Text and network based processing for financial machine learning



Dragos Gorduza

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

April 23 2025

Supervisors : Stefan Zohren and Xiaowen Dong

Aknowledgements

Writing this thesis would not have been possible without the help and support that I received throughout this time from my family, friends, and colleagues.

My thanks go to my supervisors Xiaowen Dong and Stefan Zohren whose wise advice and mentorship allowed me to progress through the stages of this work. Their guidance helped me structure my thoughts when they were in turmoil and their support allowed me the freedom to explore the different strands of research that make up this thesis. I would like to thank my collaborators and co-authors Elliott Ash, Yaxuan Kong, Adam Muhtar, Valentina Semenova and William Wildi for always being there to give advice and offer suggestions in order to bring the different papers making up this thesis to their present state. Special thanks go to the ETH's Law and economics department, in particular to Professor Elliott Ash's lab. Their help and support allowed me to develop the project that eventually became a chapter of this thesis. My thanks also extend to the Bank of England and specifically the ARTIS division which welcomed me during my internship allowing me to produce another chapter of my thesis as well as familiarize myself with the policy and research environment of a central bank.

I would like to thank my parents, Dan and Daniela Gorduza, who have stood by me and encouraged me at every stage of my studies. I would also like to mention all members of my family whose nurturing presence helped me to where I am now both personally and academically.

I want to thank my partner, Isabella Martus whom I was always able to count on to get me through the highs and the lows of writing a thesis.

My thanks extend to the community of people who accompanied me throughout my thesis. In my department at the Oxford Man Institute of Quantitative Finance, I have made meaningful ties with creative and supportive individuals. Exchanges with Pierre Osselin, Stacy Pu, Baskaran Sripathmanathan, Bohan Tang, Deborah Sulem, Dylan Sandfelder, Ning Zhang, Scott le Roux, Keyue Jiang, Ocariz Borde and Li Zhang have helped shape my thinking throughout the thesis. Thanks especially to Yin-Cong Zhi who was available ever since the Transfer to offer up his suggestions for how to best navigate the Oxford research milestones.

I would also like to express my gratitude to the ESRC for providing me with the funding that allowed me to live and complete my thesis in good conditions. Additionally I would like to thank my reviewers Professor Tomaso Aste and Stephen Roberts for their helpful comments and advice which helped improve this thesis.

Declaration

I declare that this thesis is my own work either as sole first author for the papers in Chapters 3, 4, 6 and 7 or as joint first author in Chapter 5. Both as first and joint first author I contributed to the writing, code, experiments, interpretation and discussion of the results. Below, I detail the contributions made for each paper.

I have completed :

- Chapter 3 : 'Understanding financial market stability using graph machine learning' as first author in collaboration with Stefan Zohren and Xiaowen Dong.
- Chapter 4 : 'Extracting alpha from financial analyst networks' as first author in collaboration with Yaxuan Kong, Stefan Zohren and Xiaowen Dong.
- Chapter 5 : 'Wisdom of the crowds or ignorance of the masses? A data-driven guide to WSB' as joint first author alongside Valentina Semenova in collaboration with William Wildi, Stefan Zohren and Xiaowen Dong.
- Chapter 6 : 'PRET: domain-adapted transformer for machine-assisted prudential supervision' as first author in collaboration with Adam Muhtar at the Bank of England.
- Chapter 7 : 'Common causes, common risks' as first author in collaboration with Elliott Ash at the ETH Zurich.

Abstract

Financial markets are often treated as frictionless information processors in which prices instantaneously absorb all available signals. This thesis argues that such a price-first view obscures two pervasive realities of modern finance: (i) economically meaningful interdependencies among firms, investors, and regulators that are structural and persistent, and (ii) an ever-expanding textual information environment—news, disclosures, social media, and rulebooks—through which beliefs are formed and transmitted. I develop methods that extract and combine information from these two complementary data sources: networks and texts, to produce more faithful representations of financial complexity and to generate actionable signals at the asset, market, and regulatory levels.

In this thesis I explore how leveraging network science, graph machine learning and natural language processing can allow us to better represent these interactions as networks. I contribute to two directions in the literature : 1) using network representations of inter-firm relations to build volatility and return forecasting models, thereby underscoring the importance of financial networks as financial artefacts and 2) using texts to represent novel dimensions of financial and economic interconnection by mining information from textual data through novel financial-domain specific machine learning tools.

Firstly, I introduce novel approaches for graph-based learning in a market instability monitoring context by training a graph auto-encoder to reconstruct correlation networks; reconstruction difficulty emerges as a forward-looking proxy for volatility and improves out-of-sample volatility forecasts. I then build an analyst-coverage network from the IBES database and show that graph attention networks leveraging this structure outperform alternatives for stock-level buy/sell/hold prediction, and that coverage overlap carries information beyond industry classifications and correlations. These tools serve to showcase how relational structures like networks are well suited to describe and extract insights from complex interactive relationships in financial markets. Secondly, to bridge text and network processing, I contribute new financial-domain specific natural language processing models designed to extract information from the universe of financial textual data available to investors. I perform sentiment classification and topic modelling on the online forum WallStreetBets to extract information about investor herding and the resultant inter-asset momentum. I then identify profitable trading portfolios from those herding networks, thus further highlighting how text and network processing can together enhance financial insight-gathering. Lastly bringing the focus fully on text processing, I build a financial-regulatory unsupervised retrieval model which leverages domain adaptation to outperform state of the art approaches on regulatory retrieval tasks with direct applications to regulatory supervision. I additionally develop narrative-similarity measures from news

and corporate filings and show it helps explain future inter firm correlation. All these tools serve to better measure links between financial entities present in text and more broadly to extract financially and economically significant features from large unstructured textual corpora.

Together, these contributions demonstrate that modern information extraction from networks and texts complements prices. Integrating network and textual analyses thus provides deeper insights into market dynamics, enhances risk forecasting, and offers practical applications for trading strategies and regulatory oversight.

Contents

1 Introduction **9**
- 1.1 Motivation . . . 9
- 1.2 Overview of the thesis and contributions . . . 16
- 1.3 Publications and chapters . . . 20

2 Literature review **22**
- 2.1 Financial texts as representation of markets . . . 22
  - 2.1.1 Text feature extraction from dictionaries . . . 22
  - 2.1.2 Text feature extraction from statistical and machine learning methods 23
  - 2.1.3 Text feature extractions from deep learning methods . . . 24
- 2.2 Financial networks as representations of markets . . . 27
  - 2.2.1 Extraction of networks from relational data . . . 28
  - 2.2.2 Extraction of networks from similarity measures . . . 29
  - 2.2.3 Extraction of networks from textual data . . . 30
- 2.3 Measuring financial network features . . . 31
- 2.4 Graph machine learning applied to financial networks . . . 32

3 Understanding financial market stability using graph machine learning **34**
- 3.1 Introduction . . . 35
- 3.2 Problem setting and methodology . . . 36
  - 3.2.1 Problem setting . . . 36
  - 3.2.2 Data . . . 37
  - 3.2.3 Proposed methodology . . . 38
- 3.3 Results . . . 42
  - 3.3.1 Correlation between $\mathbf{AUROC}_{t+1}$ and volatility . . . 42
  - 3.3.2 Next-day volatility forecasting . . . 43
- 3.4 Discussion and Conclusion . . . 44

4 Extracting Alpha from Financial Analyst Networks **47**
- 4.1 Introduction . . . 48
- 4.2 Literature review . . . 49
- 4.3 Methods . . . 51
  - 4.3.1 Data . . . 51
  - 4.3.2 Problem formulation . . . 52
  - 4.3.3 Proposed methodology . . . 53
  - 4.3.4 Training strategy . . . 54

4.3.5 Comparative baseline models . . . 55
4.3.6 Ablation studies . . . 55
4.3.7 Evalutation metrics . . . 56
4.4 Results . . . 56
4.4.1 Comparison of predictive performances . . . 57
4.4.2 Ablation studies . . . 58
4.4.3 Return correlation analysis . . . 61
4.4.4 Turnover and cost analysis . . . 62
4.4.5 Attention analysis . . . 63
4.5 Conclusion . . . 64

5 Wisdom of the Crowds or Ignorance of the Masses? A data-driven guide to WallStreetBets 67
5.1 Introduction . . . 68
5.2 Data . . . 71
5.3 WallStreetBets Forum Dynamics . . . 72
5.3.1 The topic landscape of WSB . . . 72
5.3.2 How are assets related to each other on WSB? . . . 74
5.3.3 Asset network structures . . . 76
5.3.4 Returns preceding and following posts . . . 80
5.3.5 Is information on WSB predictive of returns? . . . 82
5.3.6 Granger causal relationships on WSB . . . 83
5.3.7 Additional results . . . 86
5.4 Conclusion . . . 86

6 PRET: Domain-Adapted Transformer for Machine-Assisted Prudential Supervision 89
6.1 Introduction . . . 90
6.2 Literature review . . . 91
6.2.1 Information Retrieval . . . 91
6.3 Methodology . . . 94
6.3.1 Data Collection . . . 94
6.3.2 Model setup . . . 95
6.3.3 Fine-tuning using generative pseudo labelling . . . 96
6.3.4 Other methodologies . . . 98
6.4 Evaluation . . . 98
6.5 Results . . . 100
6.6 Discussion . . . 101
6.7 Conclusion . . . 103

7 Common Causes, Common Risk? 105
7.1 Introduction . . . 106
7.2 Literature . . . 107
7.3 Methods . . . 110
7.3.1 Data description . . . 110
7.3.2 Relatio parsing of news . . . 110

7.3.3 Mapping Relatio clusters to SEC Filings . . . 112
7.3.4 Using Node2Vec to extract common risk exposures . . . 112
7.3.5 Building baselines of text-based common risk exposures . . . 114
7.3.6 Regression setup . . . 114
7.4 Results . . . 115
7.5 Discussion . . . 116
7.6 Conclusion . . . 117

8 Discussion and Conclusion 119
8.1 Discussion . . . 119
8.2 Ethical considerations . . . 122
8.3 Limitations and future work . . . 124
8.4 Conclusion . . . 125

9 Appendix A Supplementary results and material for Chapter 3 127

10 Appendix B Supplementary results and material for Chapter 4 129
10.1 Hyperparameter tuning . . . 129

11 Appendix C Supplementary results and material for Chapter 5 131
11.1 Detailed Data Appendix . . . 131
11.1.1 Data scraping . . . 131
11.1.2 Due Diligence (DD) posts . . . 132
11.1.3 Sentiment detection . . . 133
11.2 Granger causality test details . . . 135
11.3 Topic Model . . . 136
11.4 CAR for Popular Stocks . . . 136
11.5 Discussion of Clusters and Returns . . . 136
11.6 Short and long term trade signals . . . 141
11.6.1 Due diligence reports . . . 141
11.6.2 Longer-term trade signals . . . 146
11.6.3 Due dilligence tables . . . 149

12 Appendix D Supplementary results and material for Chapter 6 154
12.0.1 Working Hypothesis, Problem Setup, and Embedding Generation 154

13 Appendix E Supplementary results and material for Chapter 7 158

Chapter 1

Introduction

1.1 Motivation

Information extraction and financial markets

Financial markets have become the central resource allocation mechanisms of modern economies [218]. They have achieved this celebrated status in line with Fama's hope of an Efficient Market Hypothesis (EMH) based on the expectation from policymakers, market participants and the public that "prices fully reflect all available information" [92]. Indeed, even policymakers and regulators now primarily choose to refer to financial asset prices as indicators of economic health[1] [41]. Prices in financial markets are treated not merely as the expected future discounted cashflow of an asset but real-time computational devices that seamlessly aggregate heterogeneous knowledge dispersed across millions of agents better than any central agency could and thereby should guide decisions on consumption, investment, and production as argued emphatically by Hayek [117]. From this perspective, the arrival of new data is automatically and effortlessly passed through rational "homo economicus" agents to prices. The price system in financial markets appears as a highly efficient, self-adjusting information aggregator.

Yet, the information-processing function of markets is neither frictionless nor exhaustive. Indeed, in the process of recording, documenting and anticipating changes in their business landscape, market participants create traces which we propose to call artefacts following operational research literature [175]. Financial prices often reflect only a compressed or delayed representation of the underlying information environment as the incorporation of these artefacts takes place at different speeds due to their being more or less easy to decode. Certain types of knowledge, such as relational dependencies, legal details,

[1] "*[A]sset prices and yields are potentially valuable sources of timely information about economic and financial conditions. Because the future returns on most financial assets depend sensitively on economic conditions, asset prices [...] should embody a great deal of investors' collective information and beliefs about the future course of the economy.*" [41]

strategic interactions, or qualitative assessments, for example may not be immediately or fully integrated into tradable prices. Meanwhile, markets permanently generate new information in the process of interpretation itself, as analysts, media, and institutional investors continuously attempt to forecast the beliefs of others, continuously producing textual content traces such as news, reports or regulation. This very complex and at times painful process of information extraction underpins, as Kelly et al. [131] argue, the current "production function of the modern asset management industry" centered on the competitive extraction, interpretation, and transformation of vast streams of data into investable signals. The modern asset management industry can be viewed as a vast and sophisticated apparatus for gathering, processing, and competing over informational advantage. The "production function" of finance therefore reflects not only capital allocation decisions, but also the large-scale challenge of information extraction and refinement.

These observations highlight two important limitations of a price-first perspective on financial markets, which this thesis proposes to explore. First, several economically meaningful and long-range dependencies be they between firms, investors, suppliers or regulators are structural and persistent, but are not directly observable in price series alone. Thus, they require a bespoke mathematical tool that embodies them : networks, which in turn have to be appropriately processed to be useful. Second, the amount and complexity of information that market participants must interpret has expanded dramatically, particularly through the proliferation of textual data such as financial disclosures, news reporting, social media commentary, and regulatory documents also require equally bespoke text processing tools in order to be incorporated into prices. Both of these limitations suggest that efficient information aggregation at the scale required by the current reliance on markets as forms of governance descried above, requires tools going beyond price movements alone which this thesis proposes to develop in both text and network processing.

This motivates a richer view of financial information environments along two complementary dimensions that this thesis addresses:

- The structure of financial systems as networks, where relationships among firms, investors, and markets create interdependencies that shape risk transmission, return co-movements, and systemic fragility.

- The linguistic and narrative content embedded in financial texts, which provide qualitative information, shape expectations, influence collective sentiment, and ultimately contribute to market dynamics.

The next section therefore explores how networks can be used to represent complexity in financial markets—capturing linkages such as supply chains, board-level connections,

and shared investor bases that influence co-movement and contagion dynamics. Following this, the following section examine how natural language processing (NLP) tools provide systematic methods to extract meaning from the vast textual environment in which investor beliefs are formed and reshaped. Given the crucial nature of information extraction for good quality financial decisions, and the importance of financial decision-making on economies and livelihoods, this thesis is motivated to expand the knowledge frontier by exploring the question of how can we best extract information about financial relationships and .

Together, these two perspectives—networks and texts—demonstrate how modern information extraction methods can complement traditional price-based analysis to give a more complete representation of financial market complexity.

Financial system complexity represented through networks

'To safeguard against systemic risk, the financial system needs to be managed as a system', with these words, the chief economist of the Bank of England Andrew Haldane introduced the notion of network-level monitoring of financial markets into the mainstream in the wake of the 2008 financial crisis [108]. He called for a recognition of the complex web of interdependencies that link financial actors and which became a profoundly destabilising force during the crisis even for firms whose price trajectory had seemed promising. This call prompted multiple research projects showing how 'robust-yet-fragile' [96] financial markets can display great resilience before collapsing rapidly from an otherwise 'minor' disruption to one single but central part of the system as Lehman Brothers' collapse showcased [108, 109]. This inquiry would lead researchers to question the pre-existing view of interconnections in financial markets : the idea that historical correlations could serve as an acceptable proxy for expected asset to asset shock transmission as described in the rich literature of the Modern Portfolio Theory [159] and the Capital Asset Pricing Model [211]. This view is stylistically depicted in Figure 1.1 representing a investor risk as a proportion of how 'diversified' their stock portfolio is.

This approach, though very popular to this day, has received criticisms, in part because the persistence of market-wide contractions in 1962, 1987, 1989, 2000 and 2008 saw all stock's co-movements increase significantly, making the benefits of diversification disappear precisely when they are most needed [195, 185]. These sharp increases in co-movements suggest that multiple asset holders all acted at once -usually selling- thus driving the price down [185]. Indeed, as investor's limited attention is spread across similar sources of news, this facilitates stock return spillovers and predictable lead-lag effects by orienting the attention of multiple investors at once [13, 69].

Networks, appear well designed to account for the complexity of relationships in a market which lead to spillovers, long range dependencies and significant herding behaviour.

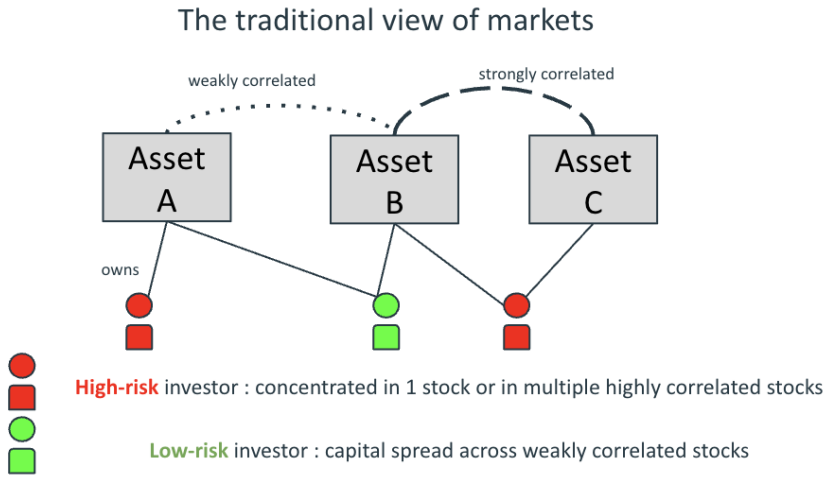


Figure 1.1: The traditional view of risk according to the MPT

A network can encode a common source of news as an edge between the two investors. If firms are under the same jurisdiction, the network can encode this with and edge connecting all relevant firms. Similarly, investors sitting on the same board and therefore serving as potential information conduits between firms as suggested in [100], can be connected by an edge. Traditional investment analysis already incorporates part of this thinking through industrial classifications such as S&P's GICS which can be thought of as edges connecting two assets. A financial network can be understood as a more generalised version of industry clusters, encoding relationships between assets and asset holders along multiple dimensions. The network perspective allows an at-a-glance measure of how stocks are related and ultimately why they might co-move. Hand-compiled resources such as supply chain knowledge graphs that relate entities through specific symbolic relationships have emerged as a good way to directly represent links between assets in a way price on its own simply could not by including explicit contracts or supply relationships between firms [164]. Figure 1.2 describes how this network could look like with a few types of relationships displayed between assets and investors.

