EFFECTIVE USAGE OF UNSTRUCTURED DATA USING NATURAL LANGUAGE PROCESSING IN THE FINANCIAL INDUSTRY

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Soukarya Ghosh

March 25, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

ADVANCEMENTS OF ADVERSARIAL NATURAL LANGUAGE PROCESSING AND ITS APPLICATIONS TO UNSTRUCTURED FINANCIAL DATA

Natural language processing (NLP) is a subfield of machine learning and artificial intelligence that deals with the interaction between computers and humans using natural human languages. The goal of NLP is to read, decipher, understand, and piece together natural language in a manner that is more valuable in terms of quantitative analysis and decision making (Belohlavek, Platek, Straka, 2015). In order to further the field of NLP through the reduction of friction experienced by newcomers, a platform was designed, built, and deployed in the form of a web application to aid the study of adversarial examples by allowing ease of access to a robust library of adversarial NLP attacks through an approachable user interface.

With research and applications expanding in the field, NLP has caught the attention of many established industries that deal with qualitative data and content with little to no structure (Huq, Pervin, 2020). As the parsing and extraction of knowledge from qualitative data can be costly to companies, NLP provides a leverage to expedite the process. However, the complexity of the data set specific to the financial industry has deterred progress from being made in the sector, with misinformation and noise from social media and misunderstandings within company filings and financial literature being a few key problems. This is where the technical project can have an indirect impact by helping potential researchers gain interest or insights into gaining leverage on the problems. An approachable user interface has been crafted to make adversarial NLP frameworks more accessible to the public, such that researchers and other interested parties can conduct studies with a lowered barrier to entry.

This STS paper will explore the relationship between the institutional traders, which include firms and hedge funds, retail or individual investors, academic researchers, whose goal is to advance NLP theoretics, and the engineers or analysts who implement such technologies in

the real world markets. Furthermore, the paper will inspect the social impact each group has on the development of NLP technologies with respect to the financial markets through the lens of Social Construction of Technology (SCOT) (Douglas, 2012) framework.

MISUSE OF QUALITATIVE DATA IN FINANCE

Quantitative research is one of the highest paying entry level fields in the entire world, offering up senior manager level salary for the ability to conduct numerical based research on financial and stock market data (Malinsky, 2020). In focusing on the quantitative analysis, many firms and researchers forego the majority of the data produced daily within the markets: qualitative data. As shown in Figure 1, this makes up roughly 80% of the data produced daily through financial statements, earnings reports, and much more (Boucher, 2020).

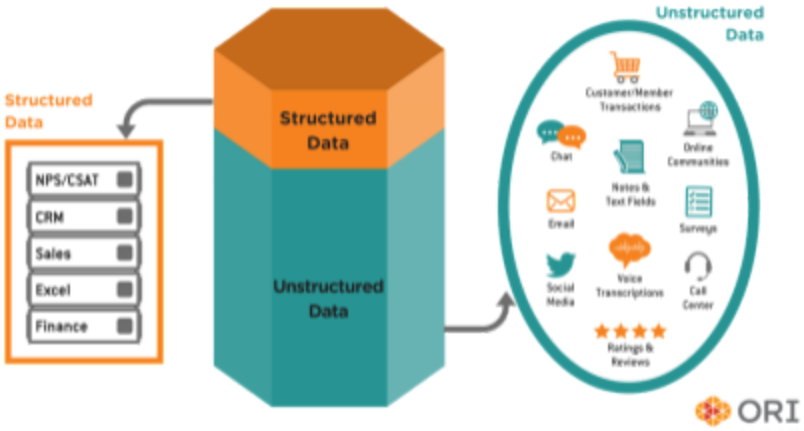

Figure 1. Ratio of structured (quantitative) to unstructured (qualitative) data in the financial industry. Accompanied by examples of each (Boucher, 2020).

