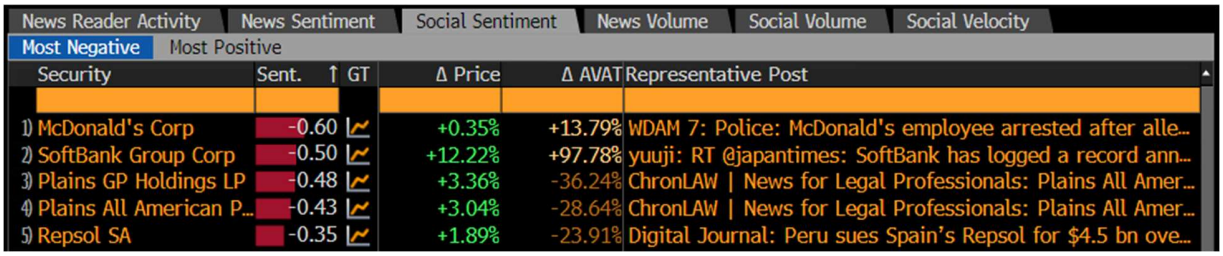
Publication volume – displays the security or security topics with the highest number of articles published or highest increase in publish volume. The correlation between publish volume and price volatility[39], allows analysts exploit the volatility to increase the returns or adjust risk score.

Reader activity – displays securities or security topics based on number of searches and readership metrics. Provides similar opportunity as publication volume.



Social media monitoring

Provides the analyst with the insight into securities with most influential tweets and the aggregated tweet sentiment.



[39] The Impact of Firm-Specific Public News on Intraday Market Dynamics: Evidence from the Turkish Stock Market

3. Natural Language Processing for Financial reports

BlackRock as many other financial institutions uses modern machine-based approaches for analyzing structured company publications such as obligatory annual reports[40] which already follow a predetermined structure in the US (Form 10-K for annual reports and 10-Q for quarterly)[41] and are in the process of standardizing the form in the EU with the European Union implementing the XBRL standard[42].

They also analyze unstructured company publications (such as press releases and letters to shareholders), news articles[43] and because of disruptions r/WallStreetBets caused with retail investors pumping GME and AMC social media sentiment and engagement on Reddit and Twitter.[44]

However, the extend of the use of different technologies isn't publicly available as this analysis process creates BlackRock's advantage over the competition.

4. Inhouse statistical and machine-learning models

As former BlackRock's VP Justin Sheetz mentions in an interview[45] BlackRock uses highly specific statistical models for classes of securities (usually similar securities that respond to similar inputs and can be considered interchangeable), in which the effect of each input on the model result must be carefully studied and understood making the development of models labor and time consuming. Additionally, BlackRock uses machine-learning models which are less rigid than statistical models meaning they can be applied to wider selection of securities and take in data with more Volume, Variety, Velocity, and Veracity.[46],[47]

[40] Video: Interview with Justin Sheetz
[41] SEC Form 10-K Structure
[42] EUROPEAN SINGLE ELECTRONIC FORMAT - XBRL
[43] Video: Interview with Stanley Chen
[44] Video: Interview with Justin Sheetz
[45] Video: Interview with Justin Sheetz
[46] The Actual Difference Between Statistics and Machine Learning
[47] Art and Science of Management module

Problem Statement

As presented above financial analysts strive to perform analysis faster and, in more detail, to gain a competitive advantage over the market.

The content of the security analysis can be structured into 3 parts: risk analysis, fundamental analysis, and news analysis.

BlackRock's industry leading risk and portfolio management system, Aladdin, already gives the analysts the best possible tool the market provides to perform risk analysis.

The structured nature of financial reports and the extensive academic coverage on analyzing them with machine-based models already provides BlackRock with advanced tools to quantify their impact.

However, there is a clear disconnect between the analysis of news and its incorporation into an analyst's workflow, with analysts most commonly still getting their news data from traditional sources such as live financial television channels or online news articles (such as CNBC, Bloomberg, YahooFinance) and from Bloomberg terminal, which as explained already has some functionalities of analyzing the text data but those are often lacking depth.

Therefore, the goal of this report will be to present different news analysis approaches and analyze their potential for implementation. Based on those findings a solution that will help the security analysts save time and perform a more in-depth news analysis by providing them with a recommendation-based system which reports various news article metrics and ranks the impact of the news will be developed.

Solution Research

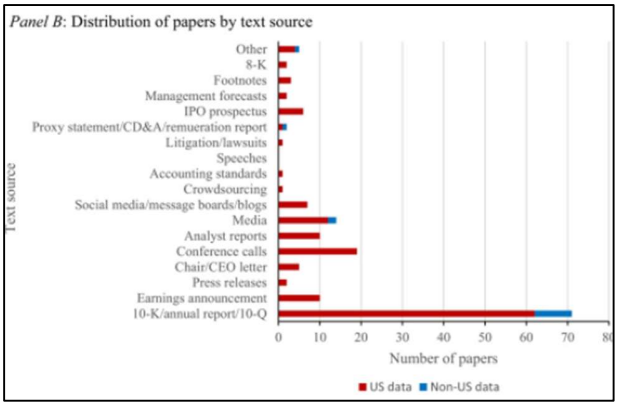
The following section will explore the current space of financial news analysis by presenting the academic research on applications of text analysis in finance, by exploring different technologies currently available for text analysis and how can they be leveraged by BlackRock, and by exploring the availability of the data and justification for using it in the developed solution.

I. Academic grounds

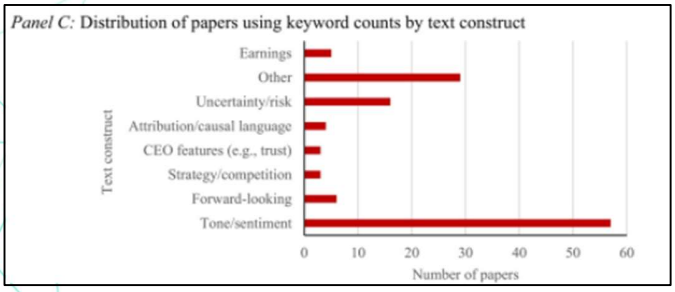
1. Academic justification for solution development

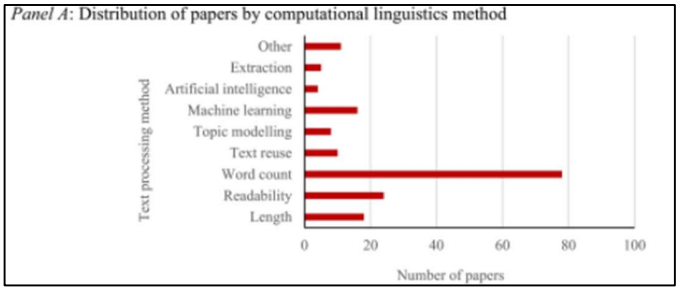
The space of research about uses of natural language processing with relation to financial text is vast. A research article by Fisher et al. (2016) analyzes 266 different articles, paper, book chapters, and dissertations in this space and categorizes them by analysis methods and use cases.

This classification further researched by El-Haj et al. (2019) suggests the space is dominated by analysis of structured annual report (10-K) text as the longer text length and predetermined structure make the comparisons between them easy to analyze.



Additionally, the most prevalent method of analysis is using word count approaches specifically for sentiment analysis.





Because the academia research space already covers the analysis of structured financial reports and BlackRock already utilized their findings this report won't be focusing on developing a solution based on this approach but rather investigate analyzing the unstructured fast moving news media space, which is

found being researched in only 15 papers with most of the papers focusing on understanding the media sentiment as according to Fischer et al. (2016).

Research papers published by Atkins et al. (2018), Alanyali et al. (2013), Emenike et al. (2020), and numerous others show that there is a positive correlation between the amount of news articles published and the securities trading volume (Alanyali et al. 2013) and its price volatility (Atkins et al. 2018).

However, the research suggests there is no correlation between the number of published media and returns (Alanyali et al. 2013). Studies from Tetlock et al. (2008), Heston et al. (2016), Dougal et al. (2012) and several others suggest that daily news can predict stock returns for the next 1 to 2 days.

Additionally, Finance II: Investment Management textbook by Dr. George Namur[48] recognizes that earnings announcement, which are for all intense and purposes a type of news, create an Efficient Market Hypothesis violation as the excess return isn't adjusted instantly but rather shows signs of abnormal returns for future periods as well.

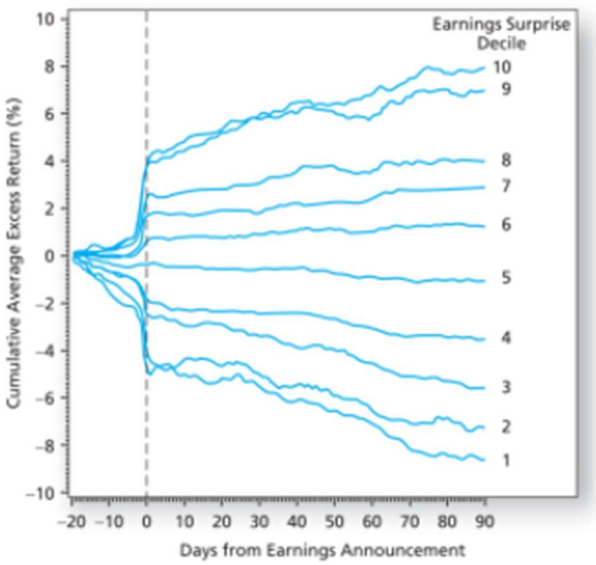


Figure 3 Cumulative abnormal returns in response to earnings announcements

These findings of news generating higher market volatility and having the potential to predict stock returns give grounds to our solution development as analysts should exploit the volatility or return predictions to increase returns and accordingly adjust risk profile of the security.

[48] Finance II: Investments Management - Book

2. Current areas of academic research for text analysis in finance

This section will create a detailed breakdown of analysis methods used for financial text analysis and explore the implications of academical papers published under each category.

According to an article by Klimczak (2021) published in *Innovation In Financial Serives: Balancing Public and Private Interest* (Lech Gąsiorkiewicz, Jan Monkiewicz, 2021) text analysis methods used in finance can be categorized into 4 main groups:

Sentiment Analysis: are methods which analyze the text and assigns it a sentiment value. These are usually categorized into 3 levels: positive, neutral, negative – each of them respectively presenting the tone or opinion in the text.

Dridi et al. (2019) in their article: *FineNews: fine-grained semantic sentiment analysis on financial microblogs and news* show that analyzing the sentiment of news articles can predict stock returns.

McGurk et al. (2019) in their article: *Stock Returns and Investor Sentiment: Textual Analysis and Social Media* also measure that the behavioral finance premise of investor sentiment influencing stock returns holds true with sentiment in news articles and social media results in abnormal stock returns.

Li et al. (2014) in the article: *News impact on stock price return via sentiment analysis*, also find that sentiment analysis improves the prediction of stock returns and that simply looking positive/negative dimension of the article (e.g. considering all positive articles as the same) don't result in useful predictions.

Additionally, numerous other research papers and articles find similar conclusions.

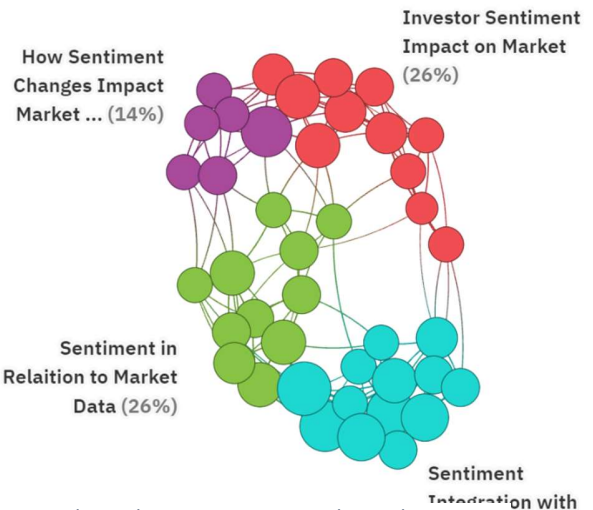


Figure 4 Quid search over sentiment analysis relating to stock return predictions

Topic Modeling: are methods which analyze the text as a bag-of-words and assigns it a specific topic breakdown based on the entire corpus of provided documents, this method helps to categorize texts and find common topics (e.g. acquisitions, mergers, earnings) throughout the documents.

Morimoto et al. (2017) in article: *Forecasting Financial Market Volatility Using a Dynamic Topic Model,* find that analyzing news data with dynamic topic models leads to improved performance of market forecasting.

