FinSight: Your Vision for Financial Insights

Seyhan Emre Gorucu, Samaksh Gulati, Sabrina Lin, Suhail Prasathong, Jinsong Zhen, Lakshmi S Gadepalli

I. Introduction

Investing in the stock market has become increasingly complex due to the vast amounts of data available and the various factors that influence stock prices and investor sentiment. Our goal is to assist investors in making informed decisions by building a smart portfolio optimizer web app. The model will assess the investor's risk profile through a survey, and this will be used to classify investors into different risk profiles. The app will then leverage securities data and incorporate social sentiment toward recent news and events to make data-driven recommendations. We will implement our sentimental analysis on financial reports which will act as a filter for the recommendations. Financial reports are widely regarded as a highly representative information source and are frequently employed by credit analysts, accountants, and investors to assess financial performance and determine investment strategies. However, there are many challenges including the complexity and technical language used in these documents, as well as the potential for bias and subjectivity in the analysis.

II. Literature Review

An early example of using big data for sentiment analysis is done by [10] where they create a list of vocabulary where each word is assigned with one of the six sentiments such as positive, negative, strong, weak, etc. However, this approach ignores the fact that the semantics of a word changes depending on the sentence it is used.

Data mining [20] and Machine Learning models can be used for sentiment analysis. Reference [17] recommends deep learning over data mining because the latter has challenges with identifying the features and filtering the best ones while the first learns the features during the training process intrinsically.

Several supervised machine learning techniques have been used for financial sentiment analysis such as support vector machines (SVM) [16], Naive-Bayes [3] or ensemble techniques [2]. Most of these supervised learning techniques typically use the bag-of-words approach. This means they are creating a list of words independent of their order in the sentence and are not able to take advantage of the sequential structure of a sentence.

Even though Convolutional neural networks (CNN) are typically used for image recognition, [17] use CNN's for financial sentiment analysis of StockTwits data, and they find CNN predict sentiment better than not only logistic regression but also recurrent neural networks (RNN) as well.

RNN uses the same neural structure one word after another within the same text. However, this technique suffers from memory loss due to vanishing gradients. A more advanced RNN technique is Long Short-Term Memory (LSTM) proposed by [6]. This method is more complex and has four gates; however, reference [17] still found it inadequate for financial sentiment analysis. Currently, state-of-the-art natural language processing techniques are transformers which are encoder-decoder architectures that contain self-attention, where all words in a text are immediately connected at one step [19].

Typically, the text is converted into vectors before feeding into Transformers as word embeddings [8], [12]. Reference [12] converts words into a vector of numbers by using neural networks. This is very useful as the whole corpus can be expressed in terms of a vector with a fixed size e.g. 512.

Existing tokenizers can be used to convert words into vectors. Then, they are fed into the transformers such as BERT, an encoder-only Transformer [5]; ELMO [14], similar to BERT but with different word embeddings; ULM-Fit [7], same architecture as BERT but with different training data and tuning procedure. FinBERT [4] is a BERT model that is pre-trained on financial documents such as financial filings, news, etc. Just like FinBERT, BERT is also pre-trained on another large dataset, but the dataset is not financially specific. Reference [13] shows that FinBERT is more successful than BERT in domain-specific financial sentiment analysis by comparing the F1-scores of several financial test sets.

Reference [18] successfully implement FinBERT after summarizing the data with BERTSUMEXT [9]. BERTSUMEXT is an NLP technique that summarizes a written text by automatically taking out unnecessary information. With the combination of BERTSUMEXT and FinBERT, they successfully make financial sentiment analyses.

A similar approach to FinBERT is RoBERTa. A team of researchers added extra transformer layers, financial sentiment dictionaries, and financial micro blogs on the pre-trained RoBERTa model to optimize the performance of sentiment analysis of business entities in a financial domain [15].

However, an effective Roberta model needs a massive amount of labels for training. In the financial domain, obtaining a large amount of labeled data is expensive since related data are scarce. Moreover, the model may not work as expected on vague texts and words, such as irony and satire.