Contributions by Social Media and the Retail Investor

Additionally, a massive portion of market movers are made up of individual or retail investors, especially with the advent of Robinhood and other easily accessible trading platforms. This recent increase in retail trading leads to the general population having a significant impact on the markets, with nearly 25% ownership (Winck, 2020). In the past, individual participation was gate-kept by hedge funds and money managers who prevented the public from having

information on trades and coordinating pump and dumps, which is the idea of artificially increasing the value of an asset in order to make a profit before it inevitably retreats to its perfect competition valuation. The information on such events are now readily available across the web, especially on social media platforms such as Reddit, Twitter, and other peer-to-peer forum based networks (Antweiler, Frank, 2004). At peak hours, there is an average of half a million active users spanning across Twitter and Reddit discussing the financial markets, generating an insurmountable amount of unstructured data for those willing to tackle it (Ghosh, 2021). As shown in Figure 1 above, social media inhabits merely a single region within the gargantuan category of unstructured data. Unstructured quantitative data like this usually go unprocessed, leading to the loss of large insight that would otherwise be trailblazing for the space.

In addition to this, most unstructured financial data is processed by hand (Beattie, McInnes, Fearnley, 2004). The manual labor of crunching a large amount of data requires a significant amount of bandwidth and time, especially with respect to crunching financial reports and quarterly earnings. The lag in time caused by the aforementioned efforts leads to inefficiencies and discrepancies between the prediction models created through the process and the blunt reality of the markets.

UNCERTAINTY AND LACK OF STRUCTURE IN FINANCIAL TEXTS

Financial data comes in all forms and as discussed above, a significant majority of it is in a form that does not bode well to quick analysis. These include texts such as company financial statements, quarterly earnings reports, mergers and acquisitions, and public opinion publications. While traded companies make this entire dataset publicly available for both firms and retail traders, the resources required to research and understand the entirety of it makes it essentially

unavailable to individuals and retail traders. The inability to utilize the data causes a gap in knowledge which has led to the decline of public opinion on money management firms and hedge funds (Ekins, 2017). This shift in recent years has had interesting effects on the market, especially with increased democratization of trading with introduction of software platforms like Robinhood.

Rallying Against Wall Street

These effects have mostly been in the form of increased market volatility, with the most recent boom being led by assets with no fundamental foundation of its high valuation namely Gamestop and AMC. These stocks saw an increase of over 1800% over the first quarter of 2021, with the main reason being a forum on Reddit called Wall Street Bets (WSB) pushing the overbought asset to be taken to new heights. The publicly advertised motivation for this was to lead hedge funds, who were shorting these assets, to bankruptcy (Phillips, Lorenz, 2021). Shorting is the act of betting against an asset, gaining profits only if said asset decreases in value. To squeeze this short for its entire value and cause the hedge funds to pay billions of dollars in losses, WSB led the charge for a historical run of the stock.

The unprecedented rally was met with authority from the government and the media, who put trading restrictions on specific assets, not allowing the public to sell or buy shares for certain periods of time. Additionally, individual traders were countered by traditional news media outlets who drove fear into the public to sell, this is an idea that has been studied and proven to be effective before (Smales, 2014). One can debate this topic through the breakdown of the actors involved in this situation. One such actor is the government, who, we can argue, prioritizes the ethics of utilitarianism, wanting to maximize the overall good effect on the nation and the stability of its economy. Allowing such chaos to ensue leaves the economy is a vulnerable state

to recede into a crash when entire industries are forced to file for bankruptcy. The media can also be tagged onto this side, with the retails and individual traders being the opposition actors. While the ethics of this group are far more questionable, a case for rights ethics can be made with a focus on closing the wealth gap between the one percenters in the Wall Street hedge funds and the rest of the nation, citing it is their constitutional right to a free market.