Xiu et al. (2019) in article: *Predicting Returns with Text Data*, report that using supervised sentiment topic model to analyze news results in significant return predictions.

Nguyen and Shirai in articles: *Topic Modeling based Sentiment Analysis on Social Media for Stock Market Predictions (2015)* and *PhraseRNN for Aspec-based Sentiment Analysis (2015)*, find that using combinations of topic modeling approaches and sentiment based approaches can predict stock market movements.

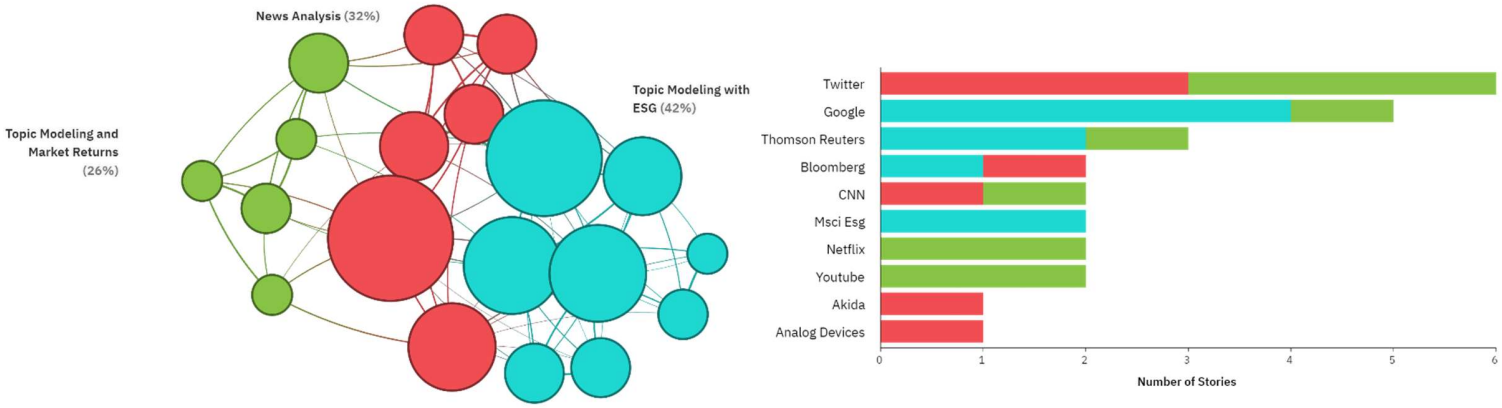


Figure 5 Quid search for topic modeling and market returns

Word Embedding: are methods which take into consideration the context a particular word is in and predict words like it. As financial texts contain plenty of jargon specific words with different meanings in traditional lexicons (e.g. bear and bull) this methods are used to understand the particular meanings of each word which reduce confusion when using the text in different models.

Hogenboom et al. (2021) in article: *The impact of word sense disambiguation on stock price prediction*, find that understanding word meanings with statistical word embeddings improve stock price prediction.

Lopez-Arevalo et al. (2016) in article: *Improving selection of synsets from WordNet for domain-specific word sense disambiguation*, implement a lexicon approach of WordNet to disambiguate financial texts with great success.

Multilingualism: process of translating and conveying meaning from news texts published in different languages. This is especially important for foreign securities (not based in US) as local financial news outlets can publish unique information faster which analysts could use to gain an edge over the market.

There is a general lack of studies surrounding the impact of foreign news on the stock market, but we can conclude with logic that keeping track of foreign news surrounding a foreign security can have an impact on securities in our market (e.g. breaking news surrounding a specific producer only present on the German market can have an effect on companies in our market that were close partners with the producer).

Bannier et al. (2019) in article: *Doing safe by doing good: ESG investing and corporate social responsibility in the U.S. and Europe*, recognize the difficulty of translating the articles from German to English as specific phrases and idioms don't translate well. This is an area that could greatly benefit from implementing the word embedding method mentioned above.

Given those 4 groups the report will focus on implementation of sentiment analysis and topic modeling as they're best researched and were found to predict market returns. However, implementing the methods from the research to only predict market returns does not necessarily help security analysts as one of their main goals is to understand how a particular news article affects the security and the fund. Getting only machine-generated return predictions misses on the intricacies of exactly understanding why and how the piece of media affects market conditions. That is why the dissertation suggests a development of a solution which will recommend influential articles and article metrics, based on semantical topic modeling analysis to analysts who can then take a deeper dive into the article to precisely understand its impact.

As analysts don't have the time to analyze every single article, they must make compromises when choosing which news article to commit their resources to. By implementing a recommendation-based system we can simplify that process of choosing the best articles and provide data-based evidence for why it's the correct decision which reduces opportunity costs created by not choosing different articles.

II. Available Technology

This section will explore the current state of the art approaches to solving the problem outlined above. It will focus primarily on sentiment analysis and topic modeling techniques.

1. Sentiment analysis

Alsaeedi et al. (2019) found there are two main methods when it comes to sentiment analysis: a lexicon-based approach and a modeling-based approach.

The lexicon based or unsupervised approach uses a predetermined lexicon, for example a popular lexicon for sentiment analysis is VADER, which includes the sentiment score for each word. The text, after cleaning, is then analyzed to find out number of times each unique word is recorded in the text and assigns the predetermined sentiment the word has from the lexicon. The overall document sentiment is then determined by summing or averaging the word sentiments from the document.[49] This approach is quick and simple to implement as creating a bag-of-words, looking up the sentiments, and calculating the average isn't a computationally complex task. However, there are a few major problems with this approach. Firstly, the underlying assumption that a word will have the same or very similar sentiment in every context (e.g. increase can be positive or negative) which doesn't necessarily hold true. Secondly, jargon words and phrases, which are very prevalent in finance, are included only in Loughran-McDonald financial sentiment lexicon[50] meaning the model captures the sentiment analysis of specific people which might not be in-line with BlackRock's analysts. Thirdly, the assumption that a bag-of-words approach, which doesn't include for word positioning and specific phrases. Lastly, news important to the market don't consist only of strictly financial news but include news from companies in different industries most of which have different meaning and sentiment behind words (e.g. importance of word trial in tech industry vs healthcare).

The superior approach is the modeling-based as it's a supervised approach to creating a model which trains from input data and makes predictions with certain confidence intervals. This approach is significantly harder and more complex to implement and create predictions with but usually provides more specific and superior results.

[49] Improved lexicon-based sentiment analysis for social media analytics
[50] Loughran-McDonald sentiment lexicon

The models are often based on a Financial Phrasebank from Malo et al. (2014) which include sentiment scores for 4840 financial sentences annotated by 16 people with financial background. As more people analyzed the sentiment the dataset also includes subsets of sentences on which's sentiment all analysts unanimously agreed, 75%, 66%, or 50% of analysts agreed upon.

```
Sales have risen in other export markets .@positive
Sales increased due to growing market rates and increased operations .@positive
The agreement strengthens our long-term partnership with Nokia Siemens Networks .@positive
```

Figure 6 Extract from Financial Phrasebank

Alternatively, models also use more specific annotated datasets (e.g. crypto focused, healthcare focused, etc.)[51] which are usually annotated by fewer analysts making them not as rigorous. Additionally, BlackRock could create a sentiment dataset in-house which would reflect their exact way of grading the sentiment information which would allow them to develop a model which specifically recognizes which news they internally value and which not.

Alsaeedi et al. (2019) explore 7 different simpler supervised machine-learning approaches to model sentiment analysis and Genc (2020) from ProsusAI evaluates 7 more complex models.

The simpler models such as Naïve Bayes, variations of classification trees (random forest, xgboosted trees, adaboosted trees, etc.), support vector machine, multinomial logistic regressions, are simple to implement but their performance suffers. A valid case can be made for their use in trial tests where we explore performances of different datasets as their training time and complexity is low but shouldn't consider them for the final application deployment as not only BlackRock can afford increased complexity and costs in exchange for accuracy but must strive towards it to keep an edge over the competition.

Out of the tested complex models a variation of Google's BERT model which is pre-trained on text from Wikipedia and BookCorpus, uses 340 million parameters, and already accounts for word embeddings.[52]

The variation created by ProsusAI is called FinBERT and is a BERT model finely tuned with Financial Phrasebank from Malo et al. (2014) which allows BERT's complex structure to adjust and evaluate the text sentiment with relation to financial sentiment specialty.

[51] Kaggle dataset search: Sentiment Analysis
[52] Google BERT

Model	All data			Data with 100% agreement		
	Loss	Accuracy	F1 Score	Loss	Accuracy	F1 Score
1. LSTM	0.81	0.71	0.64	0.57	0.81	0.74
2. LSTM with ELMo	0.72	0.75	0.7	0.50	0.84	0.77
3. ULMFit	0.41	0.83	0.79	0.20	0.93	0.91
4. LPS	-	0.71	0.71	-	0.79	0.80
5. HSC	-	0.71	0.76	-	0.83	0.86
6. FinSSLX	-	-	-	-	0.91	0.88
FinBERT	**0.37**	**0.86**	**0.84**	**0.13**	**0.97**	**0.95**

Figure 7 Model performance for sentiment analysis of financial articles

The main advantage of FinBERT is the fact that its already pretrained and can be immediately implemented or can be retrained on a specific dataset to capture a sentiment of the desired aspect (e.g. BlackRock's internal sentiment valuation).

The main disadvantage of FinBERT is its complexity and while it results in industry leading performance predicting the sentiment for each article doesn't take a trivial amount of time anymore so the model wouldn't be useful for implementation with high-frequency trading.

Additionally, to lexicon and model-based approaches a plethora of hybrid and ensemble approaches are available but none of the ones researched by Alsaeedi et al. (2014) outperform FinBERT.

2. Topic Analysis

Similarly, to sentiment analysis topic analysis can be broken down into two overarching categories supervised models and unsupervised models.

Supervised models rely on having an annotated dataset with different topics to utilize models like the ones explained above. In terms of accuracy, they offer a significant advantage over unsupervised models as training as the results are more accurate, specific, and can be measured[53] but have two severe drawbacks.

The first is the lack of available datasets, while sentiment analysis had many tested and good data sources there aren't any for topic modeling. So, to train a supervised topic modeling model BlackRock would first have to construct an exhaustive dataset with many analysts annotating news articles and assigning them topics. This would not only result in being able to construct an accurate topic model but also capture the topics BlackRock analysts value in an article which would help in determining their importance. However, constructing such a dataset wouldn't be possible for every security/industry.

[53] Supervised vs Unsupervised Learning: Algorithms and Examples

The second problem is the lack of adaptability, while sentiment generally stays the same topics present in news vary drastically so even if BlackRock constructed a detailed annotated dataset, it could only capture standard unchangeable topics (e.g. earnings, acquisitions, bankruptcy, etc.) and not topics present in current news. For example, pre-pandemic chip shortage wouldn't even be in consideration as one of the topics in tech and automotive sector but is one of the driving topics today[54], similarly war in Ukraine created many different topics across industries that heavily effect the main message of the article. Constructing a fixed topic dataset would result in the model not being able to capture recent important topics with big impact.

Unsupervised models solve some of those issues. While being less accurate and heavily relying on model analyst to derive meaningful labels for topics unsupervised models don't require a labeled data set which opens them for a far larger array of possible datatypes. Additionally, each new news source can be used to update the model helping it keep up with current topics.

The underlying model behind topic modeling is Latent Dirichlet Allocation which was developed by Blei et al. (2003). LDA is a generative probabilistic model which finds the probability that a specific word belongs to a specific topic and the probability that a specific topic belongs to a specific document[55]. The topics are machine-generated and inferred from the given array of documents.

Like LDA but less used models with the same purpose are Gaussian Mixture Model[56], Formal Concept Analysis[57] (used mainly for social media modeling, Topic Graph Markov Decision Process (TG-MDP)[58], and Neural Topic Models[59].

All the models mentioned, can process unlabeled data and infer its topics which comes with some shortcomings.

Primarily, the legibility of generated topics as they get generated by looking at which words co-occur the most the final topics are represented only by a list of words and probability that the word belongs in the topics. Words between topics often overlap making it harder to define a particular meaning of the topics so, it's up to the analyst to infer the meaning from those words as best as possible which introduces analyst subjectivity and bias.[60],[61]

Secondly, the issue with evaluating the performance of the models. Since there is no labeled data to test the model's predictions on choosing the correct model again requires the analyst to look at the results and pick the model with the most insightful ones. There are metrics such as perplexity which measure how the model reacts to new unseen data and topic coherence which measure the semantic similarity between words with the highest probability.[62] However, none of those metrics provide a concrete insight into the model's performance.