In this thesis, we argue that innovations in network science and graph machine learning make financial markets well suited for insight extraction. In a network, the relationships themselves carry crucial context that often cannot be captured by simply listing node attributes in a table. This interconnected nature means that measurements or changes to one part of the network can ripple across other parts in non-obvious ways as was highlighted in the seizure of the inter banking market in 2008 or in the long range effects of Chinese production closures on global supply chains in 2020. Extracting meaning from these interactions requires specialised tools from graph theory [34]. For instance, local interactions among nodes may lead to emergent global properties that are invisible without studying the network's topology as a whole. These global properties can be revealed via

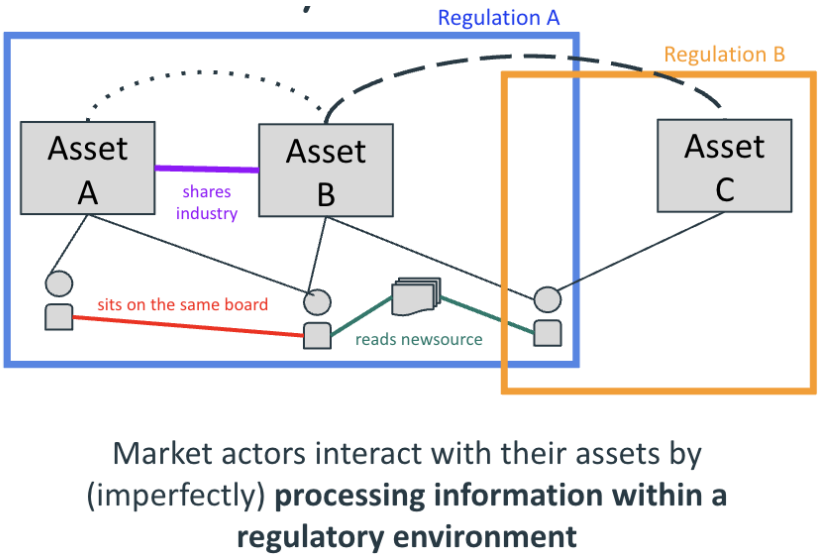


Figure 1.2: What a financial network could look like

metrics such as degree distribution, centrality, or community structures. These dedicated graph-based methods that preserve and exploit the structure of relationships ensure that valuable information is not lost in the oversimplification that results from treating network data like standard tabular data. Moving away from traditional graph-theory metrics, modern machine learning tools such as graph neural networks are adapted to work on this type of network data [113]. They are able to learn patterns directly from the network structure and outcomes of interest whether at the asset level or market level. Furthermore, graph neural networks have the advantage over traditional network statistics that they can flexibly incorporate information from nodes of the network instead of only relying on edge-level data.

Recognising the importance of networks to represent financial market complexity, this thesis proposes to build novel network representations of firm relationships from underexploited sources and leverage those network science and graph machine learning measurements.

Financial markets complexity measured through texts

This thesis also explores how new textual information, be it news, regulations or earning call transcripts, is processed by investors in order to properly represent the complexity of the financial system. Haldane, in his seminal speech references the telling example of the significantly larger number of pages a 'diligent analyst' needs to parse when assessing a complex derivative like a Collateralized Debt Obligation compared to a regular security [108]. He asserts this leads to a vulnerability and a source of risk as the price of that derivative may not reflect he actual information content of the legal documentation.

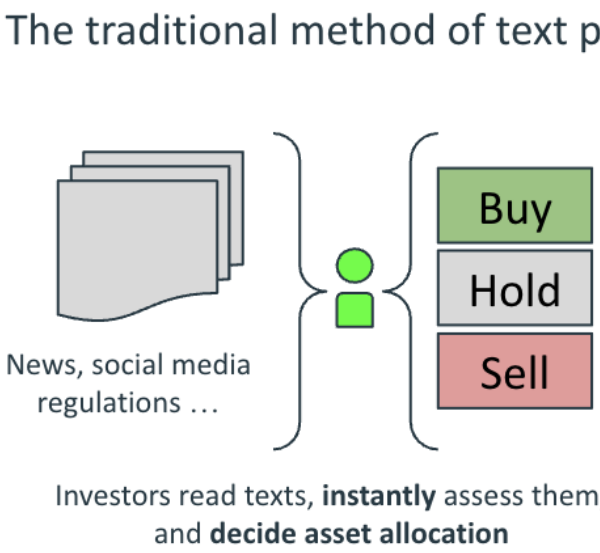


Figure 1.3: The traditional view of financial text processing

Canonically, analysts assess news and estimate their impacts on firms in a qualitative manner as in Coase's pioneering attempted to extract economic indicators of externalities from firm contracts by hand [68]. Figure 1.3 proposes a simplified description of how text processing can be thought of under the traditional view. However, there is a growing recognition that, contrary to the vision described in the Efficient Market Hypothesis (EMH) [92], price and similar quantitative metrics alone are a poor descriptor for the actual health of a company [144]. In other words, new information is not always properly interpreted and integrated into the price signal.

Indeed, on one hand, investors can not interpret the information automatically due to idiosyncratic behavioural biases [221]. On the other hand, broader 'economic narratives' influence their ability to process new information [212]. These narratives are stories which are broadly shared among investors such as notions of 'crisis' or 'economic boom' and, because they are shared, condition the intake of new information [212]. For example, Nyman et al. [177] showed that uniformity in financial news tone is predictive of increasing volatility as all agents are primed to interpret new shocks the same way by the information they consume. Furthermore the increasing informational load placed on investors by faster news cycles, increased social media usage and the increasingly large corpus of online information also limits the speed of information processing in the era of 'Big Data' [178]. This limits the ability of new information to feed into prices through only manual analysis due to the labour-intensiveness of information processing analysis. The new constraints placed on textual information processing in modern financial markets call for a renewal of text processing methods.

To appropriately deal with the scale and pace of new textual data, researchers and

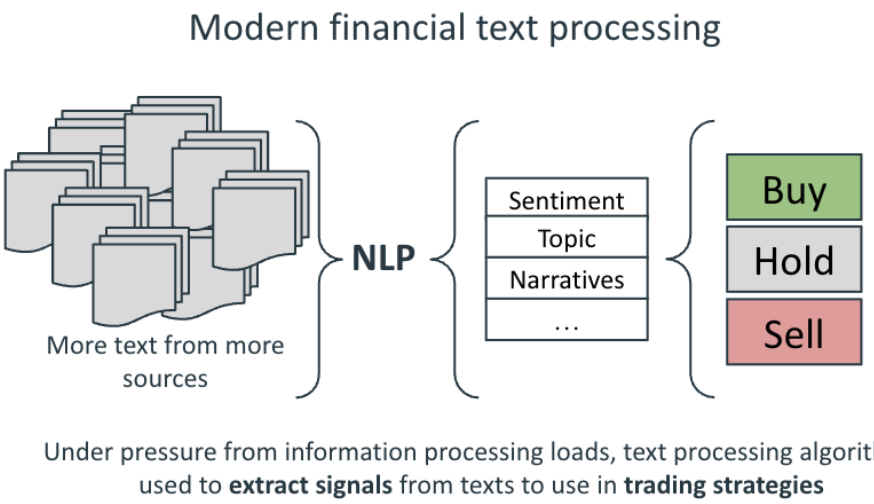


Figure 1.4: A modern description of financial text processing systems

practitioners have turned in recent decades to automated tools for text parsing. Modern natural language processing (NLP) algorithms for finance have evolved from simple dictionary models [154] composed by financial experts in to the recent extremely large modern generative models trained on a very large quantity of text [181, 241]. The usefulness of these approaches lies in transforming texts into useful downstream features such as sentiment [21, 233]. For example, if an article that talks about a firm in a positive light might imply upward price movement, the tone of the article, once properly extracted could offer a leading indicator for market movements in a similar manner to the canonical hand-parsing described above. The rise in social media as a communication vector motivates increasing interest in 'nowcasting' (ie high-frequency monitoring) this sentiment measure as a 'wisdom of the crowds' with applications in asset pricing, volatility forecasting and beyond [1, 128]. Figure 1.4 describes how the modern text processing system has evolved towards a more automated feature-extraction approach to market data.

However, despite the increase in performance of language models in recent years, financial NLP faces certain challenges which this thesis proposes to address. Firstly, these tools require careful domain adaptation during training in order to understand the particular context of a financial term such as 'investment bank' which may differ from the generic definition of that term like in 'river bank'. Models fine tuned on financial data such as FinBERT [21] or BloombergGPT [241] consequently tend to perform better on financial text processing tasks. However, their training data is in most cases either limited to regulatory fillings or financial news as those are the most easily accessible data sources. This leaves the fields of specific language models to deal with financial regulations and financial social media discussions as a fertile area for research. Secondly financial NLP applications, as suggested by Shiller [212] require capturing more than just document

sentiment, but also topic and broader 'narrative' to realistically assess the market effect of texts. This can involve looking at more complex syntactic and grammatical structures which can inform about who does the actions and to whom they are done [141, 23].

Recognising the importance of text as an information source in financial markets and the limitations of existing tools to process it, this thesis proposes a suite of novel methods adapted for new datasets and new applications from asset pricing to market regulation.

Combining text and networks in financial market information processing

Networks and textual information each provide powerful yet distinct lenses through which to view financial complexity. Financial systems are, at their core, systems of interdependent agents whose behaviour is shaped by the data they have at their disposal about their own economic health and the structural relationships with other linked actors [175]. This data is not "freely available" to these agents and information extraction challenges that motivate this thesis highlight how policymakers and financial market participants can profit from applying these tools to extract more information at scale.

In this thesis, we argue that the joint modelling of network and textual information provides a more faithful representation of financial reality. It allows for a unified view in which narratives, sentiment, and attention become quantifiable drivers of investor or firm behaviour (and are therefore ultimately reflected in prices), while structural relationships between firms, investors, and regulators also control how new information ricochets across markets. The resulting framework holds promise for both academic understanding and practical applications — from forecasting volatility and identifying contagion channels to monitoring systemic risk and designing more adaptive financial supervision tools.

1.2 Overview of the thesis and contributions

Contribution summary and research directions

This thesis draws together multiple strands of information processing in financial markets to show how they can be used to uncover novel connections between financial entities at the asset, market and regulatory levels. We provide insights on how to a) build new financial networks from underexploited datasets before exploiting traditional graph theory techniques such as community detection and graph machine learning to extract useful signals from those networks and b) build novel natural language processing applications for sentiment, topic and narrative extraction.

Literature review

Chapter 2 presents a review of the literature focused on the different aspects of processing text and networks as data in a financial context. It highlights the tools from natural language processing, network science and graph machine learning as well as the tasks and domains that this thesis covers.

Network processing approaches

Graph Auto encoder for financial volatility forecasting

In Chapter 3, we present a novel application of an unsupervised graph machine learning algorithm : a Graph auto-encoder (GAE) in an important volatility forecasting task. We train the GAE to reconstruct the network connections between firms in a market in order to improve the performance of an important task to both practitioners, researchers and regulators : volatility forecasting. We show that the reconstruction quality is inversely correlated with volatility. We develop our own metrics based on the reconstruction ability of a GAE of a correlation network. Our method displays improved performance in an autoregressive out of sample volatility forecasting setup. This approach extends the literature by yielding a new flexible measure of risk in financial markets which result from increased interconnections. Moreover, as most network based analysis of financial volatility only leverage edge-based information present in the structure of the network, the ability of the GAE to incorporate node level information opens new opportunities for inclusion of more complex information. This has important implications for designing volatility forecasting models and could be of interest for institutional policymaking who must closely monitor market-wide volatility and are interested in combining network and node level information.

Graph Attention Network and the analyst coverage network

After showing the performance of graph machine learning to extract information for market-level signals, in Chapter 4 we turn to a supervised learning task to extract trading signals at the asset-level. Our contribution is twofold. Firstly, we leverage a lesser-known dataset of financial professional analysts: IBES to extract valuable information about investor attention. This dataset is a relational table which we treat as a bipartite graph of analysts and firms the cover. We build a network between firms based on the number of analysts that cover both of them by projecting this bipartite matrix into the firm by firm dimension. This represents a unique proxy for investor attention as previous studies pointed to the influential role of analysts in opinion formation and attention direction. Secondly, we benchmark multiple graph machine learning models

on a node-level forecasting task to identify buy-sell-hold signals given their stock prices and multiple different network representations including the IBES network. This allows us to show that a graph attention network (GAT) outperforms other competing graph neural network architectures on this node-level forecasting task. Moreover, we also show that the analyst coverage information outperforms competing representations of asset to asset relationships such as GICS industrial clusters and correlation networks. This work has novel useful applications for financial signal generation with a trading perspective and it also helps underscore the importance of combining rich network representations of financial relationships with the proper extraction tool able to balance both node and network information at the same time.

Network and text processing approaches together

WallStreetBets investor forum behaviour : showcasing text and network contribution to financial analysis

In Chapter 5, we bring together both strands of analysis that motivate this work : jointly analysing the network and textual information for insight into how a community of investors functions and study consequences of those behaviours at the asset level. To do this, we leverage a novel dataset of conversations between investors from the notorious web forum R/Wallstreetbets (WSB). Our first contribution focuses on characterising the informational content of the forum's trading signals. We extract the tone of each post using a fine tuned sentiment detection algorithm and we use it to measure the implied returns of a strategy that buys and sells the stocks according to forum sentiment. We show that the forum on average lags market movements and that sentiment on its own does not lead to a profitable strategy. Our second contribution lies in proposing two different ways to visualise firm-to-firm connections. We measure the overlap between the communities discussing each stocks and the overlap between the topics used in the discussions of each stocks. We represent these overlaps as two asset to asset networks (in a similar way to the IBES coverage overlap network from Chapter 3). Lastly, we bring both text and network perspectives together by grouping the stocks in coherent communities based on their network position using the Louvain community detection algorithm [224]. We compare the return of each of these network-informed groupings to show that it is only through this network lens that the small communities of more profitable ideas can be uncovered. These novel interpretations of the informational content of WSB have useful applications for analysing the behaviour of so-called 'meme stocks', ie. assets whose return profile is heavily dictated by social media sentiment.

Text processing approaches

Fine tuned models for financial regulatory analysis

Chapter 6 highlights a different but important way forward for modern NLP tools in finance and especially in financial regulation: building high-quality datasets, evaluation targets and fine tuned models. To answer the need for labour-saving methodologies in the arduous task of financial supervision, we build a retrieval model to retrieve the most relevant passages for a given financial regulatory question. This task is analogous to the tasks that financial supervisors must undertake when they attempt to monitor complex financial documents whose complexity as argued in Haldane [108] can be a source of global risk. To solve this task, we fine-tune a transformer model to predict the most pertinent response to a passage from all passages in the rule corpus. This method leverages the underlying network structure of financial rules where passages are tightly related to one another [15]. The novelty of our approach lies in delivering the first open source model trained on the Basel III financial rulebook: an important and highly relevant dataset for financial regulation. This chapter explores the building blocks of a Retrieval Augmented Generation model for financial regulation[145, 7].

Financial news narratives as interpretable measures of inter firm co movement

Chapter 7 details how we used a narrative extraction tool called Relatio [23] to build narratives about firms in financial and economic news data. These narratives are triplets of actors, verbs and patients indicating which entities 'do' an action to which other entities, allowing us to capture more interpretable and clear relationships between entities in financial text that allowed by traditional NLP tools. This approach then combines the extracted narrative triplets into a narrative knowledge graph. The narrative entities are mapped upon the relevant risk items of SNP500 firms, giving us firm-to-firm links built from narratives. The strength of linkages between firms indicates how exposed multiple firms are to the same type of narrative. This network measure is used as an explanatory factor in a regression task to explain the correlation coefficient between two firm's stock returns. This novel approach of estimating future firm co movement contributes to the literature in financial text analysis by demonstrating the importance of a 'narrative economics' [212] approach in understanding firm risk exposure, going beyond price-based measurements.

1.3 Publications and chapters

This thesis is presented in the University of Oxford's integrated thesis format [2]. All the work presented below is a first or joint first author work. The following papers have been accepted in a journal or the proceedings of a conference (chapters 3 to 5). The remaining chapters present papers which have been featured in conferences or are under consideration for conferences (chapters 6 and 7). Section 1.2 presents the individual chapters' contributions and how they are linked by the dual mandate of text and network processing for financial markets. Figure 1.5 summarises the work done to date as well as representing how each chapter contributes methodologically to either text or network processing methodologies in finance while also highlighting what application level (asset, market or regulatory) the paper explores.

[2] https://www.mpls.ox.ac.uk/graduate-school/information-for-postgraduate-research-students/submitting-your-thesis

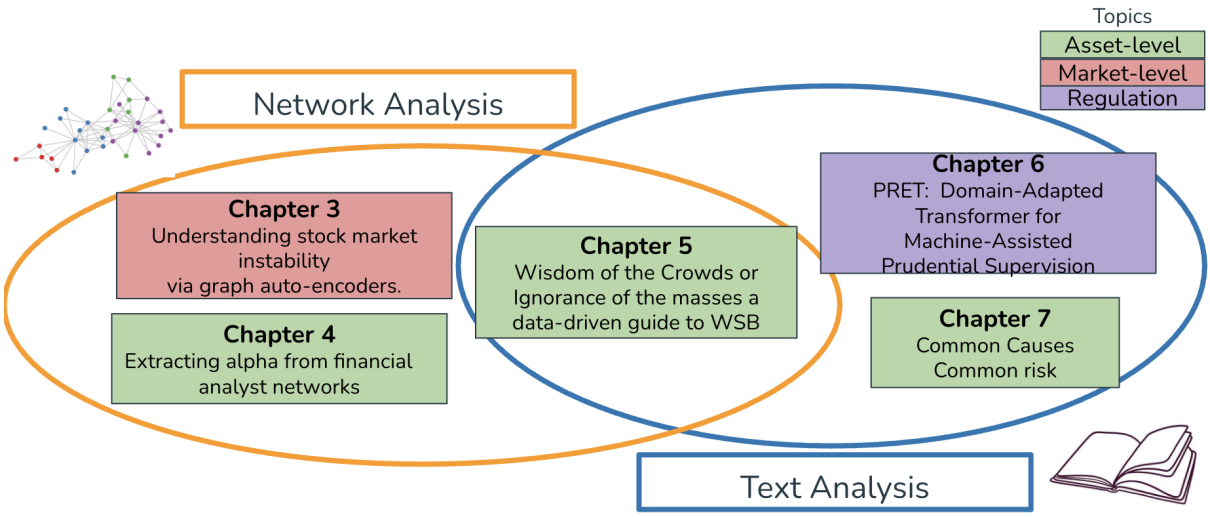


Figure 1.5: Chapters by methodology and topic

Chapter 3

Gorduza, D., Zohren, S., and Dong, X. (2025). Understanding stock market instability via graph auto-encoders. EPJ Data Science, 14(1):13.

Chapter 4

Gorduza, D., Kong, Y., Dong, X., and Zohren, S. (2024). Extracting alpha from financial analyst networks. In Proceedings of the 5th ACM International Conference on AI in Finance, pages 397–405.

Chapter 5

Semenova, V., Gorduza, D., Wildi, W., Dong, X., and Zohren, S. (2024). Wisdom of the crowds or ignorance of the masses? a data-driven guide to wallstreetbets. Journal of Portfolio Management, 50(4).

Chapter 6

Gorduza D. Muhtar A. PRET: Domain-Adapted Transformer for Machine-Assisted Prudential Supervision accepted as a non-archival publication at the 5th Workshop on ACM International Conference on AI in Finance LLMs and Generative AI for Finance

Chapter 7

Gorduza D. Elliott A. Common causes, Common Risk, Presented as a non-archival oral presentation at the 5th Annual Comptext Conference 2023

Chapter 2

Literature review

2.1 Financial texts as representation of markets

Financial text information extraction (IE) refers to the set of techniques for transforming unstructured textual data (the data artifacts we introduced above) in finance into structured information or features ready for downstream analysis. Financial markets produce a wide variety of textual artifacts, including legal documents, regulatory filings, news articles, earnings reports, social media posts, analyst research, and more. Historically, such text was analyzed manually as in Ronald Coase's early work in economics involved reading legal judgments to quantify the relation between laws and economic externalities [68]. For decades, a lack of robust computational models and limited computing power meant that financial text had to be interpreted by human experts [24, 155]. However, the past two decades have seen rapid advancements in natural language processing (NLP) and text mining, enabling automated extraction of insights at scale. Surveys of the literature note that innovations in text-mining techniques, coupled with greater data availability and faster computing, have led to text becoming a de facto data source in finance [132, 139]. Today, financial NLP is a rich field intersecting academia and industry, encompassing tasks from sentiment analysis and event extraction to question answering and regulatory compliance monitoring [85].