Abundance of Seemingly Unworkable Insights

Everything mentioned above, starting from the initial posts on WSB, to the articles and mandates posted subsequently by the media and government agencies, were all in the form of unstructured qualitative data. These natural language based pieces do not carry any quantitative metrics, but do have sentiment and emotional values which can be gauged using machine learning techniques. Various machine learning algorithms have been developed to keep up with social media sentiments towards the financial markets, with one of the largest archives being StockTwits (Grimes, Plana, 2018). However, game theoretics apply here, as the posters on this forum are aware that their sentiment and posts are being used against them to gauge market sentiment, which in turn creates a certain percentage of posts which are fabricated to throw off bots and models which automate the reading process. As shown

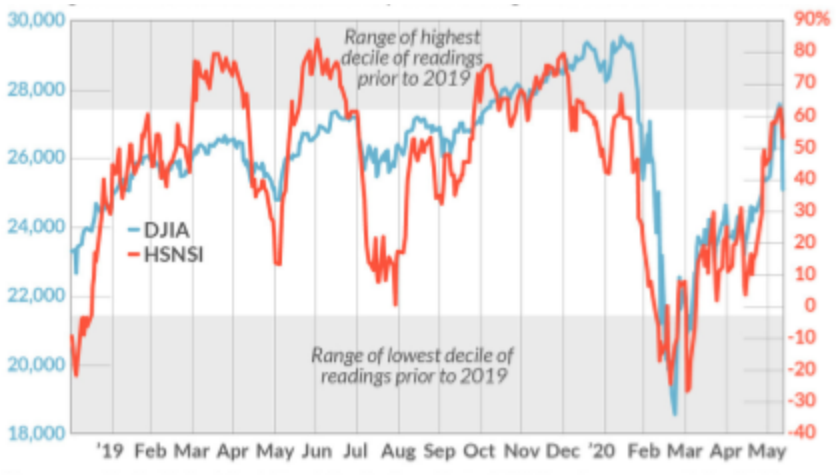

Figure 2. Value vs. Time graph of the Dow Jones Index (DJIA) and stock newsletter sentiment (HSNSI). Shows correlation between excessive bullish sentiment and corresponding crash in index value (Hulbert, 2020).

in Figure 2, with the red line indicating the Dow Jones Index and the blue line showing stock newsletter sentiment index, there are certain correlational attributes between the two. The primary one being that excessive positivity, referred to as bullishness, leads to sharp drops in the value of the index, as noticed in April 2019, July 2019, and February 2020 (Hulbert, 2020).

Not only is unstructured data difficult to crunch, it can also be unreliable without context, especially when drawing obvious conclusions, such as positive sentiment for a product means increased demand, and thus, increased company valuation. Herein lies the second most pressing issue, behind speed to decipher, of unstructured data in finance: its reliability. It is hard to get the truth out of a stream of data that is a mix of actual public opinion, contaminated by fake posts and automated reports (Antweiler, Frank, 2004). The combating process of this can potentially include automated fact checking and moderation on forum based medias and outlets. However, a study done by Reuters shows the consistent ineffectiveness of these methods as a massive 59% of false posts remain active weeks after initially being posted on Twitter. (Brennen, Simon, Howard, Nielsen, 2020).

Figure 3. Shows percentage of posts in sampled data rated as false that were still active without any fact checked flags or outright removal (Brennen, Simon, Howard, Nielsen, 2020).

More traditional forms of unstructured qualitative data include annual reports for companies, which are often written haphazardly under time pressure from corporations and investors, leading to misunderstanding and poor investment choices being made (Li, 2016). While these reports do use easily extractable quantitative data, there are often sentiments written between the lines that are foregone by most firms. Furthermore, the largest information gain that is relevant to fundamental market movements comes directly from the reports and meetings posted by the federal reserve summarizing the state of the economy (Boukus, Rosenberg, 2006). However, these more conventional sources have similar fallthrough of sluggishness and unreliability of exaggerated sentiments.

HOW CAN THE PROCESS OF QUALITATIVE ANALYSIS BE STREAMLINED?

With the overflow of unstructured qualitative data of questionable validity populating the financial space and a lack of developed and adopted technology to fully take advantage of it, questions of introducing the relatively risky machine learning frameworks becomes an increased reality. Although most risk averse industries tend to avoid usage of machine learning when concerned with revenue impacting ventures, financial firms and retail investors have slowly become increasingly open to accepting these non deterministic technologies into their portfolio determiner's core logic (Schumaker, Chen, 2009).