[54] Google news search: Chip Shortage
[55] Marketing Science Lecture User Generated Content
[56] Gaussian mixture models
[57] Formal Concept Analysis
[58] Topic Modelling: A Deep Dive Into LDA, Hybrid-LDA, And Non-LDA Approaches
[59] Neural Topic Models
[60] Topic Modelling: A Deep Dive Into LDA, Hybrid-LDA, And Non-LDA Approaches
[61] Marketing Science module
[62] Evaluate Topic Models: Latent Dirichlet Allocation (LDA)

Finally, the underlying assumption that an article can be represented as a bag-of-words meaning word positioning in the text doesn't matter.[63] This is a big assumption to make as phrases and adjective-noun word combinations impact the meaning. There are some methods of dealing with this issue, most notably including bigrams and trigrams, which are two- and three-word combinations that often co-occur in the articles (e.g. Dow Jones, bullish market). Another method in its research infancy is including the insights provided by word embeddings with LDA primarily with the lda2vec model (Moody 2016) which even the author suggests is in most cases worse than just picking one of the methods.[64]

3. Hybrids between sentiment analysis and topic modeling

Some research Rogov et al. (2020) and Nyugen and Shirai (2015) found great success with joining sentiment analysis and LDA topic modeling technique and applying it to financial articles. The latter specifically developed a Topic Sentiment Latent Dirichlet Allocation (TSLDA) model which infers topics and text sentiment simultaneously can also be used to predict market returns with satisfactory performance.

This joint approach of sentiment and topics analysis proves to be the best suited for a recommendation-based article ranking as it combines well established and trusted sentiment information about an article and insights provided by topics which creates a strong metric of understanding each topics sentiment.

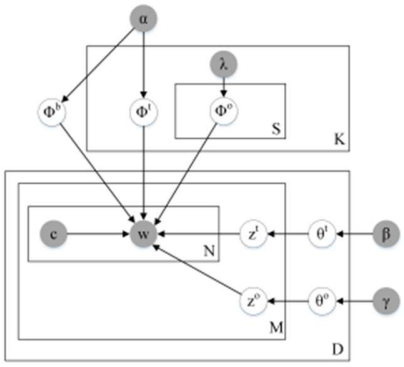


Figure 8 TSLDA representation

4. Market Return Prediction Modeling

A useful metric in building an article recommendation system is the measure of articles impact on the market (e.g. earnings or business reports having great importance on stocks price)[65]. As we're only interested in the influence the model does not need to predict exact future returns but rather the general direction in which the stock will move which additionally improves the accuracy of the model. The universe of viable prediction models is vast: from simple models like linear regression and decision trees to complex recurrent neural networks and Convolutional Long Short-Term Memory neural networks. It would require a separate report to understand and explain all advantages/disadvantages between them and evaluate their performance.

[63] Marketing Science Lecture
[64] lda2vec
[65] Netflix shares fall more than 35% after streamer loses over 200,000 subscribers

My previous research[66] which focused on evaluating models of different complexity applied to predicting market returns and research by Lu et al. (2020) which compares complex neural network-based models in forecasting stock prices both found CNN-LSTM model to be the most accurate in predicting stock returns. However, a big disadvantage of the model is its complexity and training time which are for final implementation justified by higher accuracy.

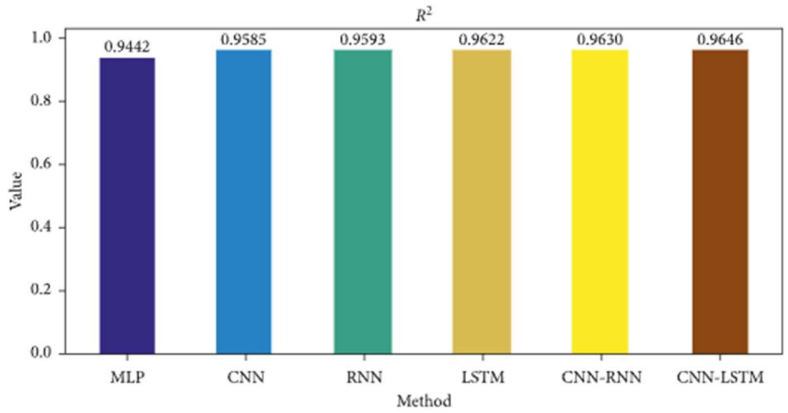


Figure 9 Results of Lu et al. (2020) study

[66] Data Analytics 2 – Individual Assignment

III. Available Data

This section will explore different streams of news data BlackRock should use in solution development it won't focus on streams of market data as BlackRock already gets it directly from multiple exchanges.

1. Breakdown of News Data

Ideally the BlackRock and recommendation solution would monitor all news outlets and as many different news types as possible but, this would unproportionate increase costs and complexity with data stream integration, data processing which when talking about text data isn't a trivial task, and data storage.

That is why BlackRock must compromise when selecting the appropriate data stream. We will explore 3 main news categories and rank which ones would be the most appropriate for BlackRock use and then show real-world news stream that could be integrated with the solution.

Financial News are news which report on market changes, public company reports (such as earnings reports), economic factors which influence the market (inflation, GDP), current news about the operations of public companies, and analyst reports. Financial news is also the type of news most used in research papers. This news type is the first news type BlackRock should implement as it directly influences the markets making it the baseline which analysts should understand intimately.

General News are news about the political landscape, legal activity, health crisis (Covid-19), scientific discoveries, education, entertainment industry, and other general news stories. The line between some general and financial news can often be blurred as political and legal issues can also affect the market (specially the currency markets)[67]. The impact of general news on the markets is less researched but generally still considered to impact the market.[68] Therefore, BlackRock should only consider researching general news after fully integrating financial news which could provide beneficial as not many investors specifically focus on analyzing general news providing BlackRock with additional information used to exploit the market.

Niche Industry News are news about very particular industries that only effect a small number of securities often in a non-obvious way. These are news such as sport matches results, video game releases, patent filings, etc. which can affect a very specific corner of the market. Naturally these types of news will be the least analyzed meaning they also provide the greatest benefit in researching them to find potential security mispricing. However, search for this information doesn't guarantee its usefulness and research of this news type should only be considered after exhausting the streams mentioned above.

[67] Black Rock interview
[68] The Impact of Non-Financial Reporting on Stock Markets in Emerging Economies

Going further we'll focus on explaining different sources of financial news BlackRock could implement into the solution. However, it's important to mention that BlackRock already has access to different news streams which they don't disclose and assuming they provide similar news types should be included in the final solution. The focused-on news streams are from companies that publicly provide news API's, any important news sources (such as CNBC) that don't readily provide an API could also be integrated given BlackRock would strike a deal with the vendor. Additionally, enterprise scale implementations of each news API are behind a paywall ranging from 80$ to 1750$ per month which, given BlackRock wouldn't require more than a few API keys, would be negligible costs to the company.

2. News API

News API is a self-proclaimed replacement for Google News API which was discontinued. The News API collects all news including financial news from all the different sources which are posted on Google News which creates an impressive catalog of different news sources thus potentially eliminating the problem of using a single API.

There are 2 endpoints (what the API returns):[69]

- Everything: returns all news in a selected time-period, from selected sources, in a selected language including a selected keyword or phrase. This is useful to create a historical backlog of news regarding a particular company or industry that is selected with keywords.
- Top Headlines: returns the top breaking headlines for a selected country, category, source. This is useful to use as a real time news tracker of most influential news.

Advantages:

- Real-time news availability
- Cross origin resource sharing enabled
- 99.95% uptime which is more than enough as our application isn't meant for HFT solutions
- Different news sources

Limitations:

- 2,000,000 requests per month (with ability to purchase additional requests)
- 4-yeah historical range, this is the biggest limitation of News API as creating a large news articles dataset is important for determining time-tested news topics and backtest model performance.

Figure 10 News API

[69] NewAPI

3. New York Times API

New York Times API is used to retrieve the news articles published by NYT. Its use was primarily designed to embed snippets of NYT articles into a website/webapp. It contains all news articles published by NYT going back to 1851, creating an impressive catalog that could be used for a historical backlog.


```
{
  "response": {
    "meta": {
      "hits": 25,
      "time": 332,
      "offset": 0
    },
    "docs": [
      {
        "web_url": "http://thecaucus.blogs.nytimes.com/2012/01/01/virginia-at
        "snippet": "Virginia's attorney general on Sunday backed off of a pro
        "lead_paragraph": "DES MOINES -- Virginia's attorney general on Sunda
        ...
      }
    ],
    "facets": {
        "day_of_week": {
            "_type": "terms",
            "missing": 1871790,
            "total": 13098462,
            "other": 3005891,
            "terms": [
              {
                "term": "Sunday",
                "count": 3122347
              },
              ...
```


There are 3 endpoints:[70]

- Archive: returns all articles for a selected month going back to 1851, however, it doesn't allow to search by keywords or phrases.
- Article Search: returns all articles for a selected time period, desk on a searched keyword or phrase. The detailed breakdown of NYT desks (news categories) allows for a very specific search which improves the data quality. This is a great resource to create a historical backlog for algorithm training purposes.
- Most Popular: returns the most popular articles from NYT, this could be very useful as its basically a labeled data source given the popularity of the article for a company/industry could be tied to market response.

Advantages:

- The main advantage is of course being able to access the entire NYT article catalog from 1851 proving especially useful when analyzing how topics about a company/industry change through time.

Limitations:

- Content limited to only the headline and leading paragraph. Currently NYT does not provide the API user access to the full content of an article but only returns the article title and leading paragraph this is due to copyright reasons (their response) and most likely due to the main use case of the API which is embedding their news. However, in the FAQ they state that they can provide the API to partner users with exclusive access to full content.[71]
- Lack of real time feed: The API documentation doesn't state if the API requests can get the most recent articles or if there is a delay. This is a major limitation as one of the analysts' goals is to perform analysis faster than the market and having news delays isn't an option.

[70] New York Times API
[71] New York Times Dev FAQ

4. Bloomberg API

Bloomberg has an API to request data from the Bloomberg Terminal which includes the news articles. Bloomberg Terminal is one of the most used financial data systems used by professional investors and investment firms. On top of providing market and fundamental data it also provides the user with news which are sentimentally labeled.

Endpoints:

Endpoints are the same as in Bloomberg Terminal.[72]

Limitations:

- Bloomberg's API requires the user to be very familiar with Bloomberg Terminal functionalities and calls as the process requires to make calls in C++ or C#. This introduces additional complexity as the nature of the API calls vastly differ other vendors.

5. EOD Historical Data

EOD Historical Data is a platform providing various API services for different types of financial data. It can return Market, Fundamental, Technical, Economic, and News Data the return is based on user determined factors. It provides a great all-round service to request different types of data.

Endpoint:[73]

- The API request can be made either on a specific keyword/phrase or on a list of 50 predefined news tags (such as balance sheet, earnings results, net income, market research reports…). The return of the request is a list of articles and their contents, date of publication, link to article, and ticker symbols mentioned in the article. Having the data on the mentioned tickers simplifies the categorization of articles.

Advantages:

- Data from different news sources
- Availability of different data variety
- 30+ years of news back catalog

Limitations:

- 1000 calls a minute which isn't a limitation as the solution isn't meant for HFT.
- Majority of news article from Yahoo Finance

[72] Getting Started on the Bloomberg Terminal
[73] EOD Historical Data – Financial News API

6. Financial Times API

Financial Times provide a readily implementable API but details about it are only discussed with interested clients.[74]

7. Yahoo Finance API

Yahoo Finance is a wildly used source of financial data by academics because it's easy to web-scrape (given permission), has an easily implementable API with RapidAPI, and provides different types of financial data.