2.1.1 Text feature extraction from dictionaries

In the early years of automated text analysis, researchers relied on manual and dictionary-based approaches to extract information from financial text. This is an example of an 'expert-system' approach, that is an attempt to survey experts and store their insights into replicable rules that can be applied at scale [75]. A prime example is the use of sentiment word lists to gauge tone in news or reports. One of the first influential studies of media text and market impact, by Tetlock [220], counted occurrences of negative words from a general-

purpose sentiment lexicon (Harvard IV-4) in a Wall Street Journal column to quantify pessimism . Tetlock's analysis showed that high pessimistic tone in news predicted short-term downward pressure on stock prices followed by reversion, demonstrating that textual sentiment contained predictive power for markets. Around the same time, researchers began examining internet forums: Antweiler and Frank [20] analyzed over 1.5 million messages from stock message boards, using simple text statistics and sentiment word counts to conclude that online discussions could predict market volatility (though not large price moves). These early studies were largely lexicon-based, using manually curated dictionaries of positive/negative terms. In the finance domain, it soon became clear that domain-specific dictionaries were needed: words like "liability" or "capital" have special meanings in finance that make a generic sentiment dictionary misleading. Loughran and Mcdonald [154] addressed this by crafting a finance-specific sentiment lexicon after examining thousands of SEC filings. Their dictionary – categorizing words into negative, positive, uncertainty, litigious, etc. – became a standard for sentiment analysis of financial reports [155]. Several other studies followed this approach building dictionaries specialized in various financial subdomains such as Correa et al. [73] for central banking stability. However, although sentiment is a very helpful feature for financial markets as it compresses a large amount of information about market outlook in a single variable, expert systems did not only focus on extracting sentiment. The second main research direction was topic extraction [24]. These could then be used to extract the frequency of certain significative keywords from large financial text corpora that became more widely available as Baker et al. [31] showed when measuring the evolution of 'Economic policy uncertainty' index from a corpus of newspapers. The use of expert dictionaries crafted with expert input helps abstract away from what the article says about these terms and instead creates an abstract measure of uncertainty about economic variables.

2.1.2 Text feature extraction from statistical and machine learning methods

As digital text data and computing resources grew in the 2000s, researchers turned to statistical machine learning to improve on manual techniques. This closely followed moves in general-purpose automated sentiment extraction methods such as for movie or consumer reviews [29, 188]. Texts became represented in high-dimensional feature vectors which counted word or group of word frequencies (e.g. bag-of-words [115], n-grams [127] or term frequency TF-IDF [202]), and algorithms like Naïve Bayes, logistic regression, or Support Vector Machines (SVM) were trained for tasks such as classification or regression [24]. The innovation here comes from the shift from expert systems where rules are designed from the top-down to explicitly codify expert heuristics to a statistical learning system where these rules are learnt from the bottom up, from large datasets labelled by

humans. For example, Kogan et al. [136] pioneered text regression on financial reports: they represented the Management Discussion & Analysis sections of annual 10-K filings as vectors of word frequencies and used penalized regression models to predict each firm's future stock volatility. Their results indicated that textual content in financial reports held significant predictive signal for realized volatility, beyond traditional quantitative metrics. Similarly, researchers applied SVM classifiers to financial news by building trading models that classify news articles as positive or negative news for a stock and then simulate market positions accordingly, finding that incorporating text-based signals could yield modestly improved returns [170, 207]. This data-driven approach was exemplified by contest challenges such as the SemEval-2017 Task 5 on financial sentiment analysis, which provided a annotated dataset of financial tweets/news and spurred development of specialized classifiers [76]. While sentiment classification was a popular task, statistical learning approaches also extended to event detection where the task was to automatically identify emerging 'events' [242].

Statistical models were not only used for classification, unsupervised techniques to discover latent structure in financial text became increasingly popular in financial text mining as demands on analyst time increased and the volume of text (in particular on social media from the latter 2000s) increased. Statistical topic modeling methods like Latent Dirichlet Allocation (LDA) originally developed for general-purpose text [43] were applied to corporate filings and news to uncover thematic content (e.g. topics related to macroeconomic risks or market sentiment) [11]. While these unsupervised approaches do not extract specific named facts, they reduce textual data to interpretable factors that could be used as features in financial prediction models.

Overall, the introduction of statistical learning enabled more nuanced and adaptive information extraction from financial text compared to lexicon-only approaches, from sentiment classification, topic extraction to detecting specific events (mergers, earnings surprises, etc.) through trained models. However these approaches began to reach limits in financial and general text-processing applications due to issues surrounding their ability to deal with context (differentiating the 'bank' in 'bank robber' from 'river bank' for instance) and in an environment of increasing text availability, this was a comparative disadvantage against more advanced deep-learning-based approaches.

2.1.3 Text feature extractions from deep learning methods

The 2010s witnessed a paradigm shift with the rise of distributed representations (embeddings) that went beyond simple word counts but instead tried to learn a good representation *before* sending the embedded representation of the document to a separate statistical learning tool. This, alongside the massive increase in data availability, and increases in performance of deep learning algorithms, significantly improved the accuracy of information

extraction from text.

First, word embeddings such as Word2Vec [169] allowed each word to be represented by a dense vector capturing semantic context using neural networks to predict the missing words in a sentence with randomly deleted words. Additionally, word embeddings produced by Word2Vec could reproduce semantic links such as 'man is to woman as king is to queen' in vector space. Building on this innovation, deep neural networks became state-of-the-art for many NLP tasks. In finance, early applications used recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to process text sequences [66]. These could for example be used to perform event extraction from news, embed each event in a vector space, and then used those representations to predict stock price movements [84]. This end-to-end deep learning approach outperformed earlier methods that relied on bag-of-words or hand-crafted features, illustrating how event-level information extraction could feed directly into market forecasting models. Similarly, researchers applied long short-term memory (LSTM) networks to financial sentiment tasks, finding that these models could capture negation and context in analyst reports or social media posts better than keyword counts. For instance, LSTMs could be used to extract sentiment from Twitter posts to measure aggregated mood features (e.g. calm vs. anxious) which correlated with and could help predict stock market indices [45].

An additional major leap in information extraction from text came with the attention mechanism and transformer architecture which emerged to solve an issue of the encoding-decoding framework in machine translation (namely that the actions of the recurrent state of the art method at the time could not be parallelised efficiently and would loose information by compressing sentence up to the evaluated token essentially leading to both computational efficiency and performance losses) [230]. Transformers enabled much deeper language understanding and context processing, which led to pretrained language models that could be fine-tuned for finance [230, 24, 21]. The watershed moment was BERT (Bidirectional Encoder Representations from Transformers) , soon followed by financially pretrained BERT models [82]. FinBERT was one of the first, created in 2019 by continuing BERT's training on a large corpus of financial filings and news [21]. FinBERT achieved state-of-the-art results in financial sentiment classification, outperforming both classical methods and general-purpose BERT on domain-specific data. Transformer models rapidly became the default for financial NLP tasks: they have been used for extracting relations from broker reports (e.g. identifying causal statements about stock moves), classifying the intent of regulatory documents, and parsing complex disclosures. Fine-tuning allowed incorporation of even fairly specialised domain knowledge – for example, researchers fine-tuned BERT to identify "greenwashing" claims in corporate sustainability reports (claims that overstate a firm's environmental responsibility) and found it could detect subtle misleading language in context [239, 217]. Similarly, in the regulatory domain,

BERT-based models have been fine-tuned to map user queries to relevant clauses in regulatory texts, enabling intelligent compliance document search [191].

Most recently, the advent of large language models (LLMs) – extremely large transformer-based models with billions of parameters has begun to influence financial text processing whether supervised or unsupervised. General LLMs such as GPT-3 and associated models demonstrated an ability to perform NLP tasks with minimal or no task-specific training which referred to as few-shot or zero-shot setup [181]. However, finance presents unique vocabulary and content, motivating specialized LLMs such as BloombergGPT a 50-billion parameter model trained on a combined corpus of 363 billion tokens of financial data (news, filings, press releases, etc.) plus additional general text [241]. These specialised financial language models achieved strong performance on financial benchmarks, including sentiment, named entity recognition (e.g. identifying companies, people, ticker symbols), and question-answering tasks that involve financial knowledge.

LLMs can perform more complex extraction tasks, such as summarizing an SEC filing in plain language or extracting a structured event list from a news article, with fairly high accuracy by virtue of their broad knowledge and language understanding. Early studies indicate that ChatGPT (GPT-4) itself, without fine-tuning, can classify financial sentiment or answer questions about financial documents with reasonable accuracy, though sometimes lagging specialized fine-tuned models [94]. A notable trend is retrieval-augmented generation (RAG) LLMs in finance which enhance an LLM with a retrieval step over a database of financial facts, to improve its accuracy on domain-specific questions [247].

Usage surveys such as Anthropic AI's AI economic index show that the latest innovations in LLM-powered information extraction have considerably lowered the difficulty of extracting text features from unstructured data which were until very recently only the domain of expertly-designed text processing algorithms whether done through expert dictionaries or from the no less specialised ML toolkits [114]. This was mostly achieved through allowing natural-language interfaces with the language model and has led researchers and practitioners to forecast the advent of 'quantfundamental investment' : LLM-powered analysis of *fundamental* aspects of firms which before required bespoke analyst work at the scale of traditional '*quant*' or quantitative processing speed and systematisation [55].

As these large models continue to improve, they are enabling new applications such as conversational assistants for investors (answering questions about portfolio performance using textual news evidence) and automated report generation. However, challenges remain, including the need to handle up-to-date information (financial data is continuously changing) and to ensure factual accuracy and compliance in generated outputs.

However, it is important to highlight the fact that there are several critiques levelled

at the current state of LLM-powered textual information extraction mostly in terms of its' lack of interpretability and explainability.

2.2 Financial networks as representations of markets

Different disciplines contribute to the view that financial markets, and specifically financial instability can be modelled effectively through networks. As Haldane [108] and Acemoglu et al. [3] argue, the markets display a 'robust yet fragile' reaction to stress akin to ecological networks. These display surprising resilience to small but regular shocks while being vulnerable to a large 'system-wide' disturbances [77]. Descriptions of finance as a network also ties in to evolutionary arguments about how financial markets evolved as fundamentally risk-sharing institutions between economic actors [129, 152]. In this view financial markets allow goods to be provided through a network-like grid of exchanges to parts of the system which, due to a temporary low-intensity shock have an unfulfilled need [4]. Markets are then liquid precisely because of their network structure. However, during high-intensity shocks when too many components of the exchange network stumble at once, the whole system shuts down, liquidity dries up as has been observed during system wide events like the 2008 crisis [108].

Beyond these sociological, political and anthropological considerations, financial networks also emerged from a different strand of literature around the physical and mathematical properties of financial correlation matrices [180, 158, 34]. This field viewed financial markets as mathematical objects where interconnections result from mathematical processes which can be approximated with functions [174, 192]. One example of this view justified the existence of financial networks by showing that the connection patterns differed from random matrix theory [140].

The notion of instability in financial markets is therefore intimately connected with an evolving network of financial assets. Recent years have shown how the correlated structure of financial assets amplifies systemic risk and helps the diffusion of catastrophic market failures [249]. Choosing how to extract and represent the links between assets and how they evolve over time is an complex but important question in using network science to understand financial stability.

Constructing a financial network involves defining what constitutes a node and a link for the problem at hand, and then using data or models to establish the network edges. To properly survey the field, this work surveys the major construction methodologies, from straightforward data-driven approaches to more advanced filtering and inference techniques.

2.2.1 Extraction of networks from relational data

The simplest case is when the financial network structure is directly observable from data artifacts - for instance in the form of direct relational databases. If one has access to records of bilateral exposures (e.g. loans between every pair of banks), or ownership stakes, or transaction flows, then building the network is a matter of reading the adjacency matrix from the data [91]. These tools are usually used in addition or in contrast to correlation, copula or other similarity-based network elicitation strategies. They offer information on dimensions of firm-to-firm similarity which cannot easily be captured by a similarity distance such as price correlations due to the aforementioned limitations on the ability for price changes to encode all existing information efficiently [227].

The network structure is recorded through explicit financial interactions. In such cases, the adjacency matrix of the network can be built directly from transaction-level or exposure-level data [123, 72]. Typical examples include interbank lending or credit exposure networks, where nodes are financial institutions and weighted edges represent outstanding bilateral loans or derivative positions [229, 16]. Many such datasets are maintained by central banks and supervisory authorities (such as the Bank of England, the European Central Bank, or the Federal Reserve), which collect large-exposure reports from regulated entities. These datasets allow the construction of empirically grounded networks capturing the true structure of financial interconnectedness.

Other canonical examples include corporate ownership and control networks, derived from shareholding databases [231], and trade and payment networks, extracted from cross-border transaction data or national payment system records [46, 122]. In these cases, the data provide a direct mapping of who transacts with whom, allowing an unambiguous and verifiable network representation of financial or economic relationships. Another source of structured direct relationship information is knowledge graphs containing information about individual companies [164]. These include common geographic locations, common ownership structures and common activity sectors [90, 125]. Recent work has also highlighted the importance of a hitherto under explored relational database at explaining common price movements: the Analyst Matrix [182, 160]. This is the network formed by simultaneous coverage of two firms by the same financial analyst [160]. Firms covered jointly more frequently by analysts show elevated comovement [160].

Nevertheless, direct-data construction methods are often constrained by data availability and confidentiality. Full bilateral exposures or transaction flows are rarely disclosed publicly, leading to incomplete network information. To mitigate this, researchers have developed network reconstruction techniques that estimate missing links from partial aggregate data (such as each institution's total interbank assets and liabilities).

The key trade-off of direct data based construction therefore lies between performance or accuracy on one hand and data accessibility and extraction cost. While directly observed

networks are the most faithful representations of financial linkages, they are expensive to collect, highly sensitive, and subject to strict confidentiality. Indeed, these networks are expensive to build and usually require domain experts to validate and maintain them up to date especially in the case of supply chain graphs [187]. As noted by Anand et al. [16], interbank exposure data are often available only to regulators, and even then under significant aggregation and anonymization.

This trade off drives the research community to complement such approaches with model-based or hybrid techniques, which we detail below combining limited empirical observations with statistical inference to recover the underlying topology of financial relations. Furthermore, we can note a recent increase in knowledge-graph inference techniques in recent years owing to the upsurge in literature around extracting relational information from the latent space of large language models [187]. This trend could ultimately compensate the time and resource limitations of relational-database-based network building strategies.

2.2.2 Extraction of networks from similarity measures

If a researcher does not have data artifacts which directly serve as edges between nodes in a network, these edges can be constructed by choosing some metric of similarity between the nodes and operating some constraint or filtering to obtain a reasonable representation of the underlying relationship.

An intuitive approach is to use the price data [118]. The earliest approaches have looked at correlation matrices and their evolutionary dynamics over time [138, 49, 195]. These correlations measure the degree of similarity between two companies' stock price returns and researchers infer from it the degree to which both companies are exposed to market events [161]. Statistical analysis of correlation matrices have indicated that correlations tend to increase sharply during crises [227]. As correlation matrices are complete matrices, they require filtering to turn them into a useable network representation. A common approach is thresholding a weighted complete graph. Given a correlation matrix of N assets, one can start with a complete graph of N nodes weighted by pairwise correlation and then prune edges below a chosen threshold (or, equivalently, retain only edges above a high threshold) [158]. Alternatively, statistical significance can define the threshold, for example by retaining only correlations with p-values below 0.01. This procedure yields a binary adjacency matrix where an entry of 1 indicates a statistically significant or strong relationship between two assets.

However, as pointed out by Mantegna [158], these correlations are unstable, noisy, backward looking and can only represent direct linear relationships between assets [209]. Moreover, because Pearson correlation is not robust to outliers, one single high variance data point causes a market to be more correlated than the implied correlation of all the

other data points [153]. This has led researchers to also considered other similarity metrics of return time series beyond correlations such as dynamic time warping (DTW) [236], mutual information [95] or Granger causality [42].

Additionally, thresholding presents drawbacks: if the threshold is too high, the resulting network may fragment into many disconnected components; if it is low, the network becomes too dense to be interpretable. Moreover, the choice of threshold is largely arbitrary and may not reflect the true multivariate dependency structure of the data [227]. As a result, threshold networks often either omit important weak ties that maintain overall connectivity or include excessive links that obscure the network's structural information content.

To overcome these limitations, more sophisticated information filtering methods have been proposed. The Minimum Spanning Tree (MST) method [158] automatically selects the $N-1$ strongest links that keep the graph connected, thus enforcing sparsity while preserving a single connected backbone. The MST optimizes a global criterion (minimum total distance) that favors the strongest correlations, but its tree topology necessarily omits secondary relationships that may carry relevant information. The Planar Maximally Filtered Graph (PMFG) [225, 26] extends the MST by adding back edges provided that planarity is maintained. The PMFG algorithm retains more edges than the MST allowing each node to connect with multiple neighbors and thereby capturing richer local clustering than an MST. A further development is the Triangulated Maximally Filtered Graph (TMFG) [163], which starts from a 4-clique and iteratively adds nodes by forming new tetrahedral cliques while ensuring both planarity and chordality. The TMFG additionally offers the possibility to represent multivariate similarity measures to represent how distant nodes are from each other which allow it to capture more dimensions of distance and thus potentially richer structures. Finally, the Maximally Filtered Clique Forest (MFCF) generalizes the TMFG by allowing cliques of arbitrary size and organizing them in a forest-like structure [162]. These information-filtering networks (IFNs) are generative and typically greedy, aiming to preserve locally the strongest multivariate dependencies while enforcing global sparsity and topological constraints. The outcome is a network that is locally dense - with tightly connected cliques of related assets - but globally sparse, a structure ideal for interpretability and visualization in high-dimensional financial data.

2.2.3 Extraction of networks from textual data

The increased availability of textual data due among others to the advent of social media have opened new pathways to extract relations between companies in a faster and more automated way : by counting co-mentions in news articles [33, 213]. This technique assumes that similar companies will be mentioned in same articles. Other approaches have since looked at improving co-mention based networks by adding different

linguistically relevant features such as sentiment to quantify the strength of connections, under the assumption that positive coverage and negative coverage in an article describing two companies indicates different types of links and should be aggregated differently [233]. This extraction method, can be seen as a special case of the similarity metric with filtering approach described above. However, this thesis highlights the different methodological subtleties surrounding how the similarity function can be defined (for instance using causality statements extracted from text, co-occurrence counts or even embedding documents directly) and thus discusses networks from text more in depth here.

Other components of discourse have surfaced as important methods to build networks from text : narratives linking financial entities together. 'Narrative economics' [212] has emerged as a helpful tool in modern economic and financial theory to capture underlying key topics in economic discourse and how these changes in topics affect markets [177]. An economic narrative is defined as 'popular stories that affect individual and collective economic behaviour' [212]. These narratives capture current 'topics of the day' in markets and can reveal what causal links between financial variables market agents expect [141]. By capturing these expectations, economic narratives can help detect instability provoking events. Nyman et al. [177] for instance showed how increased narrative agreement across a news corpus is correlated and predictive of market instability.

This builds a graph of relations between entities. It models narratives as logically arrayed sequences of entities; that similarly to a knowledge graph, whose precise implementation may vary from paper to paper but with a similar broad outline as in [165, 23]. Entities and relations between entities are identified using named entity recognition (NER) [12] and syntactic parsing from a corpus of texts. Entities are usually nouns and relations verbs between two nouns. Then co-reference resolution group together entities which refer to the same object like 'Apple', 'Apple Inc.' and 'AAPL'. The resulting knowledge graph can represent basic 'A' ,'acts on', 'B' relationships called triads. These basic building blocks can be chained together by looking at causal links between triads such as 'A' , acts on, 'B', because, 'B', 'does', C as seen in [141].

This approach offers the advantage of being relatively simple to compute in comparison with knowledge graph techniques and is able to capture more complex and causal statements about markets than simpler co-mention-based models.

2.3 Measuring financial network features

Once the financial networks are extracted whether from text or price data, different methods have been used in the literature to extract interpretable features from this data form [8].

There is a lack of consensus in the literature when investigating the link between network properties and the stability of the markets they describe. Initial observations pointed to more densely connected networks - with higher average degrees- being more robust than sparse ones as any given shock may be dispersed across more companies and thus minimising the average 'risk per company' [14]. However, other studies show that when factoring in the potential company-level -ie node-level- adverse reactions when being exposed to risk, such as reducing lending when a debtor node goes bankrupt, increased density of connections could bring more instability, not less [35]. This lack of consensus on the effect of very simple network statistics such as average degree has prompted a strand of literature which investigates more complex topological features of the financial networks. These features include shorter diameters [158], higher cluster density [158, 180, 227], or lower Ricci curvature [203, 204] which are shown to be positively linked to market instability measures such as higher volatility.