Although there already exists NLP algorithms within this space, the use is not nearly as prevalent as it could be, with only 45% of financial service firms investing heavily into the technology and considering it core logic to their business (Columbus, 2019). The other 55% air concern of introducing machine learning to their practice, citing the infancy of the technology as a potential concern. In the chase for maximizing profits in the space, there are some ethical

dilemmas that arise, namely risking life savings based on the advice of non deterministic and sentient algorithms that change with streaming data and are also vulnerable to adversarial attacks with the introduction of fake data. In their study of fraudulent data within finance and ways of avoiding its usage with machine learning, Moffit and Burns proposes the ethical dilemma of the employing such technologies that are victim to misunderstanding some natural language in the event of a large scale financial collapse or failure (Moffitt, Liao, Yang, Chang, Luo, 2010). Is there a party directly responsible if the actions were triggered by an automated system built to understand human emotions and sentiment? It is possible to point fingers to the firms who employ the technology, however, it is equally possible to blame the academics and engineers who research and build the technology with the intent of understanding such complexities.

Regardless of such dilemmas, machine learning, and specifically NLP, continues to evolve and grow at unprecedented rates. As shown in Figure 4, its uptrend in the past few decades, leading to total domination within the realm of text analysis, is evidence of positive developments and potential taming of its non deterministic behaviors. The rise in usage is a direct result of the demand in other industries, such as entertainment and tech, for such technologies, where non determinism is not as impactful. Increased social demand directs NLP's development by incentivizing investment into NLP research by institutions and academia,

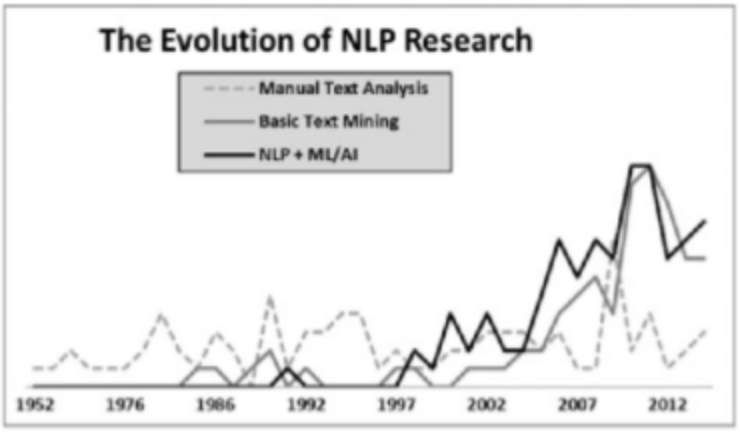

Figure 4. The Evolution of Analysis of Unstructured Content. Shows the rise of machine learning usage to analyze and understand text as opposed to manual analysis techniques. (Fisher, Garnsey, Hughes, 2016).

which in turn drives implementation of researched algorithms by engineers. The impact of this incentive loop is what is pictured in Figure 4, as the early 2000s tech boom acted as a catalyst for the loop to catch fire initially.

CONVERSION OF METRICS USING NATURAL LANGUAGE PROCESSING

Thus, considering all of the complications discussed above with respect to the usage of unstructured qualitative data in finance, let us consider introducing NLP to this industry in a new light. As of now, NLP is not considered part of the business logic for a majority of the industry and a good portion of firms do not even consider using machine learning when dealing with qualitative data. NLP steps in to bridge the gap between quantitative and qualitative sources of information, by converting qualitative data into numerical metrics that can feed into a pipeline for assisted decision making. Various forms of NLP algorithms exist to convert unstructured qualitative data into comparable quantitative data, most of which are currently not employed within the financial markets (Columbus, 2019). The most popular of these approaches is sentiment analysis, which seeks to assign a numerical positive or negative sentiment score to a sentence or statement. There is also topic extraction, which, as shown in Figure 5, can group sections of natural language into sectors to be analyzed separately.