Endpoints:[75]

- News: returns a list of news on a specific ticker or a specific news article.
- Conversation: returns a list of conversations on a given ticker

Limitations:

- 5 requests per second
- Single data source, like other API's.
- Unclear news backlog

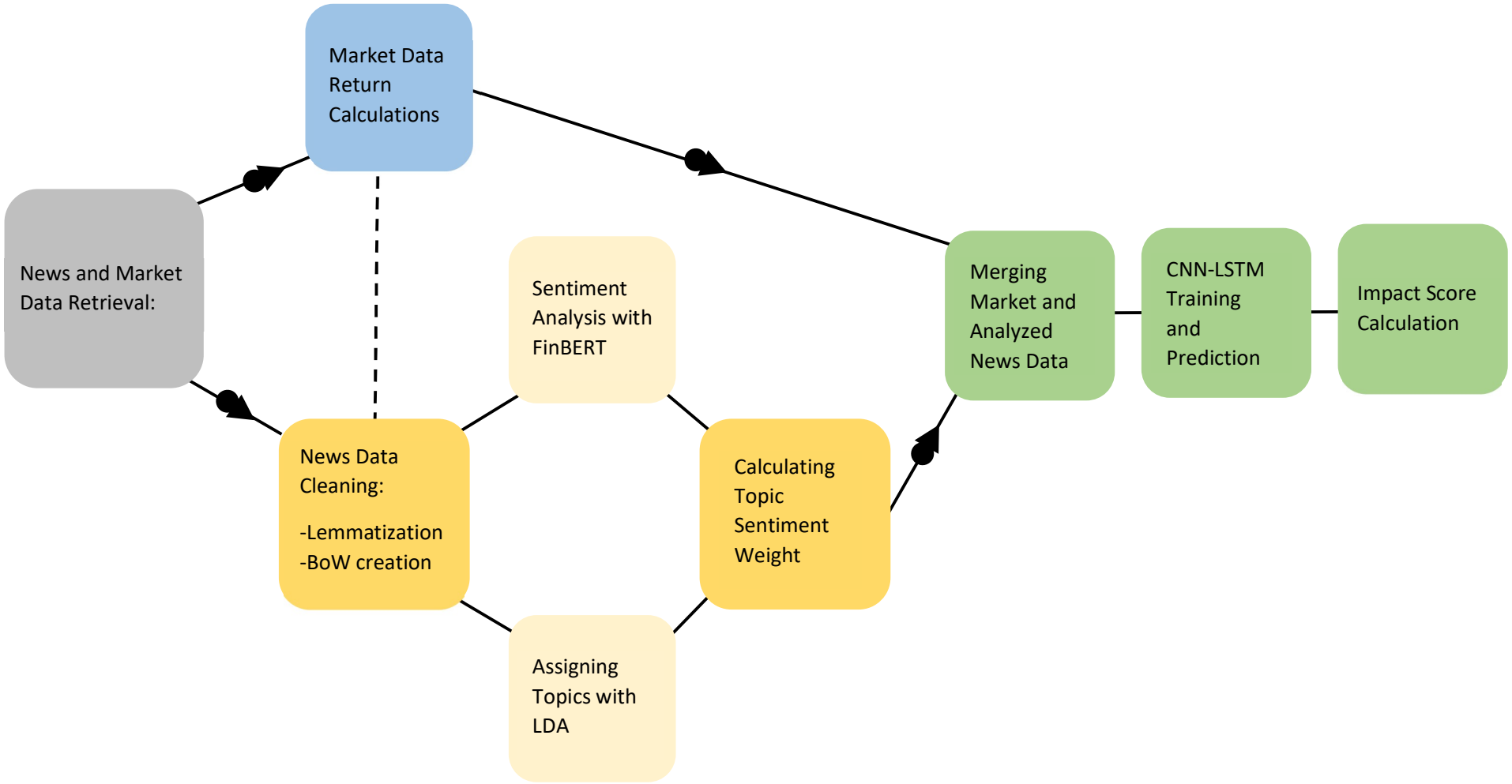
From the researched news sources BlackRock should prioritize the use of EOD Historical Data API and News API as they provide the most extensive data range with news articles being aggregated from different sources creating a big advantage over other single source APIs. Alternatively, given BlackRock already pays for Bloomberg Terminal access and if its API calls can be integrated with the solution BlackRock should certainly choose this option.

[74] Financial Times API
[75] Yahoo Finance API

Solution Development

I. Solution Overview

This section will provide a clear overview of the processing steps in the developed solution.



II. Solution Justification

As outlined in previous sections the developed solution uses a joint approach between sentiment and topic analysis as it provides both the insights into article topics and article sentiment which can be used to derive topic sentiment. This calculation of topic sentiment creates additional value for analysts by presenting a better weight breakdown of the article. For example, topic *Chip Shortage* has on average a negative presence in the articles, so article with a high presence of this topic and a large negative sentiment will be flagged as more negative. Inversely, article that will talk about end of chip shortage (positive sentiment) will have to be very assuring to be considered as truly positive.

Additionally, models calculating sentiment, topics, and market returns are models that research shows the greatest results.

Sentiment analysis is performed using FinBERT, because it's the most accurate out of model-based models, its complexity is allowable, and is more specific than lexicon-based approaches making it the best for sentiment analysis of financial articles.

Topic modeling is performed using LDA as most research finds its performance satisfactory. Ideally BlackRock would test the uses of Neural Topic Models which promise great results[76] but haven't been implemented or tested on financial articles. Because the training and reconfiguring of these models to work for financial texts would be costly and complex it's beyond the scope of this dissertation.

Market return predictions are performed with a CNN-LSTM model which as mentioned in previous sections is the most accurate model architecture for predicting market returns. These models can be fine-tuned by rigorous testing to find the optimal parameters which is again beyond the scope of this dissertation. The final model structure will be the same as one I found to be best performing when predicting market returns in my previous Data Analytics 2 individual research.

Building the recommendation system based on results from state-of-the-art text analysis instead of traditional word counting or volume/visits measuring approach[77] is used because it accounts for the differences between articles which would otherwise be lost.

[76] Neural Topic Models
[77] How to do SEO for News

III. Detailed Breakdown of Solution

This section will provide a break down of the steps took to derive the final article recommendation score.

1. Data Retrieval

Above we identified the ideal data sources for BlackRock, however not the same sources can be applied to this dissertation due to costs and exclusivity. That is why we used the free API provided by EOD to retrieve news data and a free API by Yahoo Finance to retrieve daily market data.

When it comes to data range there are three reasonable scenarios:

The first is to train the models on data ranging as far back as possible to identify constant financial topics and see which aspects are included in new news articles.

The second is to initially train the models on far reaching data and then periodically update them with recent news. This would provide a mixture of traditional topics while still accounting for only the most important current topics.

The third approach is to train the models on recent data, this would mean the models are configured to understand all the current topics. This approach comes with an important consideration of how old recent news is and how important is past news to current topics (e.g. is a topic about a unique circumstance like implementing XBRL that was significant a year ago still important).

The approach used in this solution will be the third one, as its ability to provide insights into current situation eliminates the most uncertainty analysts have with new articles (e.g. knowing importance of earnings reports is already facilitated but understanding how chip shortage affects different industries isn't) and reduces the computational complexity brought by analyzing 30+ years of data.

The specific date range will be from 1 January to 1 May 2022 providing us with 4 months' worth of data. The range was selected because it allowed us to retrieve about 1 stock's worth of news data per day.

The number of analyzed stocks chosen is 10, because of the API and computational limitations choosing 1 additional stock requires 1 days' worth of API calls and increases the execution time of the algorithm by 30min. Of course, choosing more than 10 stocks would be necessarily if the solution was implemented but for purposes of this presenting a proof-of-concept in this dissertation it's enough. The chosen companies are the top 10 tech companies in S&P 500 by index weight. Only tech companies were chosen because analyzing companies in the same industy allows to determine industry specific topics. If the analysis is performed on stocks from different industries creating industry specific topic models and using them to create an ensamble general model would be advised to understand the difference in topic between industries, have industry specific topics, and still have general financial-topics.

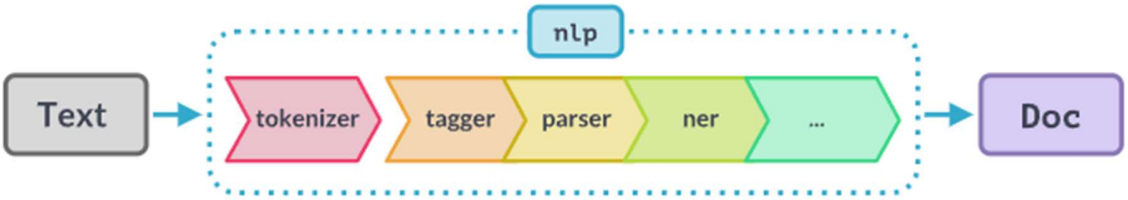
AAPL
AMZN
AVGO
FB
GOOG
MA
MSFT
NVDA
TSLA
V

2. Data preprocessing

Text data is famously difficult to preprocess because of different ways of storing it and sheer variety (e.g. different article structures).

As learned from Data Analytics 2 and Marketing Science modules there are 4 steps to clean text data:

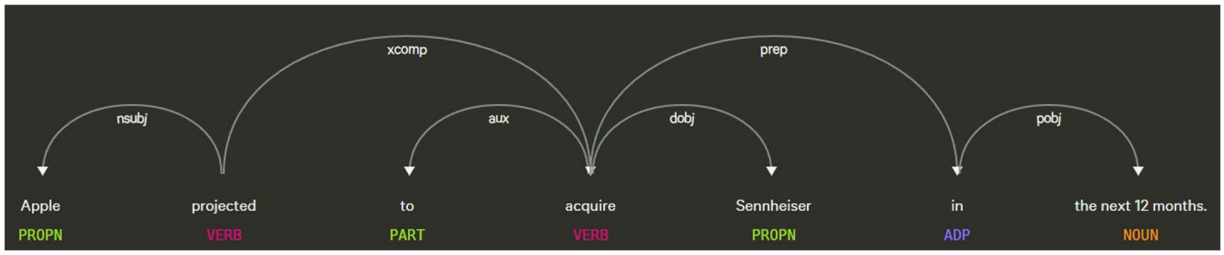
- Tokenization, is the step where text is split into separate words
- Removing stop words, removes unnecessary words such as: and, when, how, can, I, …
- Part-of-Speech tagging, assigns a POS tag to each word, a pos tag determines the word's type: Noun, Verb, Adjective, etc. A common practice that we used as well is to only keep the words with following POS tags[78]: Noun, Verb, Adjective, Adverb, Proper Noun as they're the only word types that convey meaning
- Lemmatization is the process of converting each word to its lemma, by doing this we standardize the meaning of the words (e.g. buying, bought, buy get assigned the same lemma). This step is crucial to narrow down the universe of viable words but disregards tense which is in financial context important as it signifies if a news article is reactive, reporting on past market changes, or proactive, reporting on future forecasts or present state. Including this consideration in the models would significantly increase the complexity and should be explored by BlackRock but is beyond the scope of this dissertation.



To clean the text data, we used the leading NLP python library spaCy[79], which processes the text entirely and provides additional insights that can be exploited. We will explore the two most important ones for our context and explain their uses.

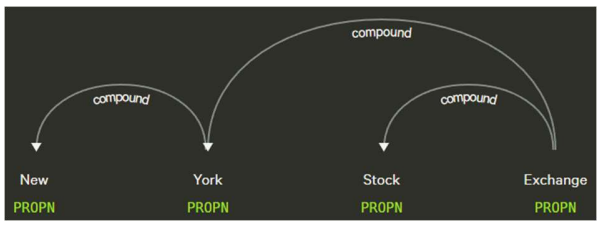
[78] spaCy – Linguistic Features
[79] spaCy – Facts & Figures

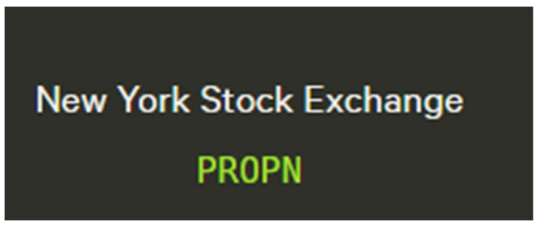
The first is dependency parser[80], which assigns which words depend on one another. We use the idea that different words have different importance levels in a sentence by looking at the number of words a word is depended by and taking that as its importance rank. This allows us to take a more heuristic approach to bag-of-words creation eliminating one of its major drawbacks which is that all words in a text are considered equal. This approach will prove important as it will improve topic legibility.



Looking at the example above clarifies our thinking, the sentence *"Apple projected to acquire Sennheiser in the next 12 months.",* to a human clearly signifies that the dominant topic is Acquisitions. Using the traditional bag-of-words approach would consider all the words to have equal impact which would lead worse topic results. Our adjusted bag-of-words approach would in this case value the word *acquire* the most making it easier for LDA to recognize it as one of the topics.

The second insight spaCy brings is noun chunks[81]. This is a way of accounting for language phrases. Take the phrase *New York Stock Exchange* for example, splitting the words traditionally would take each separately but noun chunks metric allows us to consider it as one.





[82]

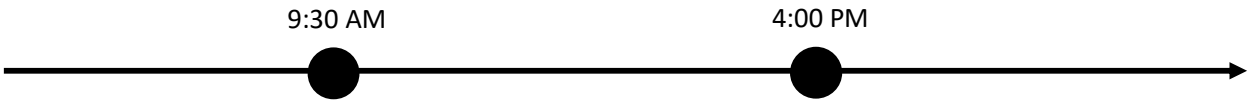
But because this approach doesn't look only at phrases common across all documents it won't be used. This is because even the slightest difference between what would be considered the same phrases: *Dow Jones* and *Dow Jones Industrial Average* would result in being considered as separate entities which would reduce their relative count.