Other more node-focused measures such as node centrality can also help expose systemically important nodes whose impairment or disruption may cause a crisis. DebtRank by Battiston et al. [36] is an eigenvector-centrality based measure that ranks companies by how central they are in a lending network, and thus how much they contribute to risk, were they to fail. RiskRank by Mezei and Sarlin [168] calculates an estimate for systemic risk from aggregating node-level statistics.

The variety of application directions of these measures, alongside the relative computational ease of there calculation makes them useful tools in the study of financial markets.

2.4 Graph machine learning applied to financial networks

As an alternative to traditional network statistical descriptors, graph neural networks have emerged as an appealing option to extract signal from edge and node data directly instead of requiring an intermediate modelling step. Machine learning is a very popular tool in finance thanks to its' ability to extract signal with very few priors, it can learn in a bottom-up manner directly from the data. Graph machine learning extends this idea by applying it not only to tabular data but to nodes and edges at once. Graph machine learning in financial networks have mostly been applied to forecasting changes in financial variables [201]. Li et al. [146] use a graph neural network to forecast overnight stock movements. Lin and Sun [150] have use a graph attention network to leverage a heterogeneous graph for stock return prediction. These models are helpful as they are able to learn node level data such as company features and edge-level data such as connection structures jointly [62]. Literature reviews such as Wang et al. [237], also highlight different types of graph

neural networks applied to financial markets. Different models are built by selecting 3 main criteria. This first is their node feature-processing strategy (whether they are taking in the node features directly or using some form of recursive network such as an LSTM or embedding textual features using a transformer for example). The second dimension consists in the different ways to model of node-to-node interactions as discussed in 2.2.1 and 2.2.3. The type of selected graph conditions a corresponding mechanism for message passing. Homogeneous undirected graphs correspond to a straightforward case where different techniques such as Graph Convolution in the case of the GCN [134], Attention in the case of architectures like the GAT [59] or even neighbourhood sampling approaches such as GraphSage [111] have been applied. However, other relation types such as the links in knowledge graphs may require heterogeneous message-passing strategies such as multirelational graph attention as can be seen in the complex task of fraud-detection on multirelational graphs by Wang et al. [235]. Similarly, bipartite graphs also require special treatment such as attention networks [248]. Lastly, Graph machine learning approaches to financial markets differ in the types of tasks they are built to solve. However, despite their wide usage Wang et al. [237] point to the lack of unsupervised models in graph neural network applications to the stock market that can learn representations of market states instead of only directly forecasting them. Some exceptions leverage unsupervised graph learning models such as the graph auto-encoder (GAE) [135]. Turner [228] for instance uses a GAE for unsupervised clustering of a market. As network data increase in popularity and as LLMs and knowledge graphs become increasingly integrated, the applicability of network-specific processing techniques to financial markets will tend to increase.

Chapter 3

Understanding financial market stability using graph machine learning

Understanding stock market instability is a key question in financial management as practitioners seek to forecast breakdowns in asset co-movements which expose portfolios to rapid and devastating collapses in value. The structure of these co-movements can be described as a graph where companies are represented by nodes and edges capture correlations between their price movements. Learning a timely indicator of co-movement breakdowns (manifested as modifications in the graph structure) is central in understanding both financial stability and volatility forecasting. We propose to use the edge-reconstruction accuracy of a graph auto-encoder (GAE) as an indicator for how spatially homogeneous connections between assets are, which, based on financial network literature, we use as a proxy to infer market volatility. Our experiments on the S&P 500 over the 2015-2022 period show that higher GAE reconstruction error values are correlated with higher volatility. We also show that out-of-sample autoregressive modeling of volatility is improved by the addition of the proposed measure. Our paper contributes to the literature of machine learning in finance particularly in the context of understanding stock market instability.

3.1 Introduction

Financial market instability has long interested investors, as, during financial collapses, previously uncorrelated companies collapse in value together, which exposes the market to a systemic level of risk that goes beyond the level of the company and requires a more global outlook [17]. Understanding this market-wide systemic risk driven by asset co-movement requires framing connections between firms as formed by evolving 'complex adaptive networks' [108]. This turns instability detection into a network analysis task. These networks are usually constructed by transforming and filtering correlation matrices of price returns [158]. They capture key structures of the correlation matrices, and, at the same time they avoid potentially noisy features derived from correlations directly [209]. Topological features of the resulting financial graphs, such as shorter diameters [158], higher cluster density [158, 180, 227], or lower Ricci curvature [203, 204], are shown to be positively linked to market instability measures such as higher volatility.

This suggests that higher market instability results from breakdowns to previously homogeneous patterns of network connectivity . Homogeneity is defined here as regular connection patterns across the entire network [244]. In financial networks, homogeneous patterns are connections between companies of similar industry and with similar returns [138]. These homogeneity-disturbing shocks arise from news announcements about financial events or asset bubbles and affect the ways in which the assets co-move by modifying the way investors form opinions about how companies are related and how they will co-evolve [54, 148, 195]. From this perspective, market instability can be seen as modifying the underlying mechanism through which inter-company connectivity structures appear. Kukreti et al. [138] points out that heterogeneously connected graphs indicates higher instability, as any shock could travel further across the entire market and not be limited to a single sector [22, 78]. Heterogeneity is characterised by high connectivity between companies which differ in industry sector, market capitalisation or long-run returns. Conversely, more homogeneous periods when strong connections only exist between similar assets would indicate lower risk, as any shock would propagate at most throughout that asset's immediate neighbourhood in the financial graph and not the entire market [203, 118].

However, the "network-only" approaches mentioned above may not fully capture homogeneity of connections between different types of nodes. Indeed, once networks are constructed, they do not make further use of any form of node features. This means they implicitly assume that all information about company returns is captured through their connectivity, and discount information from the underlying price returns which traditional volatility forecasting models like multivariate GARCH [63] or HEAVY [176] successfully used in forecasting volatility. Thus, a measure which accounts for both the

network and node features is lacking. This leads to the open question of how to measure market instability using financial networks with company characteristics such as returns.

The graph machine learning literature has provided tools for the joint learning of these features [62]. However, their applications to financial graphs remain mostly limited to return forecasting, with very few works looking at volatility forecasting directly [201]. Wang et al. [237] point to the lack of unsupervised models in graph neural network applications to the stock market that can learn representations of market states instead of only directly forecasting them. We propose a graph auto-encoder (GAE) [134] to capture both connectivity and node feature information as part of a edge-reconstruction task. Good GAE performance would indicate good node embeddings and a more stable and homogeneous market as homogeneity in edge connectivity helps reconstruction performance. Assuming the GAE is well trained, it should be able to group companies of similar returns together. If, once trained, the GAE performs poorly in reconstructing structures of financial graphs of future time steps, we argue this indicates a shift in the way connections are formed in the future graph. We interpret this shift as an increased instability in the financial network. Indeed, if the GAE performs worse, then the future graph has heterogeneous connections between companies of different returns and different sectors. A worse reconstruction performance then reflects the definition of a non-homogeneous, unstable graph from literature [138, 180, 244].

3.2 Problem setting and methodology

3.2.1 Problem setting

Consider a financial network represented by a graph $\mathcal{G} = \{\mathcal{E}, \mathcal{V}\}$, with nodes representing firms and edges the similarities or correlations between firm prices. In addition, consider a matrix $\mathbf{X}$ where the u-th column of $\mathbf{X}^T$, $\mathbf{x}_u$, is a vector of features that characterise the firm u (e.g., its stock prices at given time intervals). Our goal is to develop a measure of the instability of the financial market given $\mathcal{G}$ and $\mathbf{X}$.

Recent works have analysed market instability by investigating the connectivity of $\mathcal{G}$ alone [180, 203]. To take into account further information in $\mathbf{X}$, we utilise the GAE proposed in [134] which learns vector representations of nodes using both $\mathcal{G}$ and $\mathbf{X}$. The auto-encoder is trained on an edge reconstruction task described in Sec. 3.2.3 where we use a random subset of edges for training and the remaining edges for testing. A good performance on this task would indicate a more homogeneous pattern of edge formation across the graph.

With this understanding, we frame the problem of measuring market instability as one that quantifies heterogeneity in connectivity patterns between different areas of $\mathcal{G}$. As

shown in Kukreti et al. [138], homogeneous connectivity patterns are usually observed in low volatility periods where companies in related sectors are often exposed to similar fluctuations and overall more connected to each other than to companies outside their industries. On the other hand, more volatile markets are characterised by highly perturbed correlation structures leading to areas of the graph with highly different connectivity patterns. We therefore treat the generalisation ability of the edge reconstruction task above as a proxy for market instability. A high reconstruction performance on average with a randomly chosen split of training and test sets would indicate more stable market, while a low performance suggests the opposite due to heterogeneous connectivity patterns across the graph.

3.2.2 Data

We provide a practical context of the problem described above. Our data set consists of 7 years of stock market price data for stocks in the S&P 500 at the minute-level frequency from 2014 to 2021[1]. From these price movements which we define as $p(n,t)$ where the nt-th entry is the price for company n at t, we calculate a log return matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$ of N companies over T time periods at frequency Δt :

$$r(t,n) = \log(p(t,n)) - \log(p(t-\Delta t, n)) \tag{3.1}$$

Based on this return matrix, we calculate a Pearson correlation matrix based on a rolling window of length S days. This correlation matrix measures for all stock return pairs the covariance normalised by the square root of the product of the variances. It represents the strength of linear relationships between all the stock pairs. Given a return matrix of of N companies over T, this gives us $T - S$ correlation matrices in total. Following Kukreti et al. [138], we define an adjacency matrix $\mathbf{A}$ as $A_{uv} = 1$ if the (u, v) return pair has a correlation higher than 0.7 and otherwise we set A_{uv} to 0. The 0.7 threshold is selected to only keep strongly connected pairs and discount correlations of insufficient strengths [53][138]. This approach was selected as it is straightforward to interpret the resulting network as the subset of edges above 0.7 correlation. However, it is important to note this edge threshold method suffers from significant drawbacks which we present here and use to highlight and discuss our results and their limitations. Firstly, this approach has a tendency to produce either overly connected or sparse graphs, especially in financial networks as the correlation coefficient is highly fragile to large volatility data points, making the total number of connections potentially affected by the volatility we are attempting to measure. Secondly, threshold-based filters of correlation matrices do not

[1]The dataset is obtained from the EOD intra-day data API: `https://eodhistoricaldata.com/financial-apis/intra-day-historical-data-api/`. We had to drop some companies as they had no available data during the whole period leaving us with 401 companies.

preserve other network properties than direct one-to-one linkages which ignores weakly correlated but topologically significant structures.

We calculate the return volatility (RV) that is used throughout the following sections as a proxy for instability according to the formula set out in Eq. (3.2). It is a measure of the variance of the returns of the average price returns over Δt periods:

$$\mathrm{RV}_t^{t+\Delta t} = \sum_t^{t+\Delta t} (r(t,n)^2) \tag{3.2}$$

Because of the power law distribution of market volatility with orders of magnitude larger swings on high-volatility days, volatility is often measured as the natural logarithm of RV and termed log-RV.

We summarise the way we construct the financial network in Fig. (3.1) where a series of stock prices is transformed into a return matrix. From those returns, we calculate the correlation matrix as described above. Lastly, through a threshold applied to that correlation matrix, we obtain a network of connected companies.

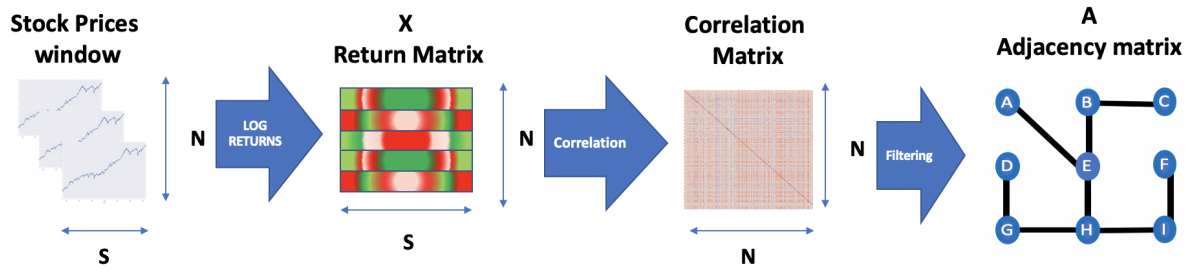


Figure 3.1: Pipeline generating networks from N returns over a window of time S

3.2.3 Proposed methodology

Based on the problem setting and data description in the previous two sections, we now formally describe our methodology. Given the financial graph $\mathcal{G}$ with adjacency matrix $\mathbf{A}$ and node features $\mathbf{X}$ defined above, we propose to use the GAE [134] as the edge reconstruction model whose generalisation ability is used to approximate market instability.

The GAE is an unsupervised representation learning model with an encoder and a decoder. The encoder is a function defined in Eq. (3.3). It describes an operation mapping the input $\mathbf{X}$ to the latent embedding $\mathbf{Z}$ via a 2-layer graph convolutional network (GCN) encoder [135]. Once the encoder generates the embedding $\mathbf{Z}$, the decoder in equation Eq. (3.4) aims to reconstruct the initial adjacency matrix. The embedding $\mathbf{Z}$ is of "good quality" if the reconstruction $\hat{\mathbf{A}}$ is close to the initial adjacency matrix. The GAE is

trained by minimising a binary cross-entropy (BCE) loss defined in Eq. (3.5) on training edges $\mathbf{A}_{uv}$:

$$\mathbf{Z} = \text{GCN}(\mathbf{A}, \mathbf{X}) = \tilde{\mathbf{A}}\text{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W_0})\mathbf{W_1} \tag{3.3}$$

$$\forall (u, v) \in \mathcal{V} \times \mathcal{V}, \hat{\mathbf{A}} = \sigma(z_u {z_v}^T) = \frac{1}{(1 + e^{-z_u z_v{}^T})} \tag{3.4}$$

$$Loss(A_{uv}, \hat{A}_{uv}) = -A_{uv} * \log(\hat{A}_{uv}) - (1 - A_{uv}) * \log(1 - \hat{A}_{uv}) \tag{3.5}$$

With the following definitions :

- $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the normalised adjacency matrix of the graph. Where $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$[2] is the diagonal matrix of the graph $\mathcal{G}$ where the values on the diagonal $(d_1, \ldots, d_n)$ correspond to the degree of the node.
- $\text{ReLU}(\mathbf{X}) = max(X_{ij}, 0)$ is the non-linear activation function of the GCN. This function is applied element-wise to each entry ij in the matrix.
- $\mathbf{W_0}$ and $\mathbf{W_1}$ are learnable weight matrices, $\mathbf{W_0} \in \mathbb{R}^{w_{in} \times w_{hid}}$ and $\mathbf{W_1} \in \mathbb{R}^{w_{hid} \times w_{out}}$. With w_{hid} and w_{out} are chosen by hyperparameter tuning described in the appendix Table 9.1 and w_{in} is T the length of the feature vector for a given firm.

The end-to-end training is done using gradient descent for which we selected a standard Adam algorithm [133]. Fig. 3.2 below summarises how the GAE operates.

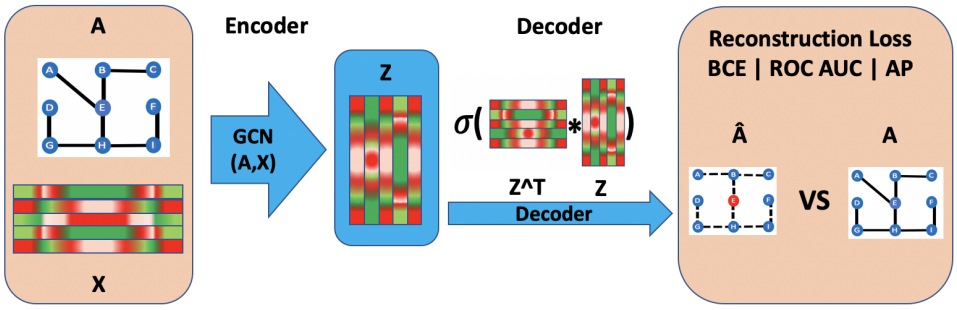


Figure 3.2: Summary of GAE operations

We now describe how we make use of the GAE model to derive a measure of market instability. We first extract the returns data $\mathbf{X}$ from a window of length $S = 20$ days. From those returns, we generate the corresponding market graph $\mathcal{G}_t$ with t denoting the last day in the time window using the methodology described in 3.2.2. We selected a span of 20 days as this represents one month of trading data, which is long enough to capture relations between companies without too much noise from unstable correlations in shorter

[2] $\text{diag}(d_1, \ldots, d_n)$ denotes the diagonal matrix with diagonal elements $d_1, \ldots, d_n$.

windows [138]. Shifting the window by one day, we generate the next day market graph $\mathcal{G}_{t+1}$.

We trained and validated the GAE on $\mathcal{G}_t$, and test its performance by applying the trained model to reconstruct edges in $\mathcal{G}_{t+1}$, the market graph of next day $t+1$. We assume the GAE is trained on a period of stable market, which means a graph with spatially homogeneous connectivity structures. Following the reasoning in Section 3.2.1, once a GAE is trained, it can reconstruct the adjacency matrix of an unobserved graph in proportion to how spatially homogeneous the unobserved graph is [134]. Consequently, any drop of the GAE testing accuracy on $\mathcal{G}_{t+1}$ can be viewed as an increase in graph heterogeneity. This increased heterogeneity measures how different the edge formation mechanism in $\mathcal{G}_{t+1}$ is different from $\mathcal{G}_t$. We interpret this change in the edge formation mechanism as resulting from shocks to investor opinion such as those from news announcements [148, 195]. These shocks also form the basis of market instability, and thus we interpret this change in heterogeneity as correlating with increased volatility on day $t+1$. The reconstruction accuracy measure which we use in our paper to signify the next day edge-reconstruction accuracy for any given test graph $\mathcal{G}_{t+1}$ is area under receiver operating characteristic curve ($\mathbf{AUROC}_{t+1}$). The process is summarised in Fig. 3.3 below, where the shift between data at time point t and $t+1$ is shown as a shift of the two time series $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$, with the former being used to train the GAE and the latter for testing.

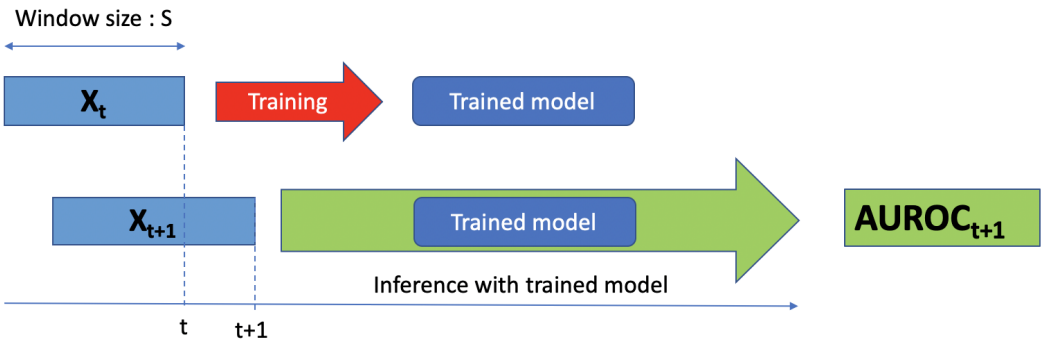


Figure 3.3: Process of generating $\mathbf{AUROC}_{t+1}$ from model trained at t

To asses the performance of the proposed measure, we will first look at correlations between $\mathbf{AUROC}_{t+1}$ (the next day GAE reconstruction) and RV_t^{t+1} (the next day return volatility). This comparison is done by observing the variation of the time series, observing the kernel density estimation (KDE) plots and the Spearman rank correlation of $\mathbf{AUROC}_{t+1}$ and RV_t^{t+1}. We chose the Spearman rank correlation as it is less vulnerable to outliers. High correlation between $\mathbf{AUROC}_{t+1}$ and RV_t^{t+1} would indicate that changes in $\mathcal{G}_{t+1}$'s homogeneity are in sync with market instability.