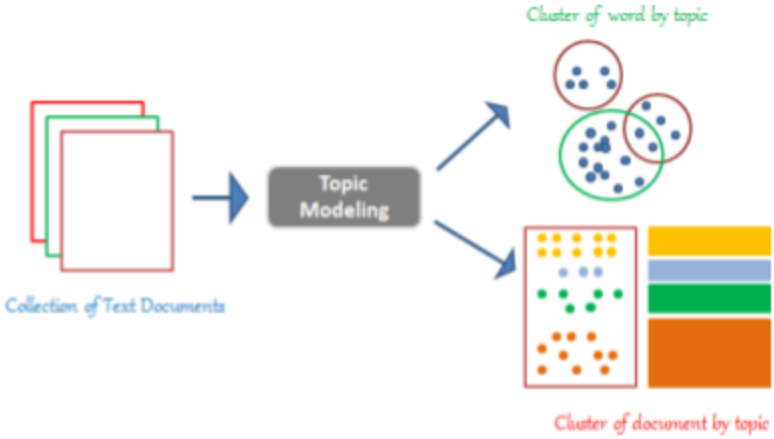

Figure 5. Topic extraction algorithm visualization in natural language processing. This algorithmic method allows for the separation of discussion by subject areas and connection of similar areas across different documents (Yiu, 2019).

These two can be combined to get a powerful insight into sentiments and emotions towards distinct entities.

Furthermore, there also exists relationship extraction and similarity analysis, which can draw conclusions of how two different statements are related based on prior knowledge and training data. This can allow the model to get a better understanding of how the market is connected and which assets often move together as complementary or substitute goods.

From the point of view of generating financial data, there currently exists unimplemented, but heavily researched, NLP technologies which allow for the generation of financial reports (Andersen, Hayes, Huettner, Schmandt, Nirenburg, Weinstein, 1992). Thus, one path to making unstructured qualitative data more streamlined and faster to decipher is to let the technology itself, NLP, generate the financial reports which can be easily reverse engineered for the underlying meaning without having to understand the subtext. However, from an ethical standpoint with respect to duty ethics, which demands that we respect an individual's rational autonomy, letting an algorithm write these reports subjects the entire population to the use of the non deterministic technologies, given that the source material is a product of it. What makes the use of NLP arguably morally sound from an analysis perspective is that it does not force the effects of the fall throughs of machine learning upon an entire population. This is given that the technology is not used blindly and not the sole decision maker within a firm's core logic.

All methods discussed above can be applied to all of the aforementioned sources discussed previously, which include everything from social media posts to official quarterly earnings reports and federal reserve reports on the state of the economy. The most cost effective of the methods will aim to convert the unstructured qualitative texts into quantitative heuristics that can be used to filter and parse massive amounts of data (Schumaker, Chen, 2009). In doing

so, both issues considered previously, time and reliability, can be alleviated, as fabricated data can be grouped together and ignored entirely and analysis can be done within a matter of minutes. In any case, the current arguments against NLP deals with the reliability and determinism of the technology itself. Since this technology is the product of social construction by the hands of those in finance, technology, academia, and to some extent, the public, by employing this technology, one is automatically helping alleviate the problem. Although not in its ideal or most powerful state right now, the increasing reliability of NLP, as exemplified by Figure 4, is directly a result of its increased use in recent years. Therefore, as more firms are convinced to the idea of using NLP, the better and less error prone the technology will inevitably become.