[80] spaCy – Linguistic Features
[81] spaCy – Linguistic Features
[82] displaCy Dependency Visualizer

Market data, processing market data is significantly easier than text data. The main metric that had to be calculated were the returns on the stocks.

With previous research showing (Heston and Sinha, 2016) that news can reliably predict stock returns for only 1 to 2 days, we chose to predict for 3 different return calculations: same day return, next day return, and next 2 days return.

However, as Yahoo Finance report market data from NYSE which is opened 5 days a week from 9:30 to 16:00 but news articles come out every day 24/7 we had to look at the publish date of the article and assign it the correct day price. Accounting for the time zone in which the news was published was also taken into consideration.



Market returns for articles published before 9:30 were calculated by using the market opening price and previous day's closing price.

$$R = \frac{Open - Last\ Day\ Close}{Last\ Day\ Close}$$

Market returns for articles published while the market was open were calculated by using open and close prices.

$$R = \frac{Close - Open}{Open}$$

Market returns for articles after 16:00 were calculated by using the closing price and next day's opening price.

$$R = \frac{Next\ Day\ Open - Close}{Close}$$

Returns for next day and next 2 days were calculated in the same manner but adjusted for end date.

Additionally, as we're creating a recommendation system predicting the exact return isn't necessary, so we group the returns into 5 distinct categories: Strong bearish, weak bearish, neutral, weak bullish, and strong bullish. Those levels are based on return statistics, bottom 5% returns are strong bearish, 5-40% are weakly bearish, 40-60% are neutral, top 40-5% are weakly bullish, and top 5% of returns are categorized as strong bullish.

3. Sentiment Analysis

To perform sentiment analysis, we used a pretrained FinBERT model downloaded from Hugging Face[83] and used PyTorch[84] to speed up its process by utilizing GPU. To achieve the fastest prediction results and not pay for external computing servers we used a Tesla T4 GPU provided by Google Collaboratory for free[85].

The model required a careful construction of the data pipeline as the data had to be broken down into chunks of 512 characters and fed article by article into the model to avoid running out of memory.

The total execution time of sentiment prediction for 20,400 articles was around 3 hours. This time can be improved by improving the pipeline and running on servers with more memory and compute units.

The prediction results were a probability spit between the positive, negative, and neutral sentiment.

stock	date	title	content	link	positive	negative	neutral
AAPL	2022-04-30 15:03:10+00:00	Warren Buffett: We didn't repurchase any Berks...	Warren Buffett hasn't used the market pullback...	https://finance.yahoo.com/news/warren-buffett-...	0.019490	0.554205	0.426305
AAPL	2022-04-30 14:28:14+00:00	12 Safe Stocks To Buy For Beginner Investors	In this article, we discuss the 12 safe stocks...	https://finance.yahoo.com/news/12-safe-stocks-...	0.282421	0.054441	0.663138
AAPL	2022-04-30 12:25:00+00:00	Buy These 2 Streaming Video Stocks Instead of ...	Netflix (NASDAQ: NFLX) was once considered the...	https://finance.yahoo.com/m/2ef7d75d-b68b-38f0...	0.008728	0.973723	0.017548
AAPL	2022-04-30 12:10:00+00:00	This Mid-Cap Stock Could Deliver Blockbuster G...	Fabless semiconductor company Cirrus Logic (NA...	https://finance.yahoo.com/m/be50e767-d2b6-3dbb...	0.884025	0.008356	0.107619
AAPL	2022-04-30 12:04:10+00:00	How Much Of Apple Inc. (NASDAQ:AAPL) Do Instit...	The big shareholder groups in Apple Inc. (NASD...	https://finance.yahoo.com/news/much-apple-inc-...	0.030488	0.032551	0.936961

Figure 11 Sentiment Analysis Results

[83] ProsusAI – FinBERT - Hugginface
[84] PyTorch
[85] Google Collaboratory – GPU Architecture

4. Topic Analysis

To perform topic analysis, we used an LDA model from gensim[86]. The model had to determine 30 unique topics and trained for 40 minutes. The number of topics was chosen by running the LDA model on randomized subsamples of the data and ranking them on the legibility of the topics. We tested the model on 10, 20, 30, 40, 50, and 100 topics and 30 topics proved the best combination between the uniqueness of the topics and readability.

To determine the topic labels, we used an interactive visualization made with pyLDAvis library[87]. This is the best way of visualizing and subsequently determining the topics as it allows for a deep exploration of key words in each topic and how unique are they to the topic.

The identified topics range from traditional finance-related topics such as stock price, company management, earnings, quarter estimates, and dividends, also tech industry-specific topics such as metaverse and gaming, Elon Musk, new technology announcements, digital transactions, electric vehicles, and chip shortage, and topics about the political space that influence the companies primarily Ukraine Invasion.

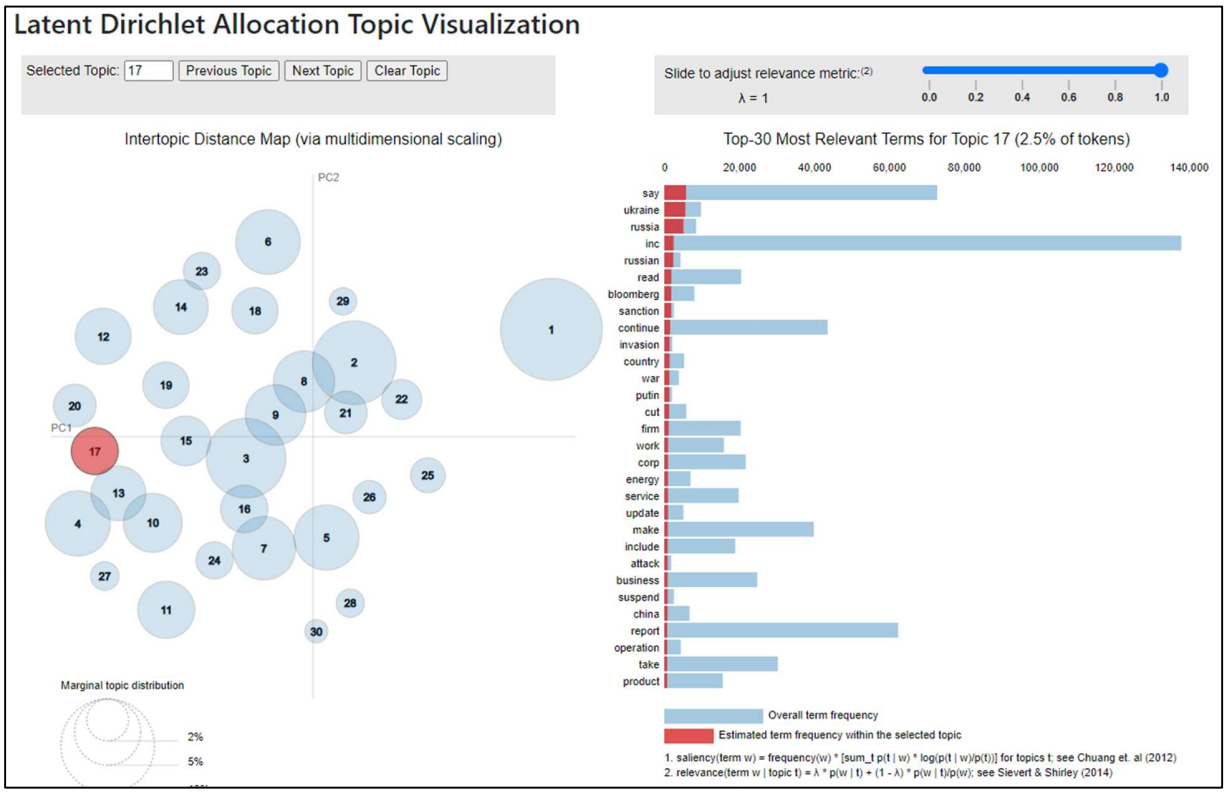


Figure 12 LDA topic 17 - Ukraine Invasion

[86] Gensm - LDA
[87] pyLDAvis

5. Topic Sentiment Allocation

To calculate the topic sentiment, we used a linear regression model provided by scikit-learn library[88], used topics as inputs, and sentiment categories as outputs. A linear regression was used because it returns coefficients that explain how topics influence the sentiment.

The coefficients were then used to calculate the mean and mean deviation from the coefficient for all topics. This calculation was performed because we're only interested in how topics in sentiment compare between each other. This topic sentiment score was calculated for positive, negative, and neutral sentiments.

$$Topic\ Sentiment\ Score = Topic\ weight * (Sentiment + Sentiment\ Topic\ Coefficnent)$$

The overall article topic sentiment was then calculated by the following formula:

$$Overall\ Topic\ Sentiment = (Topic\ Positive\ Score - Topic\ Negative\ Score) * (1 - Topic\ Neutral\ Score)$$

Usually, the sentiment score is calculated only by subtracting the negative sentiment from positive, but this doesn't account for the sentiment strength. Adjusting the sentiment score by including the neutral score favors less vague articles with stronger sentiment polarity, which are also the articles most likely to have an impact on the market. This approach can present a problem when analyzing longer articles where the overall sentiment will be more neutral as it lowers their score and inversely boosts the score of short articles which have a clearer sentiment direction.

6. Market Return Prediction

To predict the impact news has on the market returns we trained a CNN-LSTM neural network on assigned article topics and sentiment, we avoid using market data as inputs since the models with market data can easily overfit as concluded in research done for Data Analytics 2.

The model has 234,695 trainable parameters and trains for 5000 epochs which takes about 1 hour. Overall, we trained 3 separate models for three market return time ranges (daily, next day, 2 day).

Layer (type)	Output Shape	Param #
conv1d_8 (Conv1D)	(None, 33, 64)	256
max_pooling1d_8 (MaxPooling 1D)	(None, 16, 64)	0
lstm_12 (LSTM)	(None, 16, 50)	23000
dropout_30 (Dropout)	(None, 16, 50)	0
lstm_13 (LSTM)	(None, 50)	20200
dropout_31 (Dropout)	(None, 50)	0
dense_24 (Dense)	(None, 512)	26112
dropout_32 (Dropout)	(None, 512)	0
dense_25 (Dense)	(None, 256)	131328
dropout_33 (Dropout)	(None, 256)	0
dense_26 (Dense)	(None, 128)	32896
dropout_34 (Dropout)	(None, 128)	0
dense_27 (Dense)	(None, 7)	903

Total params: 234,695
Trainable params: 234,695
Non-trainable params: 0

Figure 13 CNN-LSTM model summary

[88] Sklearn – Linear Regression

Results on the training data show 88.9% accuracy but examining the model's performance on unseen data with a confusion matrix shows a grimmer picture.

The model performs poorly when determining whether the article has a positive or negative influence on the market. But performs rather well in predicting the strength of market movements. We can see that the model mostly predicts strong bearish returns which would generally be in line with the recent market movements.

Figure 14 S&P 500 Price

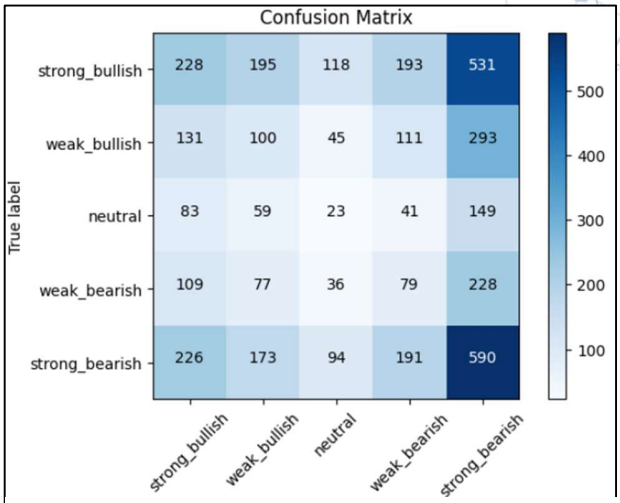


Figure 15 CNN-LSTM Confusion Matrix

As the recommendation system's main goal is to present the news articles with the biggest impact on the market and leave how exactly will the market react up to the analyst, to not influence their analysis and create overreliance on the model predictions, the results of the model are satisfactory.