III. Innovations

Our product leverages machine learning and natural language processing to analyze vast amounts of financial, social, and news data. The integration of social sentiment and news analysis about specific companies and industries into the portfolio optimization process allows the app to provide investors with a more well-rounded view of a company's potential for growth, risks, and opportunities that might not

be apparent from traditional financial data alone. We have not found an effective visualization tool for sentiment analysis using 10-K and company financial reports. Finally, by tailoring recommendations based on an investor's risk profile, our app provides a personalized and user-friendly experience that meets the needs of individual investors.

Individual investors who are looking to make informed investment decisions will find our approach particularly useful. It can also be of interest to financial advisors who are seeking to provide their clients with personalized and data-driven investment recommendations.

IV. Design and Implementation

We have developed a tool that recommends an investment strategy to a user based on several risk levels. End-to-End User Flow App Design includes the following steps:

1) Data acquisition of 10K/10Q reports by using SEC api and financial news data from Marketwatch and Motley Fool by using Beautiful Soup.
2) Sentiment analysis is carried out with FinBERT for the financial data for 613 companies.
3) 90 Stocks with best sentiment ratings are filtered.
4) The filtered 90 stocks are divided into 5 batches, each batch containing 18 stocks. The grouping of the stocks are made based on their risk level. For example, low risk stocks are used in conservative strategy, high risk stocks are used in aggressive strategy etc. The risk level of a stock can be inquired by using Barra Data Risk Score [1].
5) User fills a google sheet questionnaire to determine one of the five risk levels; conservative, moderately conservative, moderate, moderately aggressive, aggressive.
6) Tableau pulls the investment strategy from the google questionnaire.
7) Efficient frontier rebalancing strategy is used to allocate the percentage of investments for the selected investment strategy. Portfolio management gives a list of stocks and their allocation sizes to minimize risk and maximize returns.
8) Portfolio and individual stocks can be visualized by using Tableau.
9) Publish the results on cloud.

The design of data acquisition, sentiment analysis, portfolio management and visualization can be seen in Figure 1.

V. Data Acquisition

The dataset is made by scraping quarterly (10-Q) and annual (10-K) financial reports over the last 4 years for 613 publicly traded stocks. The reports are critical to organizational financial health and are prepared by domain experts. Here are the steps to extract and pre-process the data:

1. Data Scraping: We used helper functions to automate the process of extracting financial data from SEC filings based on the CIK (Central Index Key) code, filing type, and start date. We use libraries such as BeautifulSoup and urllib to parse HTML and access content from the SEC website.

2. Data Cleaning: We preprocessed the text data by removing unwanted characters and filtering out irrelevant data based on type of information. Several preprocessing steps are performed on the pandas DataFrame, including replacing hyphens and non-breaking spaces, removing null values, and filtering out rows that have less than 10 or more than 1000 words in the paragraph column.

3. Feature Generation: Named Entity Recognition We use the NLP library space to identify and tag entities in the text(named entity recognizer). The recognized entities are further used to filter out prepositional and noun phrases that are present in any of the recognized entities. The output is stored in a separate CSV file for each company. Figure 2 shows a screenshot from a sample CSV file.

Additionally, we collected financial news articles related to the top 90 active U.S. publicly traded companies based on stock data by leveraging MarketWatch, a renowned platform that provides financial information from various publishers, including The Motley Fool and Benzinga. These articles offer crucial market insights by highlighting expert sentiments toward specific stocks:

Data Scraping: We developed multiple scraper functions to extract essential information such as headlines, dates, URLs, and tickers based on the HTML structures of different publishers. We streamlined the process by employing the RegEx module to extract critical text segments and by utilizing URL strings to identify and eliminate duplicate content. To gather financial news authored by investment experts, we used the Beautiful Soup HTML parser to access and collect author names, dates, titles, and article URLs from each webpage.