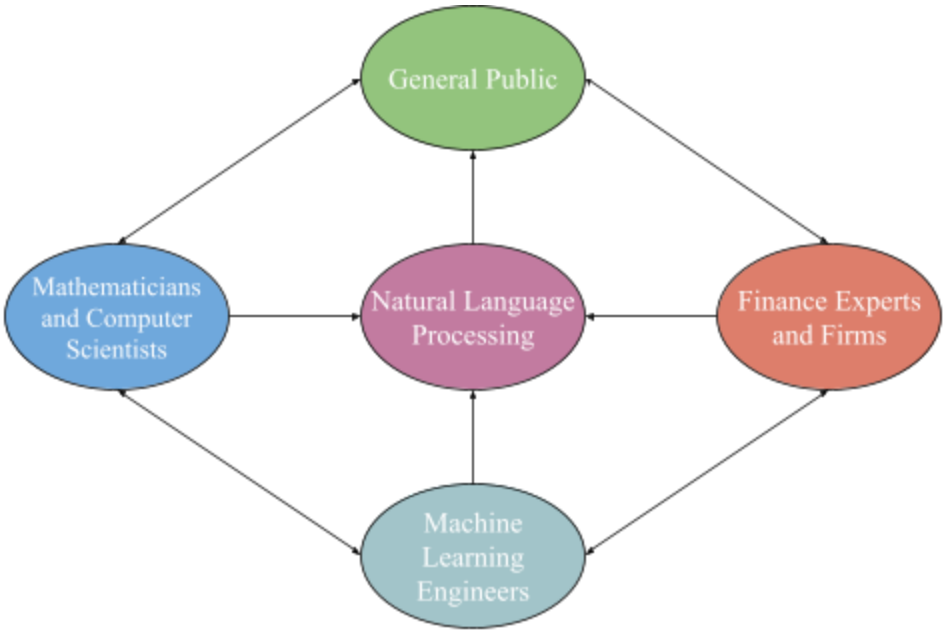

Figure 4. Natural Language Processing SCOT Model. Machine learning engineerings, academic researchers, and finance experts work together to produce the natural language processing algorithms and application, specialized towards applications of understanding unstructured data, which works to benefit the general public's wealth. The public voices approval by deciding to participate in the financial markets through firms that utilize this technology. This feedback fuels the continued research of new and innovative NLP methods (Ghosh, 2020).

Thus, as shown in Figure 6, we have that this incentive loop, which is already in motion, constructs NLP to become a more effective means to decipher unstructured data with speed. Furthermore, with advancements being made, the barrier to entry to use this technology and the required resources to fully utilize it is bound to get lower. In fact, there already exists NLP platforms that the average consumer can simply load data into to get some amount of preprocessed output, such as topic extraction and sentiment analysis (Amazon Comprehend, 2021). Software, such as AWS Comprehend, helps to close the gap discussed previously between the resource heavy institutions and the individual investors. Therefore, this social construction of NLP within finance expands to include other industries as well, who now have a vested interest and a potential profit avenue by leveraging this need.

The SCOT framework views the advancements of NLP as the summation of the conglomerate that interacts with it directly, engineers and the financial world, and those that are indirectly affected by it, the general public and retail traders. It highlights the circular nature of the incentives within this space, where each group is incentivized to help improve NLP by those whom they will affect and by those who preceded. Although there exists intergroup tensions, with individual traders revolting against institutional traders and authoritative entities stepping in to diffuse the apprehension; from a macro lens, the entire process feeds into the incentive cycle that pushes NLP further towards trustworthiness. The effect of this is a unified future where information gaps can be shortened and a larger chunk of available data can be properly utilized. To further this study and speed up the development process, increased educational ventures about machine learning, such as the platform built for my technical project, and its utilization can lead to higher adoption rates both by institutions and potential researchers.

WORKS CITED

Amazon Comprehend. (2021). Amazon Comprehend. *Amazon AWS*. doi: https://aws.amazon.com/comprehend/

Andersen, P., Hayes, P., Huettner, A., Schmandt, L., Nirenburg, I., Weinstein, S. (1992). Automatic extraction of facts from press releases to generate news stories. *Paper read at Third Conference on Applied Natural Language Processing.*

Antweiler, W., Frank, M.. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance.* doi: https://www.jstor.org/stable/3694736.