7. Recommendation Score

To create the recommendation score we use the predictions calculated with the models. The main goal of the recommendation score is to account for both the future return predictions as well as the impact the presence of different topics has. To combine them along with article sentiment we'll use the predicted market return and overall topic sentiment score. The equation for doing this was derived by looking at how different scenarios affect the score.

So, when the predicted return is bearish and average topic score also suggest bearish returns the final recommendation score should be higher than if predicted return is bearish, but the topics sentiment suggest bullish returns.

To do this we use the following equation:

$$Impact\ score = |Predicted\ Return * Overall\ Topic\ Sentiment\ Score * \frac{Predicted\ Return + Overall\ Topic\ Sentiment\ Score}{Predicted\ Return}|$$

Predicted Return	Overall Topic Sentiment	Impact Score
3 (Strong Bullish)	0.7 (Strong Positive)	2.59
3	0.7	1.61
-3 (Strong Bearish)	-0.7 (Strong Negative)	2.59
-3	-0.7	1.61

As we can see the method provides accounts for the considerations made above as positive returns with opposite topic sentiment yield lower returns than having both be positive. The problem this method faces is only having positive values; this is a passable tradeoff as the recommendation system doesn't aim to suggest how exactly will the article impact the market but rather the scale of the impact. If this requirement would change the equation will also have to be adjusted.

IV. Functionality Showcase

This section will showcase the developed application and explain and justify the visualizations.

1. Topic View

This part of the dashboard shows the topic distribution created by LDA model for all the documents. It's meant for the analysts to use to fully understand a specific topic.

For example, by looking at topic_17 the analyst can clearly see this topic is about Russia's invasion of Ukraine and the Western sanctions on Russia.

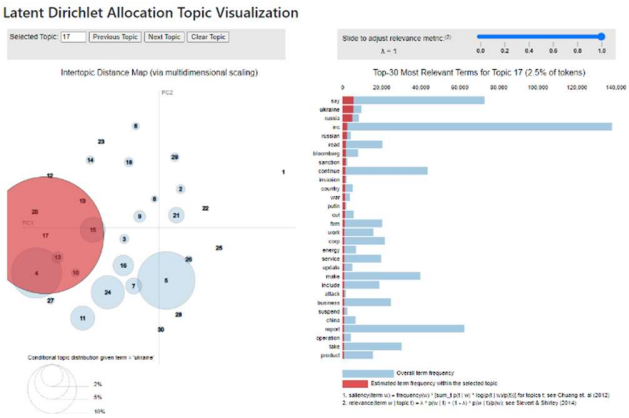


Figure 16 LDA visualization topic 17

2. Article Map

This part of the visualization allows the analyst to look at the complete picture of analyzed articles. Functionalities allow them to filter the articles by Companies and recorded market returns.

It lets them compare the companies between each other and see how the topic distribution between them varies.

Figure 17 Article Map - MA and Visa

For example, by comparing the topics of similar companies like Visa and Mastercard we can see most of the topics between them are as expected similar. However, topic_25 is distinctly present in articles about Visa and not Mastercard. By examining the topic, we see the topic is about Microsoft, Windows, and Azure – this insight is particularly useful as Microsoft integrated Visa into its Microsoft-Wallet system[89].

[89] Visa Integrates with Microsoft Wallet

The visualization also allows analysts to see how the sentiment of the articles is related to topics and daily returns.

Figure 18 Article Map - Visa Sentiment

The example above shows that topic_14 for Visa has an overly positive sentiment which are backed by bullish daily returns.

Finally, analysts can view the returns by topics for the selection of companies.

Figure 19 Article Map - Broadcom Returns

This overall visualization can help the analysts understand which topics are present for different companies, what are the topic and overall sentiments, and what are the market returns.

3. Company Tab

This part of the application is the actual article recommendation.

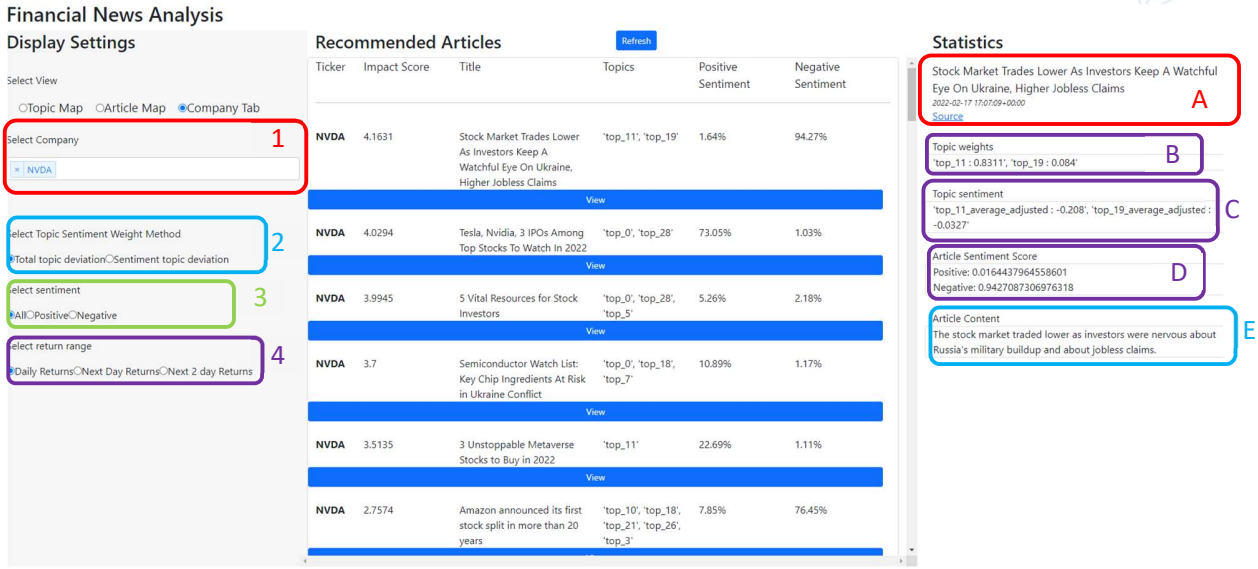


Figure 20 Company Tab

The left-hand side presents the analyst with options to filter articles by a particular company (1), by method of calculating the topic sentiment (2) (simple or adjusted), by article sentiment (3), or by return range (4).

Based on those filters the application then updates the middle part and shows them the articles and some article statistics with the highest recommendation score.

The analyst can then choose to click on the View button for an article and the right-hand side will update showing that article. This statistics tab will display the title, publish date, and link to the article source (A). Additionally, it will display the article topic breakdown (B), the article topic sentiment weight breakdown (C), and the article sentiment (D). Finally, it will also show the analyst the contents of the article so they can read it instantly (E).

This structure helps the analysts clearly define the space in which they want to search news for, shows them most important articles, and provides them with a detailed article analysis all to enable them to simplify the search for market influencing news.

V. Developed Solution Limitations

As mentioned above the main limitation of the developed solution is its scope as it only focuses on 10 stocks for news data ranging 4-months back. Another limitation is the lack of real-time news integration – this wasn't possible due to API limitations.

If BlackRock were to implement this solution the space of analyzed news would have to increase to either focus on an entire sector or entire market space, collecting both financial and general news, and the real-time news streaming with low delay would have to be implemented to stay competitive.

Another problem is not having good feedback on the performance of the solution (accurateness of topics, returns, sentiments) this could be easily corrected by implementing a feedback functionality that would allow the analysts to score how accurate or useful were the predictions and update the models based on those responses.

Additional limitation is the fact that the solution is a proof-of-concept and needs a long way to become an enterprise level solution which could be integrated with Aladdin.

The biggest limitation of the solution is the complexity as using the most advanced and accurate models for analysis creates significant delays on analyzing an article. BlackRock could reduce that delay by optimizing the pipeline and hiring sufficiently powerful servers (or use appropriate inhouse servers).

Solution Integration and Impact

I. Integration

To determine BlackRock's ability to fully develop and implement such a solution, the cost of integration, running costs, and the time frame in which they could realistically integrate it we need to understand their current resources, specifically personnel.

Figure 21 BlackRock Technology Division

Their position breakdown in technology careers suggests BlackRock values and employs both Data Scientists, Software Engineers, and people for Technology Integration which are the people with exact skills needed to develop and integrate the proposed solution. Looking at their current career offerings there's 194 open positions for data scientists[90], 124 for software engineers[91], and 20 for technology implementation experts[92]. This confirms that BlackRock already has a well-established team in these areas and see the value in expanding them which along with overall data and technology-based approach to solving problems[93] confirms BlackRock ability to undertake such a project.

The time of development is highly relative and dependent on amount of resources BlackRock would assign to the project and increased complexity that comes with scaling the solution to all news and developing it for an enterprise level.

Chi Software Development Center identifies the development of an application based on AI would require at least three-months for simple applications and at least twice as long for end-to-end solutions.[94]

[90] BlackRock Careers - Data Scientist
[91] BlackRock Careers – Software Engineers
[92] BlackRock Careers – Technology Implementation
[93] Video: BlackRock Lecture Series A Deep Dive Into BlackRock's Technology Ecosystem
[94] How to Make an AI App: A Massive Integration Guide for 2022

Taking into account this app would require a lot of collaboration between the development team and security analysts, that the solution isn't following a standard app creation steps, the complexity brought with the breath of analyzed data, and sheer importance of achieving the highest possible accuracy which would require rigorous testing; I assume the app would take from 2-5x time longer to develop than an end-to-end solution mentioned by Chi, which means around 1-2 years of development and integration time.

Expert	Approximate work time (hours)
Mobile apps developer	170
Backend developer	170
Front-end developer	120
Quality assurance engineer	130
Project Manager	40
Total	**630**

Figure 22 Worktime Hours for simple AI based app development

The app costs are highly correlated to data, server, and salary costs. Given that BlackRock doesn't reveal the extent of their computational capabilities and variety of news data sources we will assume that BlackRock would need to rent servers specifically for development but use already existing data sources.

The prices of servers for a project of this size and complexity would range from $10-20k[95] a month with cost estimates provided from Microsoft Azure[96] and Cirrascale[97] for developing, training, and testing and additional costs for integrating with Aladdin and publishing which are unknown.

The people costs would fall around $500,000 if we extrapolated the estimated made by Chi.[98]

II. Impact

The impact of this solution is hard to quantify since it would improve the workflow of analysts and improve the service offering to clients.

It would specifically reduce the time it takes the analysts to find relevant news and it would shift the selection of which articles to focus on from it being intuition and gut based to being backed by proper analysis and data – this would reduce the number of irrelevant and increase the number of relevant articles analyzed which would essentially reduce the opportunity costs which come with choosing to research a particular article. Additionally, it would allow the analysts to easily discover companies and investment opportunities that are beyond their scope of securities potentially increasing BlackRock's diversification and returns.

If we make some generous assumptions that the time it takes the analysts to constantly monitor news and read articles amounts to 10% of their work[99] the proposed solution could reduce that time spend to 6-8% but more importantly improve the quality of the analysis.

[95] Appendix 1: Computational Requirements
[96] Microsoft Azure Pricing Calculator
[97] Cirrascale: GPU Cloud Server and Storage Pricing
[98] How to Make an AI App: A Massive Integration Guide for 2022 – Development Costs
[99] What does a Financial Analyst Do?

Conclusion

As discussed throughout the report there is a lack of news recommendations based on implementing state-of-the-art news analysis techniques. The research of financial news analysis shows there is a significant potential to use it for predicting future market returns especially with sentiment analysis and topic modeling.

Those models are used to develop a final solution which is an article recommendation system that provides additional analysis for each article. The developed solution shows very promising when used to examine the top recommendations and article breakdown by topics as some of the topic article breakdowns have specific and valid reasonings behind them.

After examining the development and implementation costs and comparing them to the significant value it generates to security analysts, I recommend BlackRock pursues this solution and begins its development.

Appendix

I. Computational Requirements:

Current solution was built on Nvidia's Tesla T4 GPU's with 16GB of video memory and 6-core processor and took about 5 hours to run.

$$16GB * 6 - Cores = 5h = 10\ companies * 4\ months$$

Assuming BlackRock would want to reduce the runtime to a few, a higher core units must be considered.

Microsoft Azure Server cost = $2k/month

48 Cores, comparable GPU, 450GB ram

$$X * 450GB * 48 = 5min = 1000\ companies * 20\ years$$

$$X = number\ of\ servers \sim 9$$

$$Total\ Server\ Price = 18k$$

Cirrascale server cost = 15k/month

8 GPUs, 512GB ram, 40 cores

$$X * 450GB * 8 * 40 = 5min = 1000\ companies * 20\ years$$

$$X = number\ of\ servers \sim 1.3$$

$$Total\ Server\ Price = 20k$$

II. Code Used for Solution Development

The code I wrote to develop the solution spans many different files and Jupyter notebooks and would cause only confusion within the report. That is why I'm attaching a link to my github repository where all the code used in the report will be available.