Scraping Multiple URLs: To focus on accurate and relevant financial news rather than promotional articles, we classified the articles based on second-level domain names and URL suffixes. Initially, we collected all the URLs from the homepage of a ticker on MarketWatch. Subsequently, we used the URL patterns of regular articles published by MarketWatch, The Motley Fool, or Benzinga to create a subset. By excluding extraneous content such as live coverage articles, we efficiently stored the pertinent information in a table for each ticker, thereby saving time and resources.

VI. Sentiment Analysis

FinBERT is used for sentiment analysis. We install the transformers API into our conda environment. Transformers API contains pretrained FinBERT model parameters as well as a tokenizer specifically made for FinBERT, that is developed by [21]. There are other ways to use FinBERT for prediction such as calling an online Inference API from HuggingFace, or downloading the pre-trained model and uploading it to python. We have used the transformers API as it is easy to use and does not need internet connection for making inference. As the model is pre-trained, we only use it to make inference. The results of FinBERT inference are reliable as can be seen at Table I.

The data includes 2.2 Gb of 10-K and 10-Q reports from 613 tickers that were fed into the FinBERT model. The whole

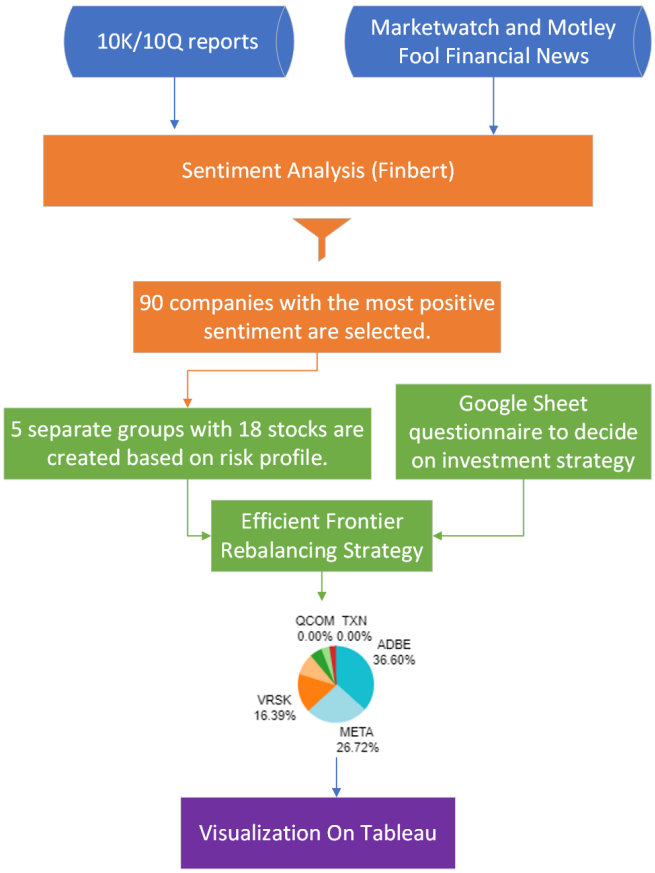

Fig. 1: Portfolio Management

TABLE I: Stocks with the most positive sentiments.

Sentence	Sentiment
Apple is a great stock.	Positive
Apple revenues have increased by 15 %.	Positive
Apple profits declined 30 %.	Negative
The stock market opened at 8 am.	Neutral
Exxon laid off 10 % of staff.	Neutral
Below is the summary of our quarterly report.	Neutral
Tesla increased its cars sales 32 % year on year.	Positive
Verizon profit margins are declining due to competition.	Negative

sentiment analysis was completed in eight days. Each ticker is represented with a CSV file. Each row in a CSV file represents a sentence. We read CSV files as pandas dataframe and then feed every 10 rows into the tokenizer. The whole document cannot be fed into the model all at once because of memory issues. Further, the number of tokens for each row is truncated to 512 tokens as the model does not accept larger size. Tokenized groups of rows are then fed into the model to make prediction. The output for each row is 0, 1 and 2 meaning neutral, positive and negative, respectively.