Beattie, V., McInnes, W., Fearnley, S. (2004). A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. *Accounting Forum.* doi: https://www.sciencedirect.com/science/article/abs/pii/S0155998204000390.

Boucher, L., (2020). What's Hiding in Your Unstructured Data? *ORI News*. https://www.oriresults.com/articles/blog-posts/whats-hiding-in-your-unstructured-data/

Boukus. E., Rosenberg, JV. (2006). The information content of FOMC minutes. *Federal Reserve Bank of New York: New York*. https://www.newyorkfed.org/medialibrary/media/research/economists/rosenberg/Bouku_and_Rosenberg_072006.pdf.

Brennen, S., Simon, F., Howard, P., Nielsen, R. (2020). Types, sources, and claims of COVID-19 misinformation. *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

Columbus, L. (2019). Why AI Is The Future of Financial Services. *Forbes*. https://www.forbes.com/sites/louiscolumbus/2019/08/15/why-ai-is-the-future-of-financial-services/?sh=11fb7d953847

Douglas, D. (2012). The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. *The MIT Press*. http://www.jstor.org/stable/j.ctt5vjrsq

Ekins, E. (2017). Wall Street vs. The Regulators: Public Attitudes on Banks, Financial Regulation, Consumer Finance, and the Federal Reserve. *Cato Institute*. https://www.cato.org/survey-reports/wall-street-vs-regulators-public-attitudes-banks-financial-regulation-consumer

Fisher, I., Garnsey, M., Hughes, M. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. xuebalib.com.5838.pdf

Ghosh, Sh. (2021). Reddit group WallStreetBets hits 6 million users overnight after a wild

week of trading antics. *Insider*. https://www.businessinsider.com/wallstreetbets-fastest-growing-subreddit-hits-58-million-users-2021-1.

Ghosh, S. (2021). *Natural Language Processing SCOT Model*. [Figure 6]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Grimes, G., Plana, A. (2018). How StockTwits Applies Social and Sentiment Data Science. *KD Nuggets*. https://www.kdnuggets.com/2018/03/stocktwits-social-sentiment-data-science.

Hulbert, M. (2020). The real reason for the stock market's 7% plunge shouldn't surprise you — and it happens every time. *Market Watch*. https://www.marketwatch.com/story/the-real-reason-for-the-stock-markets-7-plunge-shouldnt-surprise-you-and-it-happens-every-time 2020-06-11.

Li, F. (2016). Do Stock Market Investors Understand the Risk Sentiment of Corporate AnnualReports? *University of Michigan*. doi: http://dx.doi.org/10.2139/ssrn.898181.

Malinsky, F. (2020). The top 10 highest-paying entry level jobs. *Grow*. https://grow.acorns.com/highest-paying-entry-level-jobs/.

Moffitt, K., Burns, M. (2011). Using lexical bundles to discriminate between fraudulent and non-fraudulent financial reports. *Rutger University*. doi: http://raw.rutgers.edu/docs/seminars/spring11/Moffitt.pdf

Phillips, P., Lorenz, T. (2021). 'Dumb Money' Is on GameStop, and It's Beating Wall Street at Its Own Game. *New York Times*. https://www.nytimes.com/2021/01/27/business/gamestop-wall-street-bets.html.

Schumaker, R., Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: AZFinText system. *ACM Transactions on Information Systems.* doi: https://dl.acm.org/doi/10.1145/1462198.1462204

Smales, L. (2014). News sentiment and the investor fear gauge. *Finance Research Letters*. https://www.sciencedirect.com/science/article/pii/S1544612313000354.

Winck, B. (2020). Retail traders make up nearly 25% of the stock market following COVID-driven volatility, Citadel Securities says. *Markets Insider*. https://markets.businessinsider.com/news/stocks/retail-investors-quarter-of-stock-market-coronavirus-volatility-trading-citadel-2020-7-1029382035.

Yiu, T. (2019). Understanding NLP and Topic Modeling Part 1. *KDNuggets*.

https://www.kdnuggets.com/2019/11/understanding-nlp-topic-modeling-part-1.html