Solution app Github: https://github.com/weirdaxe/msin0032

Development code Github: https://github.com/weirdaxe/msin0032-dissertation-code

Bibliography

I. Interview

In 2018 I got invited to the BlackRock office in London for an exclusive open day. This was part of the "reward" the National Bank of Slovenia gave us for wining the national Generation Euro Student Awards.

We had an initial rundown of BlackRock's services, operations, and culture by their PR department and a 2-hour interview with foreign exchange analyst. He told us about the everyday workflow of analyst teams, the challenges they face and how the analyst team is closely integrated with the trading department.

We then got to spend another 2-hours with an active BlackRock trader during one of the most turbulent times for the British Pound Sterling.

Most of the ideas about the analyst's workflow and news importance come from these interviews.

II. References

1. ARTICLE:

BlackRock Surges past $10tn in AUM
Ft.com. 2022. BlackRock surges past $10tn in assets under management. [online] Available at: <https://www.ft.com/content/7603e676-779b-4c13-8f46-a964594e3c2f> [Accessed 17 May 2022].

BlackRock – About us
BlackRock. 2022. About BlackRock | BlackRock. [online] Available at: <https://www.blackrock.com/corporate/about-us> [Accessed 17 May 2022].

BlackRock Form 8-K
BlackRock. 2022. BlackRock form 8-K. [online] Available at: <https://d18rn0p25nwr6d.cloudfront.net/CIK-0001364742/1cbddb43-cbb0-46cb-b7c2-1b9c67410cbb.pdf> [Accessed 17 May 2022].

BlackRock – Investment Funds
BlackRock. 2022. Investment Funds. [online] Available at: <https://www.blackrock.com/uk/products/product-list#!type=all&style=44342&view=perfDiscrete> [Accessed 17 May 2022].

BlackRock FMA – Our Services
BlackRock. 2022. Our Services - Financial Markets Advisory (FMA) | BlackRock. [online] Available at: <https://www.blackrock.com/financial-markets-advisory/fma-services/service-offerings> [Accessed 17 May 2022].

Institutional Investor
Corporate Finance Institute. 2022. Institutional Investor. [online] Available at: <https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/institutional-investor/> [Accessed 17 May 2022].

iShares
BlackRock. 2022. Exchange-Traded Funds (ETFs) | iShares UK – BlackRock. [online] Available at: <https://www.ishares.com/uk/individual/en> [Accessed 17 May 2022].
Aladdin FAQs
BlackRock. 2022. Aladdin FAQs | BlackRock. [online] Available at: <https://www.blackrock.com/aladdin/resources/faqs> [Accessed 17 May 2022].

BlackRock Fund Charges
BlackRock. 2022. Fund charges - Products - BlackRock. [online] Available at: <https://www.blackrock.com/uk/professionals/solutions/fund-charges> [Accessed 17 May 2022].

BlackRock Active Equities
BlackRock. 2022. Active Equities | BlackRock. [online] Available at: <https://www.blackrock.com/us/individual/education/equities/active-equities> [Accessed 17 May 2022].

BlackRock ESG
BlackRock. 2022. ESG Integration and ESG Integration Statement at BlackRock. [online] Available at: <https://www.blackrock.com/us/individual/investment-ideas/sustainable-investing/esg-integration> [Accessed 17 May 2022].

AMC Vs. Game Stop
SeekingAlpha. 2021. AMC Vs. GameStop: How To Compare These 2 Short Squeeze Stocks (NYSE:GME). [online] Available at: <https://seekingalpha.com/article/4441425-amc-vs-gamestop-short-squeeze-stocks> [Accessed 17 May 2022].

Modeling Stock Returns and Risk Management in the Shipping Industry
Sunil K. Mohanty & Roar Aadland & Sjur Westgaard & Stein Frydenberg & Hilde Lillienskiold & Cecilie Kristensen, 2021. "Modelling Stock Returns and Risk Management in the Shipping Industry," JRFM, MDPI, vol. 14(4), pages 1-25, April.

Stock Market Returns and Clinical Trial Results of Investigational Compounds: An Event Study Analysis of Large Biopharmaceutical Companies
Hwang TJ (2013) Stock Market Returns and Clinical Trial Results of Investigational Compounds: An Event Study Analysis of Large Biopharmaceutical Companies. PLOS ONE 8(8): e71966. https://doi.org/10.1371/journal.pone.0071966

Customer-Based Corporate Valuation for Publicly Traded Non-Contractual Firms
McCarthy, Daniel and Fader, Peter, Customer-Based Corporate Valuation for Publicly Traded Non-Contractual Firms (March 9, 2018). Available at SSRN: https://ssrn.com/abstract=3040422 or http://dx.doi.org/10.2139/ssrn.3040422

Aladdin Risk
BlackRock. 2022. Aladdin Risk | BlackRock. [online] Available at: <https://www.blackrock.com/aladdin/products/aladdin-risk#:~:text=Aladdin%20Risk%20is%20a%20subset,act%20with%20speed%20and%20precision> [Accessed 17 May 2022].

Aladdin's Benefits to Risk Managers
BlackRock. 2022. Aladdin's Benefits to Risk Managers. [online] Available at: <https://www.blackrock.com/aladdin/benefits/risk-managers> [Accessed 17 May 2022].

Aladdin's Benefits to Portfolio Managers
BlackRock. 2022. Portfolio managers | Aladdin. [online] Available at: <https://www.blackrock.com/aladdin/benefits/portfolio-managers> [Accessed 17 May 2022].

Financial news predicts stock market volatility better than close price
Atkins, Adam, Gerding, Enrico and Niranjan, Mahesan (2018). Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science.(doi:10.1016/j.jfds.2018.02.002 <http://dx.doi.org/10.1016/j.jfds.2018.02.002>).

Bloomberg - News Headlines Powered by AI.
Bloomberg.com. 2022. Bloomberg - News Headlines Powered by AI. [online] Available at: <https://www.bloomberg.com/professional/product/news/#:~:text=With%20the%20AI%20News%20Importance,continue%20to%20see%20all%20stories> [Accessed 17 May 2022].

In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse
El-Haj, Mahmoud and Rayson, Paul and Walker, Martin and Young, Steven and Simaki, Vasiliki, In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse (March 2019). Forthcoming in Journal of Business Finance and Accounting , Available at SSRN: https://ssrn.com/abstract=3330757

Google News Ranking
Support.google.com. 2022. Ranking within Google News - News Publisher Help. [online] Available at: <https://support.google.com/news/publisher-center/answer/9606702?hl=en-GB> [Accessed 17 May 2022].

An algorithm for unsupervised topic discovery from broadcast news stories
Sista, Sreenivasa & Schwartz, Richard & Leek, Timothy & Makhoul, John. (2002). An algorithm for unsupervised topic discovery from broadcast news stories. Proceedings of the 2nd International Conference on Human Language Technology Research. 10.3115/1289189.1289267.

The Impact of Firm-Specific Public News on Intraday Market Dynamics: Evidence from the Turkish Stock Market
Baklaci, H. F., Tunc, G., Aydogan, B., & Vardar, G. (2011). The Impact of Firm-Specific Public News on Intraday Market Dynamics: Evidence from the Turkish Stock Market. Emerging Markets Finance & Trade, 47(6), 99–119. http://www.jstor.org/stable/41343443

SEC Form 10-K Structure
Sec.gov. n.d. SEC - Form 10-K Structure. [online] Available at: <https://www.sec.gov/files/form10-k.pdf> [Accessed 17 May 2022].

EUROPEAN SINGLE ELECTRONIC FORMAT - XBRL
Esma.europa.eu. 2020. European Single Electronic Format - XBRL. [online] Available at: <https://www.esma.europa.eu/policy-activities/corporate-disclosure/european-single-electronic-format> [Accessed 17 May 2022].

The Actual Difference Between Statistics and Machine Learning
Stewart, 2019. The Actual Difference Between Statistics and Machine Learning. [online] Medium. Available at: <https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3> [Accessed 17 May 2022].
Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research
Fisher, I. E., Garnsey, M. R., and Hughes, M. E. (2016) Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. Intell. Sys. Acc. Fin. Mgmt., 23: 157– 214. doi: 10.1002/isaf.1386.

In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse
El-Haj, M, Rayson, P, Walker, M, Young, S, Simaki, V. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. J Bus Fin Acc. 2019; 46: 265– 306. https://doi.org/10.1111/jbfa.12378

Financial news predicts stock market volatility better than close price
Atkins, Adam & Niranjan, Mahesan & Gerding, Enrico. (2018). Financial News Predicts Stock Market Volatility Better Than Close Price. The Journal of Finance and Data Science. 4. 10.1016/j.jfds.2018.02.002.

Quantifying the Relationship Between Financial News and the Stock Market
Alanyali, Merve & Moat, Helen & Preis, Tobias. (2013). Quantifying the Relationship Between Financial News and the Stock Market. Scientific reports. 3. 3578. 10.1038/srep03578.

How Does News Affect Stock Return Volatility in a Frontier Market?
Emenike, K. O. and Enock, O. N. (2020) 'How Does News Affect Stock Return Volatility in a Frontier Market?', Management and Labour Studies, 45(4), pp. 433–443. doi: 10.1177/0258042X20939019.

More Than Words: Quantifying Language to Measure Firms' Fundamentals
TETLOCK, P.C., SAAR-TSECHANSKY, M. and MACSKASSY, S. (2008), More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance, 63: 1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

News vs. Sentiment: Predicting Stock Returns from News Stories
Steven L. Heston & Nitish Ranjan Sinha (2017) News vs. Sentiment: Predicting Stock Returns from News Stories, Financial Analysts Journal, 73:3, 67-83, DOI: 10.2469/faj.v73.n3.3

Finance II: Investments Management - Book
Namur, G. (2020) "Finance II :Investments Management". McGraw-Hill Create.

FineNews: fine-grained semantic sentiment analysis on financial microblogs and news
Dridi, Amna & Atzeni, Mattia & Reforgiato Recupero, Diego. (2019). FineNews: fine-grained semantic sentiment analysis on financial microblogs and news. International Journal of Machine Learning and Cybernetics. 10. 10.1007/s13042-018-0805-x.

Stock returns and investor sentiment: textual analysis and social media
McGurk, Z., Nowak, A. & Hall, J.C. Stock returns and investor sentiment: textual analysis and social media. J Econ Finan 44, 458–485 (2020). https://doi.org/10.1007/s12197-019-09494-4

News Impact on Stock Price Return via Sentiment Analysis
Li, Xiaodong & Xie, Haoran & Chen, Li & Wang, Jianping & Deng, Xiaotie. (2014). News Impact on Stock Price Return via Sentiment Analysis. Knowledge-Based Systems. 69. 10.1016/j.knosys.2014.04.022.

Innovation in Financial Services: Balancing Public and Private Interests
Gąsiorkiewicz, L ; Monkiewicz, J. (2020) Innovation in Financial Services. 1st edn. Taylor and Francis. Available at: https://www.perlego.com/book/1812888/innovation-in-financial-services-pdf (Accessed: 25 September 2021).

PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis
Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2509–2514, Lisbon, Portugal. Association for Computational Linguistics.

Forecasting Financial Market Volatility Using a Dynamic Topic Model
Morimoto, T., Kawasaki, Y. Forecasting Financial Market Volatility Using a Dynamic Topic Model. Asia-Pac Financ Markets 24, 149–167 (2017). https://doi.org/10.1007/s10690-017-9228-z

Predicting Returns with Text Data
Ke, Zheng and Kelly, Bryan T. and Xiu, Dacheng, Predicting Returns with Text Data (September 30, 2020). University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-69, Yale ICF Working Paper No. 2019-10, Chicago Booth Research Paper No. 20-37, Available at SSRN: https://ssrn.com/abstract=3389884 or http://dx.doi.org/10.2139/ssrn.3389884

Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction
Nguyen, Thien & Shirai, Kiyoaki. (2015). Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. 1. 10.3115/v1/P15-1131.

The impact of word sense disambiguation on stock price prediction
Hogenboom, Alexander & Brojba-Micu, Alex & Frasincar, Flavius. (2021). The impact of word sense disambiguation on stock price prediction. Expert Systems with Applications. 184. 115568. 10.1016/j.eswa.2021.115568.