Likewise, we utilized 1 GB of financial news data from the top 90 companies to analyze the sentiment of their headlines using the FinBERT model. Our objective was to interpret the prevailing sentiment associated with each news article and subsequently quantify the proportion of articles with positive and negative sentiments for each target company.

The processed data is tabulated into a csv file as the number of rows, positive rows and negative rows. Table II shows the company name, ticker symbol, total number of rows, total number of positive and negative rows and net positive percentage for the 20 best companies. Net positive percentage equals (Positive-Negative)/Total*100.

VII. Investment Strategy Questionnaire

In this section, we couple Tableau public with a google questionnaire to determine the proper investment strategy for the person. Please click on this **link** to access the google questionnaire. We have made this questionnaire based on Lincoln Financial Group's retirement **questionnaire**. Based on the user's answers, we determine an investment strategy. For example, it wouldn't be very wise to recommend a retired elderly person to have an aggressive investment strategy.

1) Question 1: What is your age?
   a) more than 70 (1 point)
   b) 60 to 69 (3 points)
   c) 46 to 59 (7 points)
   d) less than 46 (10 points)
2) Question 2: I plan to withdraw money from my retirement plan account in:
   a) less than 5 years (1 point)

paragraph	value	label	type	company	entity_type_ext
On November 17, 2021 we declared a quarterly dividend of $0.21 per share of common stock, or approximately $63 million which will be paid on January 26, 2022 to shareholders of record as of the close of business on January 4, 2022. The timing and amounts of any future dividends are subject to determination and approval by our board of directors.	4-Jan-22	['Dividends date of record']	dateItemTyp	Apple	DATE
and bear interest at a fixed rate of 3.05% per annum. The interest is payable semi annually on March 22nd and September 22nd of each year and payments commenced March 22, 2017.	3.05	['Fixed interest rate per annum']	percentItem	Apple	PERCENT
On January 21, 2021, we redeemed $100 million of	300	['Repayments of senior debt']	monetaryIte	Apple	MONEY
Share Based Compensation. For the years ended 20	72	['Share based compensation expense']	monetaryIte	Apple	MONEY

Fig. 2: Snapshot of the data

TABLE II: Stocks with the most positive sentiments.

Company	Ticker	Total	Positive	Negative	Net positive (%)
Idacorp	ida	907	57	0	6.3%
Edwards Lifesciences	ew	1134	81	11	6.2%
Fortinet	ftnt	1476	95	5	6.1%
Ansys	anss	1057	63	0	6%
Amer Natl Insurance	anat	1088	55	0	5.1%
Motorola Solutions	msi	3371	199	38	4.8%
Navient Corp	navi	1187	60	6	4.5%
Siteone Landscape	site	2063	93	0	4.5%
Skyworks Solutions	swks	808	43	7	4.5%
Costar Group	csgp	1356	66	6	4.4%
Mdu Resources	mdu	1830	89	22	3.7%
Walgreens Boots	wba	1905	89	20	3.6%
Donaldson Company	dci	1115	40	0	3.6%
Eli Lilly And Co	lly	1635	83	27	3.4%
Mercury Systems	mrcy	1762	60	1	3.3%
Qualys Inc	qlys	963	32	1	3.2%
Intuitive Surgical	isrg	1122	41	6	3.1%
Essential Utilities	wtrg	5478	179	10	3.1%
PepsiCo	pep	1497	54	8	3.1%
Neurocrine Biosciences	nbix	1402	54	11	3.1%

b) 6 to 9 years (3 points)
c) 10 to 15 years (6 points)
d) more than 15 years (8 points)