As a second step, we then test the usefulness of $\mathbf{AUROC}_{t+1}$ for the forecasting of $\log(\mathrm{RV})$ at the following $t+2$ time step ($\log(\mathrm{RV}_{t+1}^{t+2})$). We perform this forecasting with

the heterogeneous autoregressive (HAR) model for log-RV forecasting by Corsi [74]. The HAR is a strong and industry recognised benchmark model for volatility forecasting Lim et al. [148]. It uses auto-regressive components of lower frequency past volatility to predict future volatility. Lower frequency volatility measurements are weekly and monthly volatility. Eq. 3.6 shows how the the HAR(RV) operates, it learns a parameter β_n for each previous volatility measure from the past time periods. We look at an 1-hour frequency of forecasting as it is a frequency of interest for industry practitioners [17, 180, 148].

$$HAR(\log(\mathrm{RV}_t^{t+1})) = \beta_1 \log(\mathrm{RV}_t^{t+1}) + \beta_2 \log(\mathrm{RV}_{t-6}^{t+1}) \tag{3.6}$$

Using the result in Eq. 3.6, we use the HAR-RV model for our out-of-sample volatility forecasting task to calculate the 2 hour ahead volatility $\log(\mathrm{RV}_{t+1}^{t+2})$ as a function of the autoregressive model applied on all the previous time steps' volatility information.

$$\log(\mathrm{RV}_{t+1}^{t+2}) = HAR(\log(\mathrm{RV}_t^{t+1})) \tag{3.7}$$

To assess the usefulness of our GAE-based indicator, we run these same regressions again under an 'improved' setting as described in Eq. 3.8, by adding the $\mathbf{AUROC}_{t+1}$ feature as an additional explanatory variable on top of averaged lagged volatility captured by the traditional HAR-RV. We compare the R^2 of the baseline models with the R^2 of the 'improved' models.

$$\log(\mathrm{RV}_{t+1}^{t+2}) \sim HAR(\log(\mathrm{RV}_t^{t+1})) + \beta_3 \mathbf{AUROC}_{t+1} \tag{3.8}$$

We calculate the R^2 of this estimation for a linear regression, a XGBoost tree and an MLP as well as the R^2 of the estimation to which we add the $\mathbf{AUROC}_{t+1}$ as in Eq. 3.8.

To test the hypothesis that $\mathbf{AUROC}_{t+1}$ encodes useful information, we perform a one sided t-test of of the difference between the R^2 of the regular and 'improved' approaches by testing whether the R^2 of the proposed method is higher than the regular methods' R^2. The p-value displayed on the rightmost column corresponds to a bootstrapped estimate of the significance of the difference between the R^2 of the prediction with and without the $\mathbf{AUROC}_{t+1}$, with the H_0 null hypothesis being no difference and the H_1 being a positive effect. An increase in the R^2 of the forecasting model with the GAE-based measure added would suggest that the GAE encodes useful information for out-of-sample stability forecasting.

3.3 Results

3.3.1 Correlation between $\mathbf{AUROC}_{t+1}$ and volatility

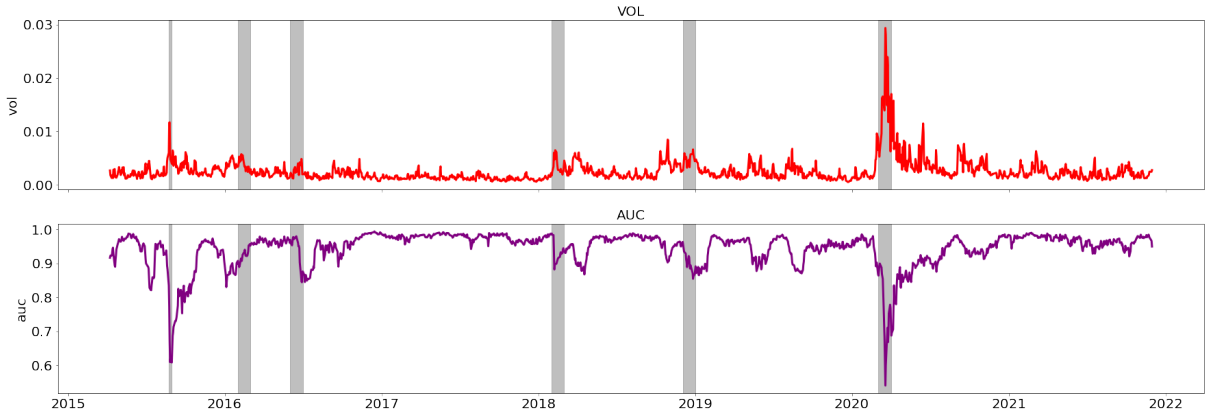


Figure 3.4: Time series of volatility vs $\mathbf{AUROC}_{t+1}$

Figure 3.4 displays the evolution over time of our indicator $\mathbf{AUROC}_{t+1}$ in purple and in red the volatility of our market as described in 3.2. When comparing the $\mathbf{AUROC}_{t+1}$ and market volatility, we remark that $\mathbf{AUROC}_{t+1}$ appears to co-move with volatility. $\mathbf{AUROC}_{t+1}$ is high in periods of relative market stability where volatility is lower and sharply fall during downturns and high-volatility markets where volatility is high. As GAE model performance is linked to the homogeneity of edge formation across a graph, the model shows lower performance during high-volatility periods. This suggests a more volatile market at $t+1$ correlates to heterogeneous edge formation in the market graph $\mathcal{G}_{t+1}$ which the GAE cannot learn as easily as those in more stable settings confirming the expectation laid out in the literature and our problem setting.

The six grey bands highlight periods of market instability, either due to so called 'flash crashes' as in late 2015 or those due to fundamental shifts in market stability such as the 2020 pandemic stock market collapse. These further allow us to remark that the $\mathbf{AUROC}_{t+1}$ seem to dip slightly before a volatility upswing, suggesting a useful application of these measures in forecasting, which is further investigated in the regression analysis results presented in Sec. 3.3.2.

The plots above show the kernel density distribution and Spearman rank correlation between log volatility and the $\mathbf{AUROC}_{t+1}$. The relationship between log volatility and the $\mathbf{AUROC}_{t+1}$ and the correlation in Fig. 3.5 displays an inverse relation between market stability and the GAE's performance. This provides evidence in support of the expectation that the GAE produces a good measure for market instability as it encodes graph homogeneity, and drops in $\mathbf{AUROC}_{t+1}$ are correlated with increased volatility.

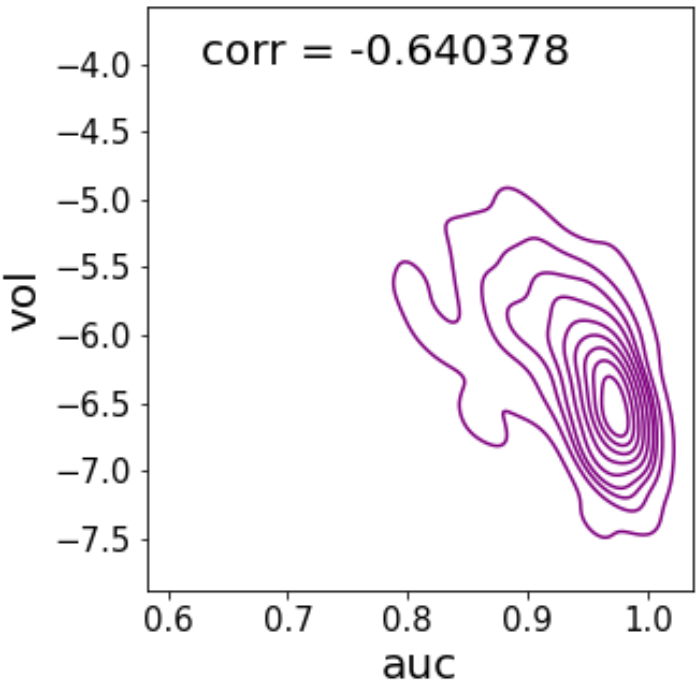


Figure 3.5: KDE of $\mathbf{AUROC}_{t+1}$ compared to volatility

3.3.2 Next-day volatility forecasting

In this section, we move on from looking at how well the $\mathbf{AUROC}_{t+1}$ represents market instability. We shift our focus to looking at the usefulness of the $\mathbf{AUROC}_{t+1}$ to predict $\log(\mathrm{RV}_{t+1}^{t+2})$ the following day's volatility.

Model	R^2 with $\mathbf{AUROC}_{t+1}$	R^2 without $\mathbf{AUROC}_{t+1}$	p-value of difference
Linear	**0.360***	0.107	0.00
Tree	**0.357***	0.310	0.043
MLP	**0.380***	0.317	0.003

Note : * and **bold font** indicate p-value < 0.05.

Table 3.1: Results of log-RV forecasting at the 1-hour frequency

Table 3.1 presents the results of the log-RV forecasting task at time period $t + 2$ defined in the Methods section (see Eq. 3.6 and Eq. 3.8). The p-value corresponds to a bootstrapped estimate of the significance of the difference between the R^2 of the prediction in a one-sided t-test test, i.e., whether the R^2 of the machine learning forecasting method used with the added $\mathbf{AUROC}_{t+1}$ is significantly larger than the R^2 without $\mathbf{AUROC}_{t+1}$ as described in Section 3.2.3. We test three different machine learning forecasting strategies this way : a linear regression referred to as the linear model, a gradient boosted tree

referred to as Tree and a multi-layer perceptron referred to as the MLP model. The results show a statistically significant positive effect of the $\mathbf{AUROC}_{t+1}$ on the forecasting of hourly volatility at time period $t+2$. The result aligns with the hypothesis of the problem setting and observations of the previous sections.

3.4 Discussion and Conclusion

Our research investigated the link between the encoding ability of an unsupervised node-embedding model and market instability. The results above support our hypothesis that the the edge-reconstruction accuracy of a GAE would serve as a good proxy to detect changes in the homogeneity of connectivity across a financial graph and thus serve as a good proxy for market volatility. The GAE's encoding of market returns and connectivity at time step $t+1$ is correlated with market instability at the same time step as shown in Section 3.3.1 and useful in forecasting out-of-sample volatility as demonstrated in the regression analysis of Section 3.3.2. The novelty of our contribution lies in underlining the importance of looking at node features as well as the adjacency matrix when deriving stability indicators from financial networks. Moreover, we extended the literature on the use of graph neural networks in finance by successfully applying an unsupervised methodology to generating good-quality latent node representations which encode financial states.

We should acknowledge the limitations of the presented methodology. Firstly, as we set out in Section 3.2.3, correlation matrix-based networks have limitations when it comes to encoding economically relevant relationships as highlighted by [226]. Further work should explore alternative pathways to representing the relationships between stocks, beyond correlations to mathematical objects which represent multiple dependencies at once such as copulas. We could also extend the work above by representing stock dependence using other types of information filtering networks such as planar-maximally filtered graphs to go beyond thresholded correlation networks which may miss 'critical dependencies' [25]. Secondly, we can observe that the strong threshold of 0.7 chosen to filter the network creates large changes in the number of edges before and during the volatility spikes. In the table presented in the Appendix 9, we measure the proportion of edges over time in periods of high and low volatility. We find a strong link between periods of high volatility and the proportion of edges in the adjacency matrix of the graph, the variations in edge structure and edge frequencies over time especially in periods of strong volatility therefore may also drive part of the variability in the $\mathbf{AUROC}_{t+1}$ as it makes the edge reconstruction task much harder by massively increasing the number of edges. This suggests that, we could in future work relax the threshold hyperparameter we use to build our network in order to further validate our results by considering other thresholds, thus reducing the variability

between high and low correlation time points.

Our approach is flexible and can be applied to other financial networks used in the literature to describe inter firm relations. Thus we aim to apply this approach to graphs constructed from news corpuses [118] [233], supply chains [164] or knowledge graphs [112]. These other types of graphs capture qualitatively different connections than correlations. That allows them to go beyond the statistical correlations we use to build the graph in this paper. This would allow us to learn instability-predictive features derived from other types of co-movement relations, and in turn refine the volatility forecasting. This same flexibility would allow us to study other edge formation structures which have seen intense interest from the standpoint of propagation of systemic risk such as inter-banking markets [3] or national stock markets indices.

We argue that by showing that homogeneity of graph connections is an important feature in market stability, this work also opens future areas for research. With that in mind, we discuss a set of extensions that address the limitations in our solution. The first proposed extension addresses a limitation of the current implementation related to static embeddings: we are currently learning node embeddings for a single static graph and testing the accuracy of the reconstruction on another static graph. In future work, we would like to explore a dynamic feature to the graph auto-encoder as in DynGAE [157]. Another extension stems from the agreement between our unsupervised method using graph features and previously observed good performance of non-graph auto-encoder models in finance which attempted to reconstruct only returns [148]. We propose extending the current implementation to reconstruct both adjacency and returns such as in GALA architecture [189].

These proposed research suggestions serve to highlight the hope that our contributions will demonstrate the usefulness of considering node information when assessing systemic risk as well as provide a helpful methodology to the community using network science tools to study systemic risk in a dynamic way.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	
Publication Status	X Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Gorduza, D., Zohren, S., and Dong, X. (2025). Understanding stock market instability via graph auto-encoders. EPJ Data Science, 14(1):13.

Student Confirmation

Student Name:	Dragos Gorduza		
Contribution to the Paper	Wrote manuscript, performed data collection and analysis, interpreted results		
Signature		Date	30 August 2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Xiaowen Dong, Associate Professor		
Supervisor comments I can confirm Dragos has made significant contribution to the paper above.			
Signature	Xiaowen Dong	Date	07/10/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Extracting Alpha from Financial Analyst Networks

We investigate the effectiveness of a momentum trading signal based on the coverage network of financial analysts. This signal builds on the key information-brokerage role financial sell-side analysts play in modern stock markets. The baskets of stocks covered by each analyst can be used to construct a network between firms whose edge weights represent the number of analysts jointly covering both firms. Although the link between financial analysts coverage and co-movement of firms' stock prices has been investigated in the literature, little effort has been made to systematically learn the most effective combination of signals from firms covered jointly by analysts in order to benefit from any spillover effect. To fill this gap, we build a trading strategy which leverages the analyst coverage network using a graph attention network. More specifically, our model learns to aggregate information from individual firm features and signals from neighbouring firms in a node-level forecasting task. We develop a portfolio based on those predictions which we demonstrate to exhibit an annualized returns of 29.44% and a Sharpe ratio of 4.06 substantially outperforming market baselines and existing graph machine learning based frameworks. We further investigate the performance and robustness of this strategy through extensive empirical analysis. Our paper represents one of the first attempts in using graph machine learning to extract actionable knowledge from the analyst coverage network for practical financial applications.

4.1 Introduction

Financial analysts play an important information-processing role in modern financial markets. They provide signals about firm health, and their predictions about future firm outcomes have traditionally been regarded as a shorthand for stock pickers to easily keep their finger on the 'pulse' of the market. However, simultaneous coverage of two firms by one or more analysts is also shown to explain higher levels of return correlation between those firms' returns [13, 126, 246]. This suggests that beyond providing information, sell-side analyst coverage also has an effect on investor attention which we propose to leverage in order to build a trading strategy.

Indeed, as investors follow specific analyst reports, they tend to focus their finite attention on specific baskets of firms covered by sell-side analysts [124]. Limited investor attention has been identified as a source of investment opportunities [13] as information overload leads to delayed recognition of the impact that news affecting one company should have on other economically related companies. This slow information diffusion can result in predictable lead-lag effects between multiple firm's returns [69]. By acting as conduits for investor attention, 'analyst coverage networks' have been used to explain uneven patterns of investor attention between firms in the market [124, 126]. Analyst coverage networks refer to networks where nodes are firms and the weighted edges between two firms denote the number of analysts covering both firms in a certain time frame such as the last year or the last quarter [13]. For instance, if Analyst A covers both Google and Apple, a link with weight 1 is established between these two firms in the network. If Analyst B also covers Google and Apple, the link between the two companies will have a weight 2. These networks are interesting to financial research beyond just directing investor attention. Analyst coverage networks are shown to also identify fundamental economic linkages between firms, as analysts tend to cover related firms [13]. High analyst coverage offers strong predictive power for financial linkages [101] as well as changes in firms' fundamental information [142, 39, 110]. This motivates our interest in leveraging the analyst coverage network to build a profitable momentum-spillover strategy.

This increased co-movement between stocks linked to analyst coverage has led to initial interest in creating a 'model-free' strategy to profit from this effect [182]. This approach is easy to interpret and implement as it involves simple weighted averages of neighbour returns as indicators of future profitability. The trading signal then takes the form of averaging momentum of a firms' 1-hop neighbours in the network as an indicator for future performance. However, the neighbourhood momentum indicator - as we investigate in this work - may be too weak on its own and fail to take full advantage of the nonlinear and multi-hop relationships between a firm and its wider neighbourhood. Moreover, existing approaches mainly focus on the pre-defined network structure, and hence lack

the flexibility of adapting the strength of the relationships between firms, especially given node-level information coming from the stock price evolution of firms in the portfolio.

Lead-lag models developed in research and industry meanwhile have begun including machine learning in uncovering firm-level momentum to build trading indicators [38]. Inspired by this, we propose to leverage the additional flexibility allowed by graph machine learning models to 1) include both firm level and network level data and 2) adjust and adapt in real-time the strength of relationships from the initial analyst coverage network, in order to build a trading signals for out-of-sample returns. This allows us to profit from the rich information incorporated in the analyst network which uncovers relationships between stocks not traditionally captured by industry or correlation-based linkages. In addition, we demonstrate the potential of a multi-layer graph attention network (GAT) [59] in capturing the non-linear and multi-hop relationships in the analyst network in order to build an improved trading strategy in real time.

Our research contributes to the existing literature in several ways. First, to the best of our knowledge, our work represents one of the first attempts in using graph machine learning techniques to model the analyst coverage network and predict stock returns.Second, we demonstrate the superiority of our GAT-based approach over traditional aggregation methods, highlighting the importance of capturing complex, non-linear relationships between firms in the analyst coverage network. Third, we show the benefits of incorporating higher-frequency firm-level data into the model, enabling it to adapt more effectively to changing market conditions, thereby generating more robust trading strategies.

The outline of this paper is as follows. We first discuss the relevant literature which led to the different components of our approach. Subsequently, we formulate the trading-strategy problem as a graph machine learning question and present the model itself as well as the methodology we built for evaluating the model against a suite of industry benchmarks and ablation studies. Lastly, we present and discuss the experimental results before concluding and presenting the potential avenues for future research.

4.2 Literature review

The concept of analyst coverage networks and their potential to explain various momentum spillover effects in the stock market has captured significant attention in recent literature.

Several studies have investigated the economic and fundamental linkages between companies to better understand and capitalize on the lead-lag effect in investment strategies. The effect is evident among companies within the same industry [120, 172], firms offering similar products [70, 119], entities connected through supplier-customer chains [69, 166], firms sharing common technological innovations [39, 143], those located in the same geographic region [65, 190], companies with overlapping institutional ownership [97], and

those with the same strategic alliances [56].

In addition to the relationships above, Ali and Hirshleifer [13] introduce shared analyst coverage as an additional overarching approach to determine firm relatedness. They specifically highlight its potential to explain various previously established cross-asset momentum effects such as industry momentum, geographic momentum or customer momentum. They argue that analyst linkages are particularly adept at uncovering fundamental relationships between companies, more so than other methods for identifying firm linkages such as industry networks and correlations. Investors that overlook the analyst coverage network may be underestimating fundamental channels of shock transmission between firms thus explaining the observed momentum spillover effect between strongly connected neighbours on the analyst coverage network [182, 124]. Ali and Hirshleifer [13] also argue that shared analyst coverage could quantify the strength of company relationships more accurately than simple binary variables or sector groupings, and address the challenges faced by other methods such as difficulty in accessing complete supplier information for a given company [69].

Recent papers on shared analyst coverage have further explored its potential to unify momentum spillover effects and predict returns. Key studies include Gomes et al. [101], who investigated the role of analyst coverage networks in corporate financial policies. Yi and Guo [245] provided evidence from China on how common analyst links predict returns, while Jiang et al. [126] explored connected-firm momentum spillover in China. Oyeniyi et al. [182] discussed profiting from sell-side analysts' coverage networks, Israelsen [124] suggested that investors make correlated information processing errors which can be tracked using the analyst coverage network and help explain the link between it and excess comovement, and Martens and Sextroh [160] examined interfirm information spillovers due to overlapping analyst coverage.