Improving selection of synsets from WordNet for domain-specific word sense disambiguation
Lopez-Arevalo, Ivan & Sosa-Sosa, Victor & Rojas-Lopez, Franco & Tello, Edgar. (2016). Improving selection of synsets from WordNet for domain-specific word sense disambiguation. Computer Speech & Language. 41. 10.1016/j.csl.2016.06.003.

Doing safe by doing good: ESG investing and corporate social responsibility in the U.S. and Europe
Bannier, Christina E.; Bofinger, Yannik; Rock, Björn (2019) : Doing safe by doing good: ESG investing and corporate social responsibility in the U.S. and Europe, CFS Working Paper Series, No. 621, Goethe University Frankfurt, Center for Financial Studies (CFS), Frankfurt a. M., https://nbn-resolving.de/urn:nbn:de:hebis:30:3-480587

A Study on Sentiment Analysis Techniques of Twitter Data
Alsaeedi, Abdullah & Khan, Mohammad. (2019). A Study on Sentiment Analysis Techniques of Twitter Data. International Journal of Advanced Computer Science and Applications. 10. 361-374. 10.14569/IJACSA.2019.0100248.

Improved lexicon-based sentiment analysis for social media analytics
Jurek, A., Mulvenna, M.D. & Bi, Y. Improved lexicon-based sentiment analysis for social media analytics. Secur Inform 4, 9 (2015). https://doi.org/10.1186/s13388-015-0024-x

Loughran-McDonald sentiment lexicon
Loughran, T. and McDonald, B. (2011), ``When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.'' The Journal of Finance, 66: 35-65.

Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts
Malo, Pekka & Sinha, Ankur & Takala, Pyry & Korhonen, Pekka & Wallenius, Jyrki. (2014). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. Journal of the American Society for Information Science and Technology. 10.1002/asi.23062.

Kaggle dataset search: Sentiment Analysis
Kaggle.com. 2022. Kaggle dataset search: Sentiment Analysis. [online] Available at: <https://www.kaggle.com/search?q=sentiment+analysis+in%3Adatasets> [Accessed 17 May 2022].

Google BERT
GitHub. 2020. GitHub - google-research/bert: TensorFlow code and pre-trained models for BERT. [online] Available at: <https://github.com/google-research/bert> [Accessed 17 May 2022].

Supervised vs Unsupervised Learning: Algorithms and Examples
Valcheva, S., n.d. Supervised vs Unsupervised Learning: algorithms, example, difference. [online] Intellspot. Available at: <https://www.intellspot.com/unsupervised-vs-supervised-learning/> [Accessed 17 May 2022].

Google news search: Chip Shortage
News.google.com. 2022. Google news search: Chip Shortage. [online] Available at: <https://news.google.com/search?for=chip+shortage&hl=en-GB&gl=GB&ceid=GB%3Aen> [Accessed 17 May 2022].

Latent Dirichlet Allocation
David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

Gaussian mixture models
scikit-learn. n.d. 2.1. Gaussian mixture models. [online] Available at: <https://scikit-learn.org/stable/modules/mixture.html#:~:text=A%20Gaussian%20mixture%20model%20is,Gaussian%20distributions%20with%20unknown%20parameters> [Accessed 17 May 2022].

Formal Concept Analysis
Cimiano, P., 2005. Formal Concept Analysis. [online] Cs.cmu.edu. Available at: <https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume24/cimiano05a-html/node3.html#:~:text=Formal%20Concept%20Analysis%20(FCA)%20is,these%20attributes%20on%20the%20other> [Accessed 17 May 2022].

Topic Modelling: A Deep Dive Into LDA, Hybrid-LDA, And Non-LDA Approaches
Stoy, L., 2021. Topic Modelling: A Deep Dive into LDA, hybrid-LDA, and non-LDA Approaches - LAZARINA STOY. [online] LAZARINA STOY. Available at: <https://lazarinastoy.com/topic-modelling-lda/#3-lda-alternatives> [Accessed 17 May 2022].

Neural Topic Models
Leilan, 2022. GitHub - zll17/Neural_Topic_Models: Implementation of topic models based on neural network approaches.. [online] GitHub. Available at: <https://github.com/zll17/Neural_Topic_Models> [Accessed 17 May 2022].

Evaluate Topic Models: Latent Dirichlet Allocation (LDA)
Kapadia, S., 2019. Evaluate Topic Models: Latent Dirichlet Allocation (LDA). [online] Medium. Available at: <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0> [Accessed 17 May 2022].

lda2vec
Moody, C., 2016. lda2vec. [online] MultiThreaded. Available at: <https://multithreaded.stitchfix.com/blog/2016/05/27/lda2vec/#topic=38&lambda=1&term=> [Accessed 17 May 2022].

Latent Dirichlet allocation assisted time series of financial news sentiments
Rogov, O., Fedorova, E. and Demin, I., 2020. Latent Dirichlet allocation assisted time series of financial news sentiments. [online] Moscow Institute of Physics and Technology, Dolgoprudniy, Russia fintech@gmx.ch 2 Financial University under the Government of the Russian Federation, Moscow ecolena@mail.ru. Available at: <https://events-files-bpm.hse.ru/files/8D54705A-8FE1-4D9E-80CA-E14C04DBB0FE/hse_rogov_lda.pdf> [Accessed 17 May 2022].

Netflix shares fall more than 35% after streamer loses over 200,000 subscribers
Helmore, E., 2022. Netflix shares fall more than 35% after streamer loses over 200,000 subscribers. [online] the Guardian. Available at: <https://www.theguardian.com/media/2022/apr/20/netflix-shares-fall-losing-subscribers> [Accessed 17 May 2022].

A CNN-LSTM-Based Model to Forecast Stock Prices
Wenjie Lu, Jiazheng Li, Yifan Li, Aijun Sun, Jingyang Wang, "A CNN-LSTM-Based Model to Forecast Stock Prices", Complexity, vol. 2020, Article ID 6622927, 10 pages, 2020. https://doi.org/10.1155/2020/6622927

The Impact of Non-Financial Reporting on Stock Markets in Emerging Economies
Filip, Adrian & Spatacean, Ioan & Nistor, Paula. (2012). The Impact of Non-Financial Reporting on Stock Markets in Emerging Economies. Procedia Economics and Finance. 3. 781-785. 10.1016/S2212-5671(12)00230-4.

NewsAPI
Newsapi.org. n.d. Documentation - News API. [online] Available at: <https://newsapi.org/docs> [Accessed 17 May 2022].

New York Times API
Developer.nytimes.com. 2022. New York Times API. [online] Available at: <https://developer.nytimes.com/apis> [Accessed 17 May 2022].

New York Times Dev FAQ
Developer.nytimes.com. 2022. New York Times Dev FAQ. [online] Available at: <https://developer.nytimes.com/faq> [Accessed 17 May 2022].

Getting Started on the Bloomberg Terminal
Data.bloomberglp.com. 2022. Getting started on the Bloomberg Terminal. [online] Available at: <https://data.bloomberglp.com/professional/sites/10/Getting-Started-Guide-for-Students-English.pdf> [Accessed 17 May 2022].

EOD Historical Data – Financial News API
EOD Historical Data. 2022. Stock Price Data, Financial and Stock Market API. [online] Available at: <https://eodhistoricaldata.com/financial-apis/> [Accessed 17 May 2022].

Financial Times API
Developer.ft.com. 2022. FT Developer Programme. [online] Available at: <https://developer.ft.com/portal> [Accessed 17 May 2022].

Yahoo Finance API
RapidAPI. 2022. Yahoo Finance API. [online] Available at: <https://www.yahoofinanceapi.com/> [Accessed 17 May 2022].

How to do SEO for News
Gareth, B., 2021. How to Do SEO for News Websites in 6 Essential Steps. [online] SEO tools & Online Marketing Tips Blog | WebCEO. Available at: <https://www.webceo.com/blog/how-to-do-seo-for-news-websites/> [Accessed 17 May 2022].

spaCy – Linguistic Features
Spacy.io. 2022. spaCy - Linguistic Features. [online] Available at: <https://spacy.io/usage/linguistic-features> [Accessed 17 May 2022].

spaCy – Facts & Figures
Spacy.io. 2022. spaCy – Facts & Figures. [online] Available at: < https://spacy.io/usage/facts-figures> [Accessed 17 May 2022].

displaCy Dependency Visualizer
Explosion - spaCy. 2022. displaCy Dependency Visualizer. [online] Available at: <https://explosion.ai/demos/displacy> [Accessed 17 May 2022].

ProsusAI – FinBERT - Hugginface
Huggingface.co. 2021. ProsusAI - FinBERT. [online] Available at: <https://huggingface.co/ProsusAI/finbert> [Accessed 17 May 2022].

PyTorch
Pytorch.org. 2022. PyTorch Documentation. [online] Available at: <https://pytorch.org/get-started/locally/> [Accessed 17 May 2022].

Google Collaboratory – GPU Architecture
Colab.research.google.com. n.d. Google Colaboratory - GPU Architecture. [online] Available at: <https://colab.research.google.com/github/d2l-ai/d2l-tvm-colab/blob/master/chapter_gpu_schedules/arch.ipynb> [Accessed 17 May 2022].

Gensm - LDA
Rehurek, R., 2022. Gensim: topic modelling for humans. [online] Radimrehurek.com. Available at: <https://radimrehurek.com/gensim/models/ldamodel.html> [Accessed 17 May 2022].

pyLDAvis
Mabey, B., 2015. pyLDAvis — pyLDAvis 2.1.2 documentation. [online] Pyldavis.readthedocs.io. Available at: <https://pyldavis.readthedocs.io/en/latest/readme.html> [Accessed 17 May 2022].

Sklearn – Linear Regression
scikit-learn. n.d. sklearn.linear_model.LinearRegression. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> [Accessed 17 May 2022].

Visa Integrates with Microsoft Wallet
Usa.visa.com. 2022. Visa and Microsoft Wallet. [online] Available at: <https://usa.visa.com/visa-everywhere/innovation/new-way-to-pay.html> [Accessed 17 May 2022].

BlackRock Careers - Data Scientist
BlackRock Careers. 2022. BlackRock Careers - Data Scientist. [online] Available at: <https://careers.blackrock.com/job-search-results/?parent_category=Technology> [Accessed 17 May 2022].

BlackRock Careers – Software Engineers
BlackRock Careers. 2022. BlackRock Careers - Software Engineers. [online] Available at: <https://careers.blackrock.com/job-search-results/?parent_category=Technology&sub_category=Software%20Engineering> [Accessed 17 May 2022].

BlackRock Careers – Technology Implementation
BlackRock Careers. 2022. BlackRock Careers – Technology Implementation. [online] Available at: <https://careers.blackrock.com/job-search-results/?parent_category=Technology&sub_category=Technology%20Implementation> [Accessed 17 May 2022].

How to Make an AI App: A Massive Integration Guide for 2022
Melnik, Y., 2021. How to Create an AI Application: Implementation Use Case. [online] CHI Software. Available at: <https://chisw.com/how-to-build-an-ai-app/#The_Development_Cost> [Accessed 17 May 2022].

Microsoft Azure Pricing Calculator
Microsoft Azure. 2022. Microsoft Azure Pricing Calculator. [online] Available at: <https://azure.microsoft.com/en-gb/pricing/calculator/> [Accessed 17 May 2022].

Cirrascale: GPU Cloud Server and Storage Pricing
Cirrascale. 2022. GPU Cloud Server and Storage Pricing. [online] Available at: <https://cirrascale.com/cirrascale-cloud-pricing.php> [Accessed 17 May 2022].

What does a Financial Analyst Do?
Corporate Finance Institute. n.d. What Does a Financial Analyst Do. [online] Available at: <https://corporatefinanceinstitute.com/resources/careers/jobs/what-does-a-financial-analyst-do-day-in-the-life/> [Accessed 17 May 2022].

2. VIDEOS:

Interview with Justin Sheetz
Breaking into Data by Promotable - YT. 2021. The Quantitative Equity Analyst's Data Science Toolkit w/ Fmr Blackrock VP. [online] Available at: <https://www.youtube.com/watch?v=OpHqPU_2XBw&t=2054s> [Accessed 17 May 2022].

Interview with Stanley Chen
Subtle Asian Investors - YT. 2020. Stanley Chen | CFA | Ex-Blackrock Interview. [online] Available at: <https://www.youtube.com/watch?v=qKqmD-E5TTg> [Accessed 17 May 2022].

BlackRock Lecture Series A Deep Dive Into BlackRock's Technology Ecosystem
KSU College of Computing and Software Engineering - YT. 2021. BlackRock Lecture Series A Deep Dive Into BlackRock's Technology Ecosystem. [online] Available at: <https://www.youtube.com/watch?v=M3nF55IiMnE> [Accessed 17 May 2022].