3) Question 3: I should have enough savings and stable/guaranteed income (that is, Social Security, pension, retirement plan, annuities) to maintain my planned standard of living in retirement:
   a) Not confident (1 point)
   b) Somewhat confident (2 points)
   c) Confident (4 points)
   d) Very confident (6 points)
4) Question 4: The following statement best describes my willingness to take risk:
   a) I am more concerned with avoiding losses in my account value than with experiencing growth. (1 point)
   b) I desire growth of my account value, but I am more concerned with avoiding losses. (3 points)
   c) I am concerned with avoiding losses, but this is outweighed by my desire to achieve growth. (5 points)
   d) To maximize the chance of experiencing high growth, I'm willing to accept losses. (7 points)
5) Question 5: If I invested $100,000 and my portfolio value decreased to $70,000 in just a few months, I would:
   a) be very concerned and sell my investments. (1 point)
   b) be somewhat concerned and consider allocating to lower risk investments. (2 points)
   c) be unconcerned about the temporary fluctuations in my returns. (4 points)
   d) invest more in my current portfolio. (5 points)
6) Question 6: My assets (excluding home and car) are invested in:
   a) I don't know how my assets are invested. (1 point)

b) My pension, certificates of deposit (CDs), annuities, IRA, and savings accounts. (2 points)
c) A mix of stocks and bonds, including mutual funds. (3 points)
d) Stocks or stock mutual funds. (4 points)

Based on the user's answer, the corresponding points are summed and one of the following investment approaches is selected; conservative (0-12 points), moderately conservative (13-20 points), moderate (21-28 points), moderately aggressive (29-34 points), aggressive 35-40 points). The questionnaire recommends one of the following options; retake the quiz, or proceed to visualization. The "proceed to visualization" selection opens a new tab which is also hosted by Tableau. However, before visualization, some back-end calculations are carried out in order to use an efficient frontier rebalancing strategy to allocate the stocks.

VIII. Portfolio Management

Thus far, we have prepared a vetted list of securities, based on sentiment analysis, consisting of 90 stocks. The 90 stocks are further divided into five batches based on their risk factor. The riskiest 18 stocks are used in the aggressive investment strategy while the safest 18 stocks are used in the conservative investment strategy etc. The risk level of a security can be found from Barra data risk score [1]. One of our members has access to a paid **data source** that provides the Barra data risk score. Therefore, we were able to reorganize the 90 stocks into 5 batches based on their risk level. The rest of this section outlines the efficient frontier rebalancing strategy that is used to optimize the allocation weights of the final filtered 18 stocks.

The key balancing strategy being applied in our scenario is the Efficient Frontier Rebalancing Strategy which also is the cornerstone of the Modern Portfolio Theory developed by Harry Markowitz [11].

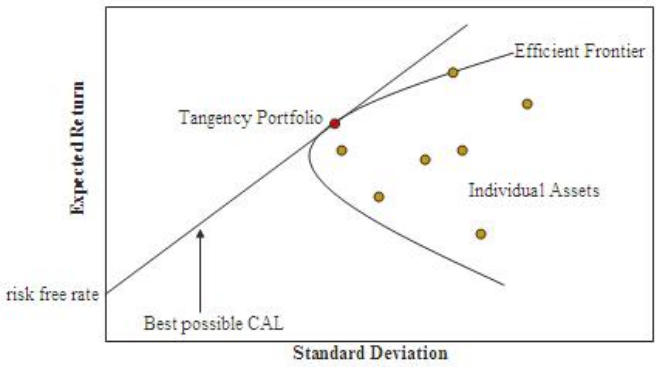

Fig. 3: Portfolio Management - Efficient Frontier Chart

The Efficient Frontier Strategy calculates the optimal combination to maximize rewards while minimizing risk. This is done so by attaining diversification and having realistic returns goals. Figure 3 describes how the optimal portfolio is attained by finding the best fit when plotting various portfolio range outcomes.

In theory, the Efficient Frontier leverages the compound annual growth rate (CAGR) of an investment as the return component while using annualized standard deviation as the risk metric.

$$Er(p) = w_a * Er(A) + w_b * Er(B) \tag{1}$$

$$SD_P = (w_A * SD_A)^2 + (w_B * SD_B)^2 + 2 * w_A * w_B * SD_A * SD_B * cor(A, B) \tag{2}$$

For instance, for a two asset portfolio, the risk and return is calculated as described in Eq. 1. The Efficient Frontier plots the risk and returns described in Eq. 2 represent maximized returns with assumed risk for n number of assets [11]. The goal here for an investor is to fill a portfolio with securities offering better returns with a combined standard deviation lower than that of the standard deviations of individual securities.