While the existing literature has made significant progress in understanding the role of analyst coverage networks in explaining momentum spillover effects, there remain several gaps that our research aims to address. Previous studies have primarily focused on using simple aggregation methods, such as weighted averages of a firm's direct neighbours' momentum, to create lead-lag portfolios [13]. However, these approaches may not fully capture the complex and non-linear relationships between firms in the analyst coverage network, limiting their ability to predict future stock returns accurately. Moreover, existing methods lack the flexibility to adapt the strength of the aggregation coefficients based on higher-frequency information from the firms in the portfolio, hindering the strategies' ability to respond to rapidly changing market conditions and firm-specific events.

Our research aims to address these gaps by leveraging the power of graph machine learning techniques, specifically graph attention networks (GATs) [59], to develop a more sophisticated and adaptive approach to modeling the analyst coverage network

and predicting stock returns. The GAT's attention based architecture allows our model to assign different importance weights to each firm's neighbors in the analyst coverage network, thus making its' representation of firm to firm interactions more flexible than the static weights used in model-free approaches to building portfolios from the original analyst coverage network [182].

4.3 Methods

4.3.1 Data

Data sources

We investigate a dataset of stock prices extracted from the CRSPR/COMPUSTAT datasource hosted by WRDS covering the 2006-2022 period[1]. The dataset is composed of 495 firms of the SNP500 accross a diverse array of industries. We combine this dataset with information from Institutional Brokers' Estimate System (IBES) analyst estimates from 2006 to 2022[2]. These estimates are recorded as they are produced by each analyst covering firms in our sample over the 2006-2022 period. These forward looking predictions are made at yearly quarterly or monthly horizons.

Momentum indicators from stock prices

We calculate $r_{i,t-\Delta t}$, the log-return over Δ time periods of firm i as described in Equation 4.1. We select 5 Δ to represent several return horizons $\Delta \in [1, 21, 63, 126, 252]$ corresponding to the returns $r_{i,t-\Delta t}$ over the past 1, 21, 63, 126 and 252 days respectively:

$$r_{i,t-\Delta t} = \log\left(\frac{p_{i,t}}{p_{i,t-\Delta}}\right) \tag{4.1}$$

With $p_{i,t}$ and $p_{i,t-\Delta}$ the price of security i at times t and $t-\Delta$ respectively. Those 5 return horizons serve as 5 common measurements for firm information as described in [38].

We define 3 additional firm-level indicators following Baz et al. [38] $s_{i,t}(S, L)$ in Equation 4.2 based on varying short (S) and long (L) time scales belonging to $(S, L) \in (8, 24), (16, 48), (32, 96)$.

$$s_{i,t}(S, L) = \frac{m(i, t, S) - m(i, t, L)}{\text{std}(r_{i,t-63t})} \text{std}\left(\frac{m(i, t, S) - m(i, t, L)}{\text{std}(r_{i,t-63})}\right)^{-1} \tag{4.2}$$

[1]Compustat prices are available at : https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/annual-update/crspcompustat-merged/security-daily/

[2]IBES coverages are available at : https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/vendor-partner-ibes/

Here $m(i,t,S) = \gamma * p_{i,t} + (1-\gamma) * m(i,t-1,S)$ is the exponential weighted moving average of the price of asset i at time t with a scaling factor $\gamma = \frac{1}{S}$ and $\text{std}(r_{i,t-63t})$ is the standard deviation of prices over the past 63 days. We group these 8 indicators for each firm i into a vector $x_{i,t} \in (1,8)$. With $x_{i,t} = [r_{i,t-1t}, r_{i,t-5t}, ..., s_{i,t}(32,96)]$ of length 8. Combining each $x_{i,t}$ into a matrix with all firms, we obtain the feature matrix $\mathbf{X}_t \in (N,8)$.

Network extraction from IBES ratings

We represent the overlapping coverage portfolios of the analysts in our sample as an undirected network. Each trading day in our sample, we record the IBES estimates made by each analyst for all the 495 firms in our sample over the past 252 days. This creates an evolving bipartite analyst-firm coverage record which we transform into an evolving firm-to-firm network by counting the number of analysts that cover every pair of firms as described in Figure 4.1. The nodes of the proposed analyst coverage network represent all

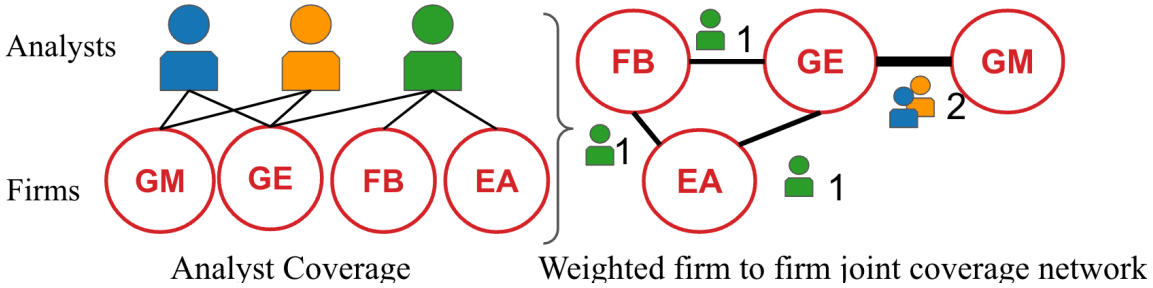


Figure 4.1: Building the analyst network

the firms in our sample. Meanwhile an edge between two nodes represents the number of analysts that cover both firms within a given look back window of 252 days. This approach to defining analyst coverage is common in the literature studying the statistical properties of analysts' choices [13]. This allows us to create at each time step an adjacency matrix $\mathbf{A}_t$ that captures the topological information contained in the analyst coverage network on day t.

4.3.2 Problem formulation

We propose to frame the task of constructing a portfolio of stocks as a classification task. We define the categorical target variable $y_{i,t+21}$ for each firm as the out-of-sample excessive returns compared to the average return of all firms in the market which takes as value either 1 (overperformance) or 0 (underperformance) as defined in Equation 4.3:

$$y_{i,t+21} = \begin{cases} 1 & \text{if } r_{i,t+21-t} > \frac{\sum_i^N r_{i,t+21-t}}{N} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

Following [182], we select a monthly target as the statistical inquiries into the analyst matrix as we assume the monthly frequency is a reasonable timeframe for the effect of analyst coverage on investors to manifest itself.

We stack each of the target variables $y_{i,t+21}$ into a target vector $\mathbf{y}_{t+21}$. We are aiming to define a function that produces the best $\hat{\mathbf{y}}_{t+21}$, the out-of-sample forecast of over and under-performance. This forecast takes the form of a predicted probability of belonging to class 1 or 0.

Following [198], we then transform $\hat{\mathbf{y}}_{t+21}$ into an investment strategy by buying the stocks corresponding to the entries in $\hat{\mathbf{y}}_{t+21}$ with the 25% highest predicted probability to over-perform. Similarly, we sell the stocks corresponding to the entries in $\hat{\mathbf{y}}_{t+21}$ with the 25% highest predicted probability to under-perform.

4.3.3 Proposed methodology

For a given trading day t in our sample, we have a feature matrix $\mathbf{X}_t$, an adjacency matrix $\mathbf{A}_t$ and a target vector $\mathbf{y}_{t+21}$. We use these in order to build the graph $G = \mathbf{A}_t, \mathbf{X}_t$ and train a GAT that learns the mapping between network information and the target following the form $GAT(\mathbf{A}_t, \mathbf{X}_t) = \mathbf{y}_{t+21}$. This triplet $\mathbf{A}_t$,$\mathbf{X}_t$ and $\mathbf{y}_{t+21}$ defines one sample, subsection 4.3.4 describes how we combine these samples to form our training-validation-testing sets.

The GAT layer takes as input a matrix of node features $\mathbf{X}_t = [x_{1t}, x_{2,t}, ..., x_{Nt}]$ where N is the number of nodes and $x_{1,t}$ is a vector of dimension D where D is the number of features in each input node. In our setup, as discussed in Section 4.3.1, the number of dimensions is 8. We describe the mechanism for transforming each node feature vector $x_{1,t}$ to their updated value $x'_{1,t}$ after applying each GAT layer.

Given learnable weights $\mathbf{W} \in \mathbb{R}^{F \times D}$ and a learnable attention vector $\mathbf{a} \in \mathbb{R}^{2F}$, we define the attention scoring function $a : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$ and a learnable attention vector $a \in \mathbb{R}^{2F}$ in Equation 4.4. It calculates the attention score between two vectors using a shared attentional mechanism $a : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$:

$$e_{ij,t} \;=\; a(\mathbf{W}\mathbf{x}_{i,t},\, \mathbf{W}\mathbf{x}_{j,t}) \;=\; \mathrm{ReLU}\big(a^\top\big[\,\mathbf{W}\mathbf{x}_{i,t} \,\|\, \mathbf{W}\mathbf{x}_{j,t}\,\big]\big)\,, \tag{4.4}$$

where $[\cdot\|\cdot]$ denotes concatenation. The GAT also normalizes the attention scores using the softmax function to make them comparable between layers.

$$\alpha_{ij,t} \;=\; \frac{\exp\big(e_{ij,t}\big)}{\sum\limits_{k \in \mathcal{N}(i)} \exp\big(e_{ik,t}\big)} \tag{4.5}$$

Having acquired the attention scores α, the GAT layer applies a parametrization using a weight matrix $\mathbf{W}$ and a non-linear transformation with the *ReLU* function in order to

obtain the updated feature representation $\mathbf{x}'_i$ for each node as described in Equation for each node 4.6.

$$\mathbf{x}'_{i,t} \;=\; \mathrm{ReLU}\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij,t}\,\mathbf{W}\mathbf{x}_{j,t}\right), \tag{4.6}$$

The model we present is set up to use an arbitrary number of GAT layers followed by a linear layer with learnable weight vector $\mathbf{w}_{\text{linear}}$ as described in Equation 4.7. We map each node's representation to a scalar target via a shared linear head. Let $\hat{\mathbf{y}}_{t+21}$ be our prediction vector.

$$\hat{\mathbf{y}}_{t+21} \;=\; \mathrm{ReLU}\big(\mathbf{X}'_t\,\mathbf{w}_{\text{linear}}\big) \tag{4.7}$$

We describe the overall pipeline of our model in Figure 4.2.

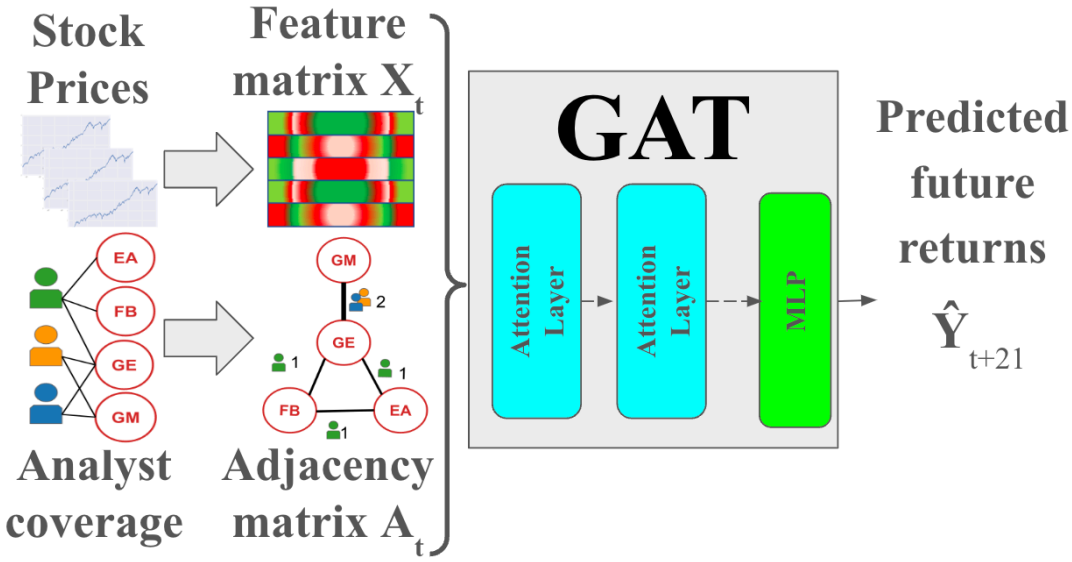


Figure 4.2: Pipeline of the proposed GAT trading model

4.3.4 Training strategy

We split our 17 year dataset into 204 1-month trading periods, at the end of each period, we retrain and validate our model. In order to obtain training, validation and testing sets we group multiple triplets $\mathbf{A}_t$,$\mathbf{X}_t$ and $\mathbf{y}_{t+21}$ together. We group the first 10 samples in each trading period (from $t = 0$ to $t = 9$) to form our training set, we group the following 10 samples (from $t = 10$ to $t = 20$) to form our validation set and lastly, we test our model on the 21st sample corresponding to t=21.

For validation, we perform hyperparameter tuning using grid search with the following settings: the learning rate $\in$ {1e-2,1e-3,1e-4}, the layer size $\in$ {64,128}, the number of

layers ∈ {1,2}, the weight decay regularization ∈ {1e-4,1e-5,1e-6}, and Attention heads ∈ {2,8}. The result of the hyperparameter tuning is described in the Appendix Table 10.1.

4.3.5 Comparative baseline models

We also consider several candidate approaches against which to compare the performance of our model. These all leverage either the feature matrix $\mathbf{X}_t$ or the network information $\mathbf{A}_t$ and serve as alternative formulations present in the literature for how to build a trading signal. These are summarized in Table 4.1 :

1. Market Long Only: buying all the stocks in the market with equal weights
2. MACD Momentum: averaging the momentum indicators defined in Equation 4.2 to use as a trading indicator.
3. Analyst Matrix: following Oyeniyi et al. [182], averaging the momentum of the 1-hop neighbours on the analyst coverage matrix.
4. Neural Network (NN): Using a 2-layer feed-forward neural network to predict $\hat{Y}_{t+21}$

Table 4.1: Comparison of features between our proposed model and benchmarks

Model Name	uses $\mathbf{X}_t$	uses $\mathbf{A}_t$	Learning
Long Only	Yes	No	No
MACD Momentum Averaging	Yes	No	No
Analyst Matrix	No	Yes	No
Neural Network	Yes	No	Yes
Ours ($\text{GAT}_{\text{analysts}}$)	**Yes**	**Yes**	**Yes**

4.3.6 Ablation studies

We also perform a series of ablation studies on the performance of our graph attention based model in different setups to understand what drives its performance and better grasp what features add value. To do this we take the basic setup described in 4.3.3 and replace different components of the basic $\text{GAT}_{\text{analysts}}$ model:

- *GCN*: A graph convolutional Network [135] learning model that does not use attention for propagating information.
- $\text{GAT}_{1_\text{layer}}$: A GAT model with only 1-layer instead of the 2 of our initial setup.

- GAT_{corr} : A GAT model which uses as neighbourhood information a correlation matrix which we transform into an adjacency matrix by eliminating edges with correlation inferior to the 90-th percentile.
- $\text{GAT}_{\text{industries}}$: A GAT model which uses as neighbourhood information a the GICS industrial classifications with firms being connected if and only if they are in the same industry
- $\text{GAT}_{\text{del_edge}}$: A GAT model which uses the original analyst network from which 60% were randomly removed.

4.3.7 Evalutation metrics

We evaluate these strategies by calculating several features displayed by the returns of their corresponding portfolios. Those are:

- Returns: the annualized average gross percent returns of strategy
- Volatility: their annualized average standard deviation of percent returns over the time period denoted as Vol.
- Sharpe Ratio: the ratio of the average annualized returns minus the risk-free interest rate divided by the standard deviation, a measure of risk-adjusted returns of the portfolio.
- Maximum Drawdown (MD): the maximum peak-to-trough span reached by the portfolio
- Maximum Drawdown duration (MDD): the maximum number of consecutive periods the portfolio was in drawdown expressed as a percentage of the entire number of periods

Moreover, we also calculate the cumulative log returns of each strategy by adding up the log returns at each trading period. This gives us a final measure of how well the strategy performed in gross returns.

4.4 Results

This result section will first present the main predictive comparisons in Section 4.4.1, then discuss the ablation studies in Section 4.4.2. We analyse the correlation between the returns and the market in Section 4.4.3. We perform a turnover analysis of the different strategies in Section 4.4.4. Lastly, we present discuss the behaviour of the attention weights in the model in Section 4.4.5.

4.4.1 Comparison of predictive performances

Table 4.2: Performance metrics of different portfolios

	Returns (%)	Vol. (%)	Sharpe	MD (%)	MDD (%)
Market	6.89	11.88	0.411	-39.4	21.0
Analyst Matrix	1.83	8.58	0.069	-22.8	51.0
MACD	9.56	11.46	0.672	-35.3	19.0
NN	15.44	8.32	1.753	-6.4	4.0
GAT $_{\text{analysts}}$	**29.44**	**7.07**	**4.069**	**-6.0**	**1.0**

Table 4.2 presents the results of the different strategies against the financial tests we use to evaluate performance. The market long only shows a Sharpe ratio of 0.411 and returns of 6.89% displaying relatively worse performances. It also displays the longest maximum drawdown with -39.4% of value lost at the trough. The Analyst Matrix strategy consisting of a weighted average of neighbour momentum is the worst performing strategy in both returns (1.83%) and risk adjusted returns (annualized Sharpe ratio of 0.069). It also performs the worst in terms of maximum drawdown duration with 51% of the trading backtest period (corresponding to 103 months). The MACD strategy performs better than the Analyst Matrix and the market long-only strategy with a Sharpe ratio of 0.672, and it shows a maximum drawdown of -35% and a maximum drawdown duration of 39 trading periods (21%). The Neural Network outperforms all the previously introduced strategies with a Sharpe ratio of 1.753 (more than double that of the next best MACD) and a much lower maximum drawdown of -6.42%. The Neural network strategy displays a shorter MDD than the MACD with 4.0% compared to 19%. Lastly, the $\text{GAT}_{\text{analysts}}$ method we introduce displays higher log-returns (at 29.44% annualized returns) and a Sharpe ratio of 4.069, more than double the previous best Sharpe ratio of 1.753 for the NN strategy. The $\text{GAT}_{\text{analysts}}$ also features lower drawdowns (-6%) and shorter MDD than any other competing strategy with 1% of the trading periods corresponding to 2 months of trading time being spent in continuous drawdown. The $\text{GAT}_{\text{analysts}}$ portfolio, moreover, displays lower volatility than every other strategy with 7% against 8.32% for the NN, the next lowest volatility strategy.

The superiority of the $\text{GAT}_{\text{analysts}}$ model compared to its competitors can also be observed in Figure 4.3 which presents the cumulative log-returns. The $\text{GAT}_{\text{analysts}}$ strategy performs the best amongst all the presented approaches with a cumulative log-return of 5 over the 18 year period. The Neural network is the second best performing strategy throughout the evaluation period. It lags the market at the start of the period and quickly outperforms it and all the other approaches. This temporal evolution of the returns also

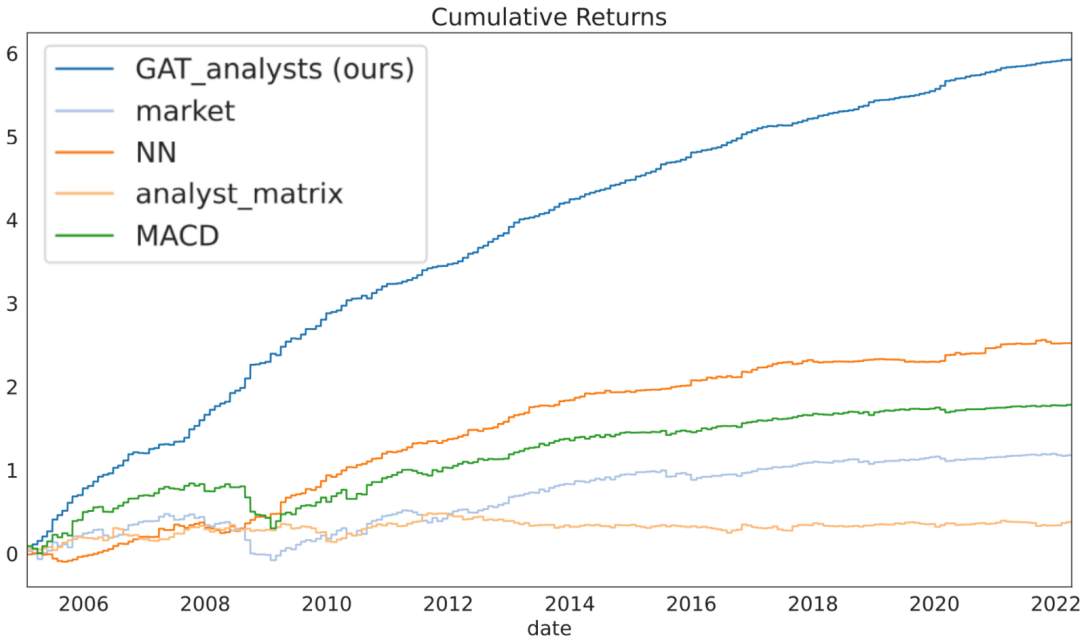


Figure 4.3: Cumulative returns

serves to highlight the relative superiority of the $\text{GAT}_{\text{analysts}}$ after the 2008 financial crisis, suggesting that it was able to identify promising lead-lag clusters despite the prevailing financial perturbation. Moreover, it also shows that the node and network information combined in the $\text{GAT}_{\text{analysts}}$ strongly outperforms the simple network aggregation strategy based on the Analyst matrix, which confirms our starting hypotesis that analyst coverage network topology and firm momentum features can be learnt jointly to extract alpha.