Our implementation is triggered by the strategy decision from the google questionnaire. By using python and several api's, we pull the 18 companies' 1 year stock data on the back-end, and carry out standard deviation and Sharpe ratio calculations. Sharpe ratio gives a portfolio's return divided by the standard deviation of its historical performance. Therefore, a higher Sharpe ratio allows higher returns with minimal instability. We vary the allocation of the companies such that they sum up to 1 and carry out 15000 simulations -with varying allocations- that compute the Sharpe ratio. We also use Scipy's optimizer to minimize the negative of the Sharpe ratio as a function of the allocations. Out of all of these calculations, we pick the allocation that gives the highest Sharpe ratio. We then recommend the user an optimized security allocation strategy based on their investment strategy. The user can see the details of the recommended strategy in the visualization tab.

IX. Visualization

Visualization starts with a Google questionnaire embedded in Tableau where the survey automatically calculates user's Risk score based on their selections.

Once the questionnaire is completed, the user is assigned an investment strategy based on their risk category, and back-end calculations are carried out to optimize the percentages of 18 stocks. The user has the option to retake the quiz or proceed to the dashboard. The dashboard shows 2 sections - the user's portfolio summary and market summary. Under the portfolio summary, user can view the stocks in their portfolio, their respective percentages, and a visualization showing the Monte Carlo simulations run in the back end as shown in figure 4. In the market trends section, users can see how various tickers are performing in the market (as seen in Figure 8) and the volume being traded. Users can also see the positive and negative sentiment counts for each ticker in the portfolio (as seen in Figure 6. Figure 7 shows the scatter plot mapping counts of positive vs. negative sentiments for all tickers. These tickers are also clustered using Tableau's clustering feature. We selected 4 clusters to show the quadrants where the sentiments lie. Figure 9 shows the trade volume of the selected tickers.

However, it is possible to select the desired stocks on the dropdown list on the right.

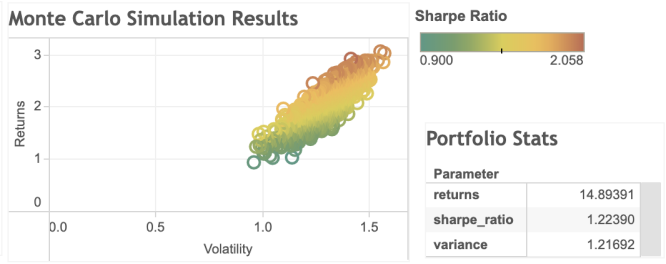

Fig. 4: Monte Carlo Simulation

Fig. 5: Recommended portfolio shown as a list.

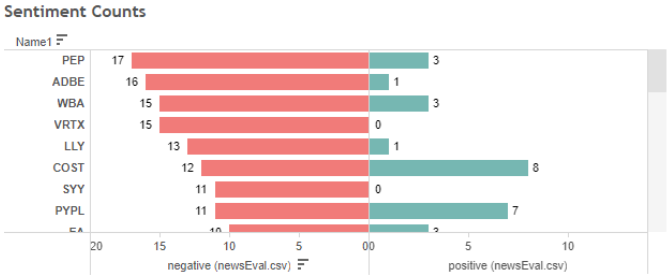

Fig. 6: Positive and negative sentiment counts for each ticker in the portfolio.

X. Conclusions and Discussion

We have built an end-to-end cloud based app that uses a pre-trained nlp api to estimate sentiment of 613 tickers. We have picked the best 90 tickers for further analysis. We have distributed these stocks into five different investment strategies. Further, we optimized portfolio allocation by using efficient frontier rebalancing strategy. Finally, we visualize our optimum portfolio by using Tableau enabled website. In the future, we recommend comparing our results with other investment strategies such as S&P 500 index both in terms of return and risk.

All team members have contributed a similar amount of effort to the project.

References

[1] Barra global equity model, investopedia.

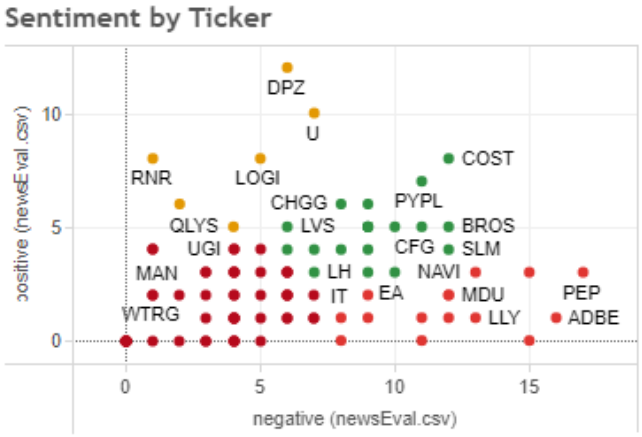

Fig. 7: A two-dimensional plot that shows the negative and positive sentiments for several stocks.

Fig. 8: Aggregate stock performance of the portfolio.

[2] Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 540–546, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[3] Ghaith Alkubaisi, Siti Kamaruddin, and Husniza Husni. Stock market classification model using sentiment analysis on twitter based on hybrid naive bayes classifiers. *Computer and Information Science*, 11:52, 01 2018.

[4] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.

[9] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders, 2019.

[10] Tim Loughran and Bill Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66:35 – 65, 02 2011.

[11] Harry Markowitz. Portfolio Selection. *Journal of Finance*, 7(1):77–91, March 1952.

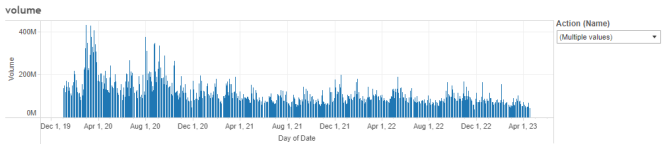

Fig. 9: Trade volume of the selected tickers.

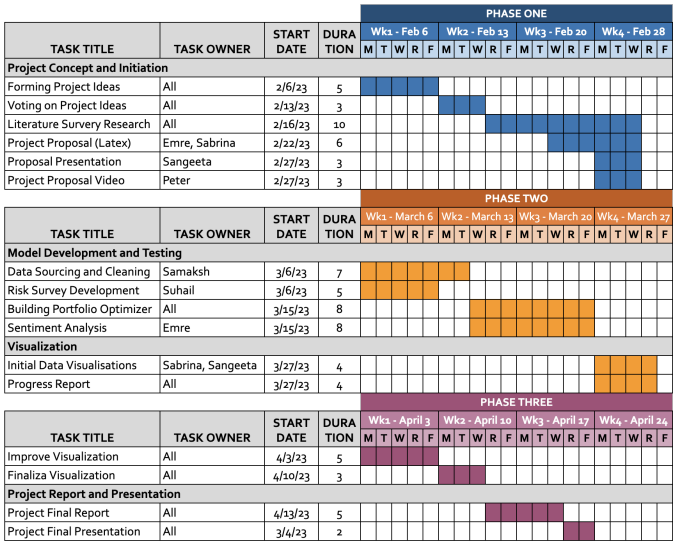

Fig. 10: Gantt Chart

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[13] Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. Is domain adaptation worth your investment? comparing BERT and FinBERT on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[14] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[15] E. Pontes and M. Benjannet. Contextual sentence analysis for the sentiment prediction on financial data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4570–4577, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[16] Rui Ren, Desheng Dash Wu, and Tianxiang Liu. Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, 13(1):760–770, 2019.

[17] Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M. Khoshgoftaar. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5, 2018.

[18] Gim Soong and Tan Chye Cheah. Sentiment analysis on 10-k financial reports using machine learning approaches. pages 124–129, 11 2021.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[20] Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y. Zhao. Crowds on wall street: Extracting value from social investing platforms, 2014.

[21] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020.