4.4.2 Ablation studies

Table 4.3: Ablation study

	Returns (%)	Vol. (%)	Sharpe	MD (%)	MDD (%)
$\text{GAT}_{\text{analysts}}$	29.44	**7.07**	**4.069**	**-6.0**	**1.0**
GAT_{corr}	**33.81**	9.10	3.757	-4.1	2.4
$\text{GAT}_{\text{industries}}$	19.21	8.26	2.250	-8.1	2.9
$\text{GAT}_{\text{del_edge}}$	19.57	8.26	2.265	-10.3	5.3
$\text{GAT}_{\text{1_layer}}$	28.54	8.02	3.417	**-3.6**	1.5
GCN	17.67	7.73	2.205	-3.9	2.4

Table 4.3 presents the performance of the basic GAT framework exposed to different sources of network information instead of the analyst matrix We can observe that the $\text{GAT}_{\text{analysts}}$ outperforms all the alternative representations of firm to firm relationships

with a Sharpe ratio of 4.069 compared with 3.757 for the GAT_{corr} and less than 2.25 and 2.26 for the industry and edge_delete versions. The $GAT_{analysts}$ displays slightly lower returns than the GAT_{corr} (29.44% vs 33.81%), however the basic approach also displays lower volatility than any of the other ablations. The $GAT_{analysts}$ model which only uses the analyst matrix also displays a lower Maximum Drawdown duration (MDD) with only 1.0% of the trading period being spent in drawdown, which is 50% less than the next longest at 1.5% of the GAT_{1_layer}. However, we can observe that the peak-to-trough maximum drawdown of the GAT_{corr} is less (-4.11%) than the maximum drawdown of the basic GAT (-6.01%). The table above suggest that the information content of the alternative network formulations such as correlations, industries and edge-deletion do not allow the GAT to generate better risk-adjusted forecasts. The GAT_{corr} displays slightly higher returns compared to the GAT_analysts, however the higher volatility incurred by the correlation suggests that the analyst matrix helps the $GAT_{analysts}$ model to select firms with slightly lower return volatility, leading to better risk-adjusted returns.

Table 4.4: Improvements from different model setups

Model Name	Cum. returns	Message Passing	Adjacency	# of Layers
$GAT_{analysts}$	**5.9**	Attention	Analyst	2
GAT_{corr}	5.4	Attention	Correlation	2
GAT_{1_layer}	4.7	Attention	Analyst	1
$GAT_{industries}$	4.7	Attention	Industries	2
GAT_{del_edge}	3.6	Attention	Perturbed	2
GCN	3.0	Convolution	Analyst	2
Market	1.2	-	-	-

Table 4.4 summarizes how much improvement we obtain in terms of cumulative returns for different message passing, adjacency information and number of layers for the models proposed in Section 4.3.6 as ablation studies. Cumulative returns can be interpreted as the ability of a model to extract a trading signal from the node and network information. We can observe that the largest improvementis achieved by introducing attention instead of graph convolutions for message passing. This significant improvement corresponds to a near doubling of cumulative returns over the trading period (+96% between $GAT_{analysts}$ and GCN). In addition to this, we observe that the going from a one layer model (GAT_{1_layer}) to a two layer model ($GAT_{analysts}$) led to an a 25% increase in returnsMeanwhile having the complete analyst matrix instead of the randomly perturbed one leads to 63% more returns. Replacing the industrial-GICS network of firms with the analyst matrix leads to a 25% improvement in cumulative log returns. Lastly, replacing the correlation-based adjacency matrix with the analyst matrix nets a comparatively smaller but still meaningful 9% increase in cumulative returns. These results comfort the

initial hypothesis that the analyst matrix contains useful information for the building of a portfolio. This improvement can be attributed to the structural information present in the adjacency as removing edges from the analyst matrix and replacing the matrix with other firm-to-firm networks leads to strongly degraded results. Another useful observation can be drawn from the single vs multi-hop setup which suggests that it is not only the single-hop neighbourhood aggregation that brings value as suggested in Oyeniyi et al. [182], Ali and Hirshleifer [13]. Rather, the proper leveraging of the complex and informative relationships present in the analyst matrix requires more complex models able to aggregate information from wider neighbourhoods. Combining information from 2-hop neighbours would allow the GAT model to update firm representations to include This could be the result of the analyst matrix helping to uncover latent relationships which are hard to detect otherwise, such as alignment of economic and actuarial practices amongst covered firms and the higher probability for an analyst to follow firms that use similar technological tools [160].

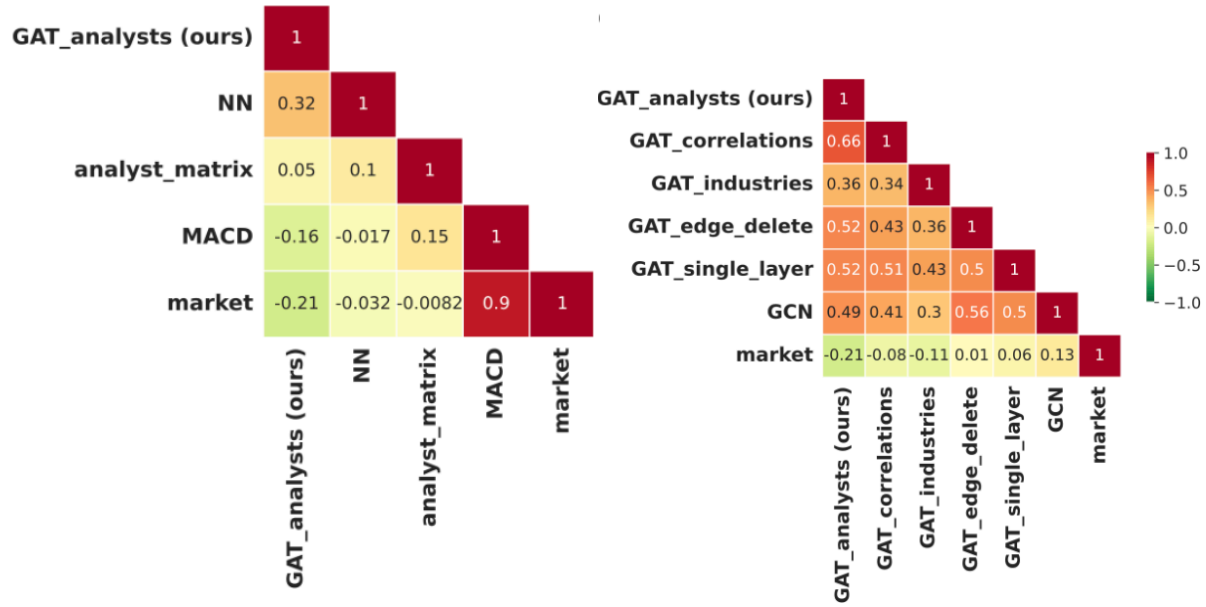
Table 4.5: Topology comparisons between different networks

Name	Jaccard	Diameter	Transitivity
Industry	1.0	1.0	1.0
Correlation	0.34	5.4	0.66
Analyst	0.98	11.29	0.67

Table 4.5 presents a comparison between the three network topologies considered in the ablation study. The Jaccard similarity (percentage of edges in common), Diameter (longest path between two nodes) and Transitivity (fraction of all possible triangles in the graph) are calculated at each time t for each graph. They are then averaged to produce on measure for each graph type and metric. The transitivity comparison shows that the correlation and analyst networks are less clustered than the industry network. That is expected as the industry network is a set of fully connected components. Moreover, the analyst and correlation networks both display similar levels of transitivity at 0.67 and 0.66 respectively, which signifies that over 60% of open triangles are connected. A high transitivity suggests both the analyst and correlation networks are strongly clustered.The network diameter in the Diameter column describes the maximum distance between two nodes in the graph. Both the analyst and correlation matrices have a higher diameter (11 on average for the analyst network and 6 on average for the correlation network) compared to the industry network. A network with a higher diameter can help in reducing the likelihood of over smoothing in GAT models which may explain the better performance of the $\text{GAT}_{\text{analysts}}$. Moreover, a higher diameter suggests that the analyst coverage connects firms along 'longer' chains. As discussed in Oyeniyi et al. [182], Ali and Hirshleifer [13], Jiang et al. [126], these longer chains closely represent fundamental real-life links between firms which tend to be missed by the correlation matrix. This helps explains

the added value of the analyst network : it captures different fundamental relationships [69]. The Jaccard index of the different networks presented the Jaccard column show that the analyst matrix remains consistently self-similar (over 90%) as compared to the correlation matrix. The structural properties of the industry network will not evolve as the structure of the industry classifications which explains the high Jaccard similarity of 1.0 (meaning the industry network stays on average constant through time). Meanwhile the correlation network displays a much lower Jaccard index of 34% implying it evolves faster period-to-period than the analyst matrix. A more stable network period to period can help with training a graph machine learning model and contributes to the better performance of the $\text{GAT}_{\text{analysts}}$ strategy. Existing literature [184] performs quantitative comparisons between industry and correlation based models as these are more commonly used to build graph-based trading strategies. However, few compare these networks with the analyst matrix as it is less often used as a building block of trading strategies. We show that these differences are profound and can serve to explain the different performance of a graph machine learning tool on each of these firm to firm networks.

4.4.3 Return correlation analysis



(a) Benchmark methods (b) Ablation study

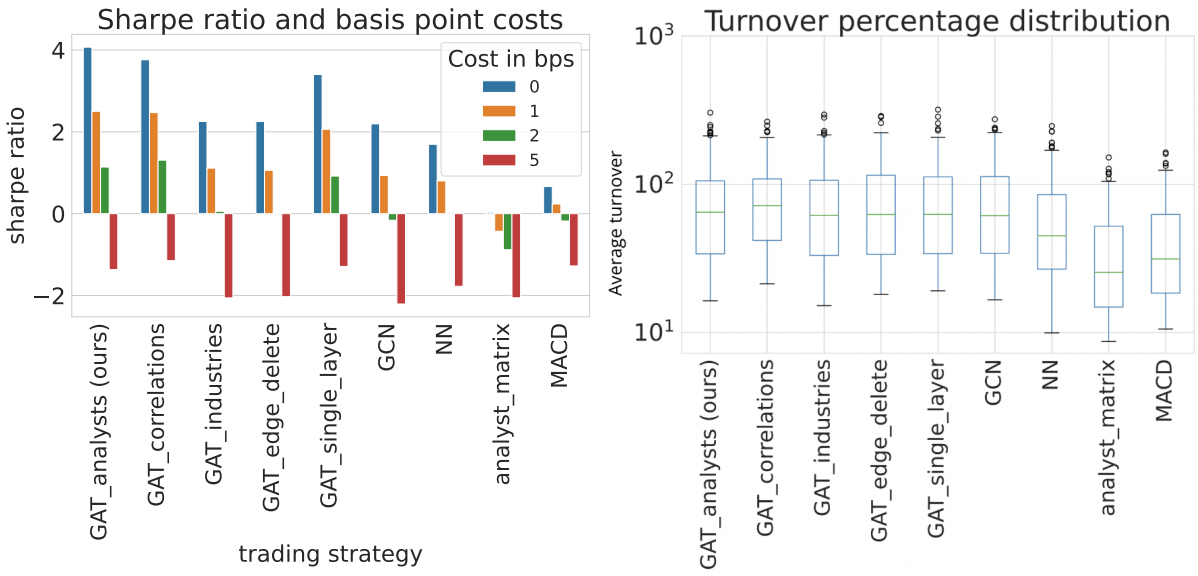
Figure 4.4: Correlation matrices of studied signals

Figure 4.4a displays the correlations of returns between the different evaluated trading strategies. A high correlation coefficient in red indicates that the two models' returns follow similar movements. Low correlations are generally sought by investors as it allows for diversification. All else being equal investors can limit the concentration of their

portfolio and increase their returns by investing in the least correlated signals. The GAT model appears most anti-correlated with the market with a -0.21 coefficient. The different models' returns appear uncorrelated with each other aside from a slightly higher return correlation between the $\text{GAT}_{\text{analysts}}$ and the NN of 0.32. These plots suggest that the returns of $\text{GAT}_{\text{analysts}}$ follow a pattern that responds to different market signals compared to the other competing models. This supports the hypothesis that the analyst network is a useful and informative prior with which to diversify a trading strategy as the signal it produces is different from the other existing ones. Figure 4.4b represents the correlation of the returns for strategies based on $\text{GAT}_{\text{analysts}}$ and all of the different ablations under investigation. The correlation of all of these approaches with the market are also presented. The original $\text{GAT}_{\text{analysts}}$ displays the lowest correlation with the market at -0.21 with the second lowest being the $\text{GAT}_{\text{industries}}$ with -0.11. Moreover, the original GAT's returns and signal is most strongly correlated with the GAT_{corr}'s returns and signal at 0.65 and 0.28 respectively compared with 0.52 and 0.28 for the $\text{GAT}_{\text{del_edge}}$. This suggests the signal of the $\text{GAT}_{\text{analysts}}$ is reasonably different to other setups. The addition of attention has a significant effect on the signal as evidenced by the comparatively low correlations signal between the GCN and all of the other models which use the analyst matrix and especially the $\text{GAT}_{\text{analysts}}$. This can be interpreted as an example of the differing effect of network information contained in each of the different networks upon the final trading signal: the analyst network helps the $\text{GAT}_{\text{analysts}}$ learn links between firms that are similar to lagged correlations and to a lesser extent industry linkages and yet distinct from both.

4.4.4 Turnover and cost analysis

We investigate how the strategy performs when exposed to transaction costs in Figure 4.5a. This figure describes the annualized Sharpe ratio of the different strategies under increasing trading costs from 0 basis points increasing to 1, 2 and 5 basis points. This penalises strategies with larger turnovers in portfolio content as depicted in Figure 4.5b. We observe that the returns of all of the models under consideration a sizeable decay in performance as the trading costs increase. All of the strategies yield negative Sharpe ratios when exposed to 5 basis points of trading costs. The $\text{GAT}_{\text{analysts}}$ and GAT_{corr} display similar turnover than other strategies with comparable Sharpe ratios such as $\text{GAT}_{\text{1_layer}}$. Only $\text{GAT}_{\text{analysts}}$, GAT_{corr} and $\text{GAT}_{\text{1_layer}}$ maintain positive Sharpe ratios when exposed to 2 basis points of trading costs: every other strategy exhibits a null or negative Sharpe ratio. This suggests that the analyst and correlation based strategies are more robust to trading frictions despite the drop in performance but it is the initial strong-risk adjusted performance that ensures they do well under tthese frictions. The $\text{GAT}_{\text{del_edge}}$, $\text{GAT}_{\text{industries}}$ and GCN display comparable turnover ratios. The analyst matrix's Sharpe ratio decays quickly from 0.069 to -2, however since it displays relatively



(a) Trading cost and Sharpe

(b) Turnover by strategy

Figure 4.5: Effect of trading cost on different strategies given turnovers

less turnover, it outperforms certain more complex strategies such as GCN or $\text{GAT}_{\text{industries}}$ under the highest cost regime. Lastly, we note that the Sharpe ratio of the MACD strategy fares the best out of all of the when exposed to the highest level of transaction costs, it experiences the least sharp drop going from 0.66 to -1.27. However, we also note that the $\text{GAT}_{\text{analysts}}$ approach still displays a better Sharpe ratio at a 2 basis point cost than the 0 transaction cost MACD portfolio and still outperforms it at the 5 basis points trading scenario. This suggests that the model-free approach, though interesting in terms of turnover-limitation, can be substituted under most cost scenarios with the proposed solution of this paper. Model-based approaches such as the NN, the $\text{GAT}_{\text{analysts}}$ and the ablated GAT models have on average a slightly higher average turnover (77$) than the model-free approaches (40$) such as the MACD and the analyst-matrix confirming findings in the literature [149]. Consequently, this explains why these model-free strategies are less affected by the increase in trading costs.

4.4.5 Attention analysis

Figure 4.6 represents the subset of the analyst matrix where the edges had the strongest attention coefficient in the $\text{GAT}_{\text{analysts}}$ model for one trading period in December 2016. The $\text{GAT}_{\text{analysts}}$ model learns to place high weight on the relationship between those firms when determining how to best forecast future value. The firm nodes are coloured by industry. This figure reinforces the idea that the attention function is helpful in allowing the model to learn informative and interpretable links between firms . For instance, the KO (Coca-cola) - STT (State Street Corp) link in the middle corresponds to a link

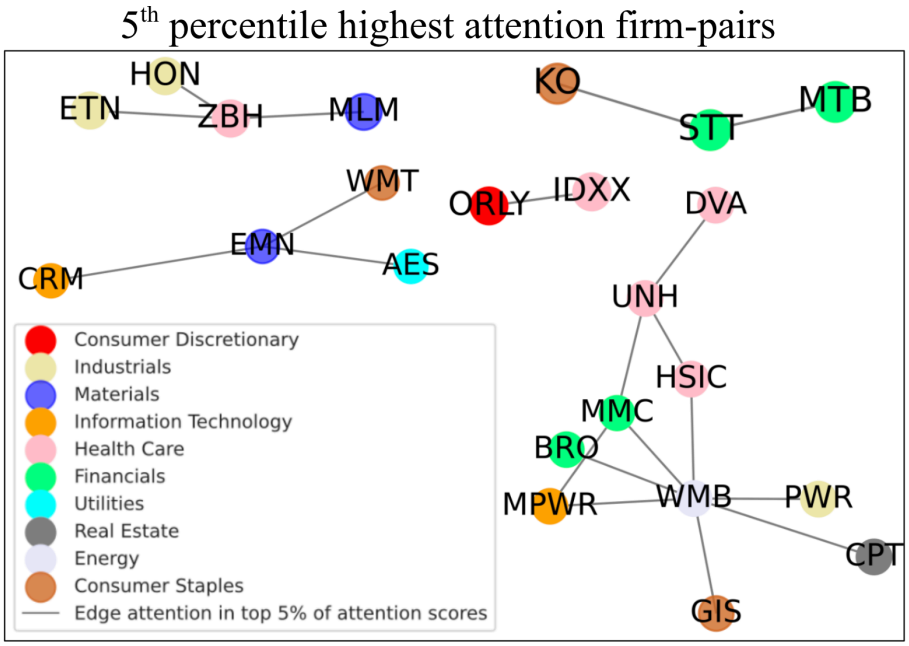


Figure 4.6: Firm links with strongest attention

between an investment firm (STT) with a very high stake in Coca-Cola. Similarly, the WMB cluster on the bottom right corresponds to an energy provider (WMB) connected to several firms who require energy provision, from the agro-industrial company GIS, to healthcare firms like HSIC. These links support the hypothesis that the GAT model identifies links between firms which correspond to fundamental economic linkages likely to lead to momentum spillover effects from investor under-reaction [13, 142, 39]. Moreover, this graph shows the strongest attention weights being quite spread out by industry, i.e., the frequency of connection between industries is higher than within industries, suggesting that the model is able to flexibly learn inter and intra-industry patterns in a richer way than just focusing on the industry matrix as suggested by Martens and Sextroh [160]. Lastly, one important point to note is that none of these entries with highest attention are present in the correlation matrix as the strength of these correlations is lower than the 90-th percentile of correlation used in defining the correlation matrix. This implies that the attention mechanism applied to the analyst matrix is flexible enough to allow the model to reveal meaningful 'economic relations' between firms [69]. These links go beyond simple correlation based measures which have been criticized in the literature for an inability to capture non-linear relationships [161].

4.5 Conclusion

In this paper, we have explored the ability of a graph attention network to systematically learn a novel trading signal from a metric of firm to firm momentum spillovers: the

analyst coverage network. We have shown that the information contained in the analyst network can be efficiently extracted by the graph attention network in order to produce positive and persistent out-of-sample trading returns. Our method outperforms existing benchmarks in terms of average returns and drawdowns. Furthermore, we have shown the strategy's robustness through ablation studies and turnover analysis. Moreover, we have shown the link between topological information contained in the analyst network and the performance of our strategy by comparing the performance of the model trained on the analyst network to that of the same model trained with industry and correlation based networks. This work represents the first step in incorporating analyst coverage networks into financial graph machine learning. We demonstrate the profitability of the strategy and the wealth of insights which can be leveraged from this setup. We outline several pathways for future work. The first is extending the proposed model to incorporate richer temporal information, for instance by making the edge information into a time series whose features could be modelled explicitly by a dedicated temporal learning block. Moreover, given the link between analyst estimation error and volatility, the GAT model could be used to model volatility spillover effects between firms across edges of the analyst network.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Extracting alpha from financial analyst networks.
Publication Status	X Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Gorduza, D., Kong, Y., Dong, X., and Zohren, S. (2024). Extracting alpha from financial analyst networks. *In Proceedings of the 5th ACM International Conference on AI in Finance, pages 397–405.*

Student Confirmation

Student Name:	Dragos Gorduza		
Contribution to the Paper	Wrote manuscript, performed data collection and analysis, interpreted results		
Signature		Date	30 August 2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Xiaowen Dong, Associate Professor	
Supervisor comments I can confirm that Dragos made significant contribution to the paper above.		
Signature Xiaowen Dong	Date	07/10/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Wisdom of the Crowds or Ignorance of the Masses? A data-driven guide to WallStreetBets

A trite yet fundamental question in economics is: What causes large asset price fluctuations? A tenfold rise in the price of GameStop equity, between the 22nd and 28th of January 2021, demonstrated that herding behaviour among retail investors is an important contributing factor. This paper presents a data-driven guide to the forum that started the hype – WallStreetBets (WSB). The paper's initial experiments decompose the forum using a large language topic model and network tools. The topic model describes the evolution of the forum over time and shows the persistence of certain topics (such as the market / S&P500 discussion), and the sporadic interest in others, such as COVID or crude oil. The authors use network analysis to decompose the landscape of retail investors into clusters based on their posting and discussion habits; several large, correlated asset discussion clusters emerge, surrounded by smaller, niche ones. A second set of experiments assesses the impact that WSB discussions have had on the market. The authors show that forum activity has a Granger-causal relationship with the returns of several assets, some of which are now commonly classified as 'meme stocks', while others have gone under the radar. The paper extracts a set of short-term trade signals from posts and long-term (monthly and weekly) trade signals from forum dynamics, and considers their predictive power at different time horizons. In addition to the analysis, the paper presents the dataset, as well as an interactive dashboard, in order to promote further research.

5.1 Introduction

Social media has changed the structure of our society. As many as 4.9 billion people, or 61% of the world population, are active social media users, each just a few clicks away from the next viral phenomenon. People turn to their accounts for everything from news to product suggestions. Now, a growing audience turns to social media for promising stock market gambles. Even though investor discussion forums have existed for decades, the `r/WallStreetBets` (WSB) subreddit was arguably the first to reach an unprecedented retail following – since it's creation in 2012, the forum grew exponentially in membership, attracting followers not only through lucrative trade ideas, but also through the promise that coordination among smaller retail traders could unseat investment giants. Their aspirations became a reality in January 2021 when the forum's cherished stock, GameStop (GME), experienced a 22-fold rise in asset price, while *Melvin Capital*, an investment fund with a large short position in GME, experienced a 30% decline in its value.

A single trader with $1,000 in her bank account couldn't move the markets. 750 thousand individuals on the other hand, each with $1,000 in their bank accounts, could have bought all of the floating shares of GameStop at its early January price of $17.25 per share. However, how did they overcome the colossal coordination challenge? What was the structure of the discussion on `r/WallStreetBets` and how did the conversation evolve? In this paper, we dive under the hood with the goal of shedding light on the dynamics of the forum which started the hype.

Even though the literature has explored specific phenomena within the WSB forum, we believe that our work is the first to take a broad approach.[1] Our goal with this paper is to thoroughly characterize the trading signals and behaviours on WallStreetBets. We focus on the following questions:

- What are the primary topics of interest to WSB users, and how do they relate to each other in the forum discussion landscape?
- What assets are forum participants discussing, and can their conversations be used as indicators of market moves?
- Are WSB users trend-followers or predictors of asset returns?

We answer these questions by applying several machine learning tools to the dataset of WallStreetBets submissions and comments:

1. We extract text features from the forum data, such as sentiments and topics using large language models, as well as tickers mentioned within posts.[2]

[1] See [47], [208], [52], [240], [60]
[2] See [21], [19]

2. We estimate the relationship between assets using a network approach applied to both our extracted topics, and user submissions.

3. We approximate the interaction of asset returns and user sentiment by estimating the cumulative abnormal returns in the days preceding and following submissions.

4. We study whether WSB discussions forecast market returns by estimating a Granger causal relationship between the two.

The text features within our data allow us to uncover several important characteristics. Our topic model shows how the interests of the forum have evolved over time. Certain topics, such as the discussion of Tesla and Elon Musk, retain the interest of WSB users. Others, such as oil and or COVID, briefly grab the forum's attention, but are promptly replaced by other discussion themes. Our sentiment model allows us to extract a perspective taken by users (bullish, bearish, or neutral) about the future movement of the asset mentioned in a submission. This sentiment measure, in turn, allows us to understand the predictive power of the forum and user reactions to market moves.

The network approach allows us to cluster assets based on: i) whether they are frequently mentioned together while discussing the same topic (*Topic Network*), and ii) whether the same group of users are interested in both assets (*Submission Network*). We observe that distinct clusters of similar assets emerge. Posts within the large asset clusters generally tend to incorrectly forecast future asset prices. However, smaller, niche clusters have better market prediction performance, perhaps indicating that close-knit groups of well-informed retail investors may have market insights.

We analyse the seven-day Cumulative Abnormal Returns (CAR) in the fourteen days preceding and following a post about an asset. The CAR is the asset log-return, which is distinct from simultaneous market returns, and is typically used to model sudden price shocks and event detection in financial markets. We observe that, overall, WSB appears to be reactive to, rather than predictive of, CAR in an asset. The average CAR appears to run up gradually immediately before a post, and subsequently experiences a sharp decline. The pattern is particularly distinctive in 'meme' stocks, where high asset returns potentially reinforce the hype. In stocks with a broader following, such as MSFT or AAPL, CAR appears flat around WSB posts.

In a final exercise, we study a Granger causal relationship between asset returns and sentiments in posts. A time series A is said to Granger-cause B if it can be shown that the *lagged* values of A provide statistically significant information about *future* values of B. Our analysis highlights that a Granger causal relationship exists between the sentiments on WSB of many 'meme' assets and their returns – the result is not surprising, since conventional wisdom is that the WSB discussions heavily influenced 'meme' stocks. The Granger causal analysis allows us to look at the comparative degree of influence between

different tickers, and also to identify new assets which may have been influenced by WSB but have gone under the radar - such as SnapChat.

Beyond the immediate quantitative insights, an additional goal of the paper is to provide tools and data for researchers and practitioners. We hope that this research will contribute to a broader effort within the financial industry to take advantage of and evaluate tradeoffs in non-stationary and quickly-evolving signals. We make the dataset used within the paper, manual annotations, as well as code database, public. Furthermore, in an effort to encourage a more interactive data experience for practitioners, we publish our insights in a dashboard, allowing users to explore various aspects of the dataset and evaluate opportunities for market returns.[3]

Relevant Literature

We believe that this paper may be of interest to those looking into social media in finance and the WallStreetBets forum specifically, as well as to those studying new, noisy signals within the finance community.

Our paper takes inspiration from an emerging literature looking at the relationship between social media activity and stock market prices. It is well understood that new information about a company (such as from a news report) can affect its stock price, as well as the stock price of similar or competing companies.[4] Social media data provides information beyond news: shedding light on how individuals engage with financial information, which can reflect the sentiment of the market more broadly. Many researchers have exploited this to find associations between publicly expressed sentiment on Twitter / StockTwits and stock market prices.[5] Social media has thereby provided rich new text data for quantitative finance researchers to exploit, and is now a key input (along with news articles and company reports) for text analysis in finance.[6] We contribute to this literature both through our analysis of a specific social media dataset, the WallStreetBets forum, and through providing annotated data for future research. A more targeted set of papers focus on specific characteristics of the forum. We add to this literature by taking a more broad approach to analysing the forum.

[3] `https://sites.google.com/view/wsbtrialsite`

[4] See [233]

[5] See [219], [186], [88], [27], [9]

[6] See [98] for a review

5.2 Data

Data collection

Submission and comment data from WSB is collected through the Pushshift Reddit database, which provides an archive of all posts and comments on Reddit.[7] We collect data from the inception of the forum in April 2012 up to 24 June 2022. Posts on WSB typically refer to companies by their stock symbol, and we use the work of [208] to obtain a list of stock symbols (from Yahoo Finance and Compustat) and extract these from the submission text. Stock price data is sourced from from AlphaVantage.

Due diligence posts

Due diligence (DD) posts have attracted attention in the literature as 'higher quality' posts from WSB, which contain meaningful stock-level insights. To obtain the subset of due diligence (DD) reports in the forum, we firstly filter to posts that have the DD flair. We supplement the posts containing the DD flair with posts that contain the DD acronym. We manually review every single potential DD post (i.e. flaired posts that are not removed by moderators, or contain the text 'DD'), and we remove any post that does not appear to be a valid DD – i.e. does not contain market insight. The DD posts are sufficiently small in number that we can manually label each post as bullish/positive or bearish/negative manually (neutral posts are removed). We find 77 percent of posts are bullish, and 23 percent are bearish.

Sentiment classification

For the forum as a whole there are far too many posts for manual classification. Previous papers have either relied on counting keywords, or they have used off-the-shelf rule based sentiment classifiers, resulting in certain mis-classification issues, discussed in our Supplementary Materials, posted to our dashboard site.[8] To resolve these issues, we firstly hand-code a random sample of 4,000 WSB posts, categorising them as bullish, bearish, or neutral depending on what the author's attitude is on future price increases. Of the posts in the sample 41 percent are bullish, 37 percent are neutral and 22 percent are bearish.[9] We then take a pre-trained BERT model called FinBERT, which we fine-tune using our manually labelled sample.[10] Our final model had a test accuracy of 69.2 percent, which we consider sufficient due to the challenging task of predicting future-facing sentiment.

Further details on data sourcing, cleaning and annotation are presented in Supplemen-

[7] See [37]

[8] `https://sites.google.com/view/wsbtrialsite`

[9] Since bearish posts tend to be a minority on the forum, bearish posts were up-sampled to reduce class imbalance.

[10] See [80], [21]

tary Materials, posted to our dashboard site.[11]

5.3 WallStreetBets Forum Dynamics

In this section, we characterise the discussion landscape of the WSB forum. We consider how assets are related on WSB in terms of topic and user interest. Additionally, we explore the relationship between WSB and markets through studying the returns preceding and following posts. The section that follows discusses predicting market movements using data from WSB.

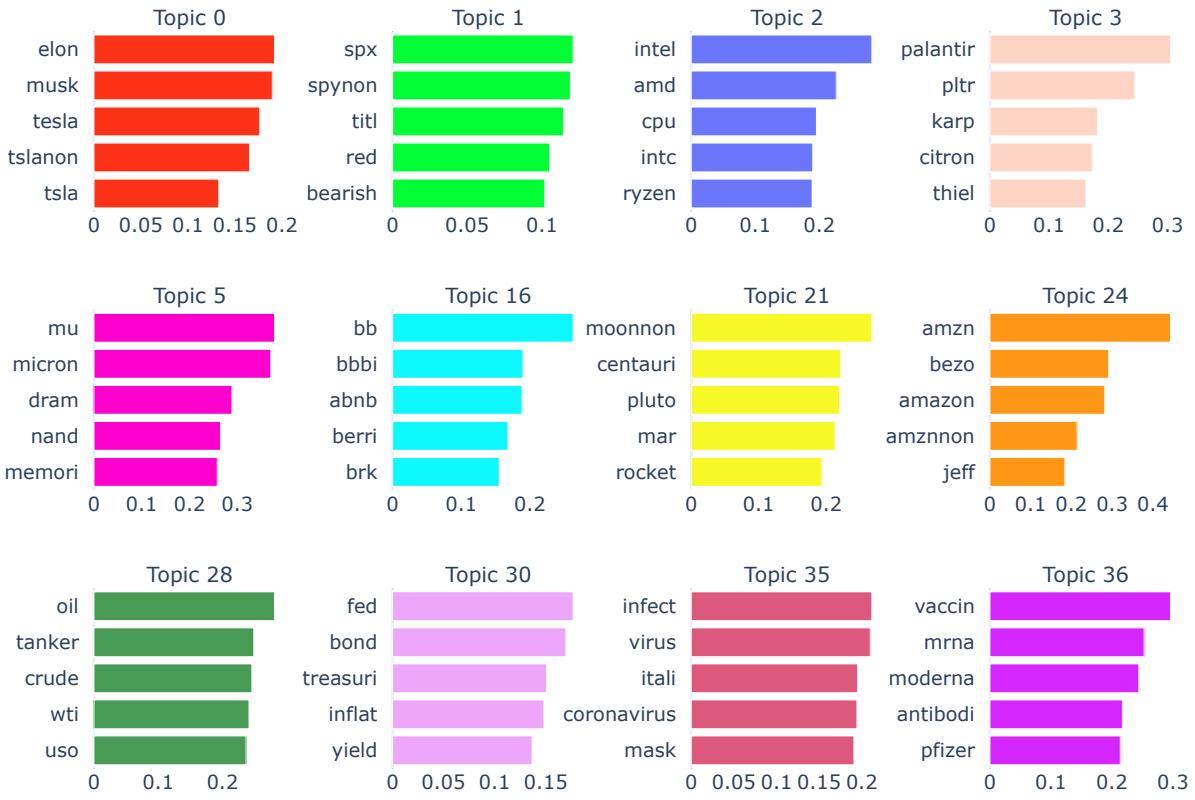
5.3.1 The topic landscape of WSB

Topic models are often employed to map out the distribution of information in a textual corpus.[12] In economics and finance, they are used to uncover the latent discursive directions that individuals can follow when expressing an opinion, which allows researchers to quickly detect salient points in the discourse and use them as signals for downstream analysis.[13] However, prior work leverages older topic models which do not make full use of the new modelling capacities offered by large language models (LLMs).[14] We use the Bert LLM Topic package to perform topic modelling on WSB, which accounts for non-linear semantic similarities between texts which go beyond word-level co-occurrences (a limitation of traditional topic models).[15] This model gives us a mapping from each post in the sample to a learnt set of topics. Each topic has a unique set of representative terms associated with it, and posts are mapped to topics based on how well the text within the post matches that of a particular topic.

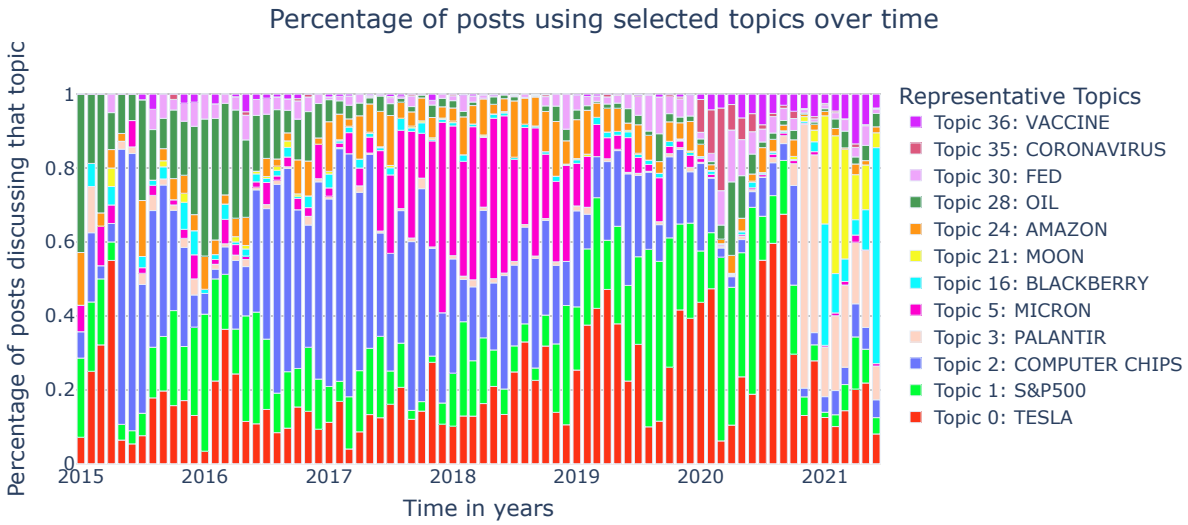
Figure 5.1 highlights our key findings from employing the BERT topic model: Figure 5.1a present some of the key topics of discussion on WSB, while Figure 5.1b displays their prevalence on the forum over time. Each post within WSB is given a probability vector whose length is equal to the number of extracted topics, identifying the extent to which the post is associated with different topics. For example, as post which discusses pharmaceutical companies and how they will profit from the COVID vaccine will be given a higher probability of being associated with topics 35 and 36.

Figure 5.1b is constructed by first splitting our dataset into months and then calculating the fraction of the discourse dedicated to that topic within that month for the ten most popular topics. The span of topics covered here is a subset of the overall topics identified by the model, which serve to illustrate the key areas of attention for the forum across time.

[11] `https://sites.google.com/view/wsbtrialsite`
[12] See [44], [19]
[13] See [99]
[14] See [206]
[15] See [19]



(a) Sample of Representative Topics Discussed on WSB; the figure displays topics and the frequency with which the most popular words appear within the topics (ranked by the categorical term frequency - inverse document frequency score, described in [19]).



(b) Topic Distribution over Time

Figure 5.1: The WallStreetBets Discussion; we present several representative topics from WSB in Figure 5.1a, as well as their relative importance within the forum in Figure 5.1b.

We note that several topics and their prevalence follow news, such as CORONAVIRUS (topic 35) which appears in March 2020 and VACCINE (topic 36), which are more prevalent during the introductions of the first COVID-19 vaccines in end of the year 2021. The topic model also tracks the emerging focus of the forum on so called 'meme stocks' such as TESLA (topic 0), PALANTIR (topic 3), and BLACKBERRY (topic 16). Contrary to other transient topics, we detect a persistent interest of the forum in SPY (S&P500) as noted by the prevalence of topic 1, which represents attention toward the overall state of the market.

5.3.2 How are assets related to each other on WSB?

The WSB forum is uniquely suited to study the relationships between assets, as they are perceived and discussed together by retail investors. We choose two methods to map the relationships between assets and to create a ticker-to-ticker network structure: the *Topic Network* approach and the *Submission Network* approach. The *Topic Network* uses the frequency with which assets are discussed within the same topics to create inter-asset connections, indicating that investors perceive the assets to be influenced by the same general financial trends. The *Submission Network* tracks which groups of investors are interested in, and create submissions about, the same assets. We describe both approaches below.

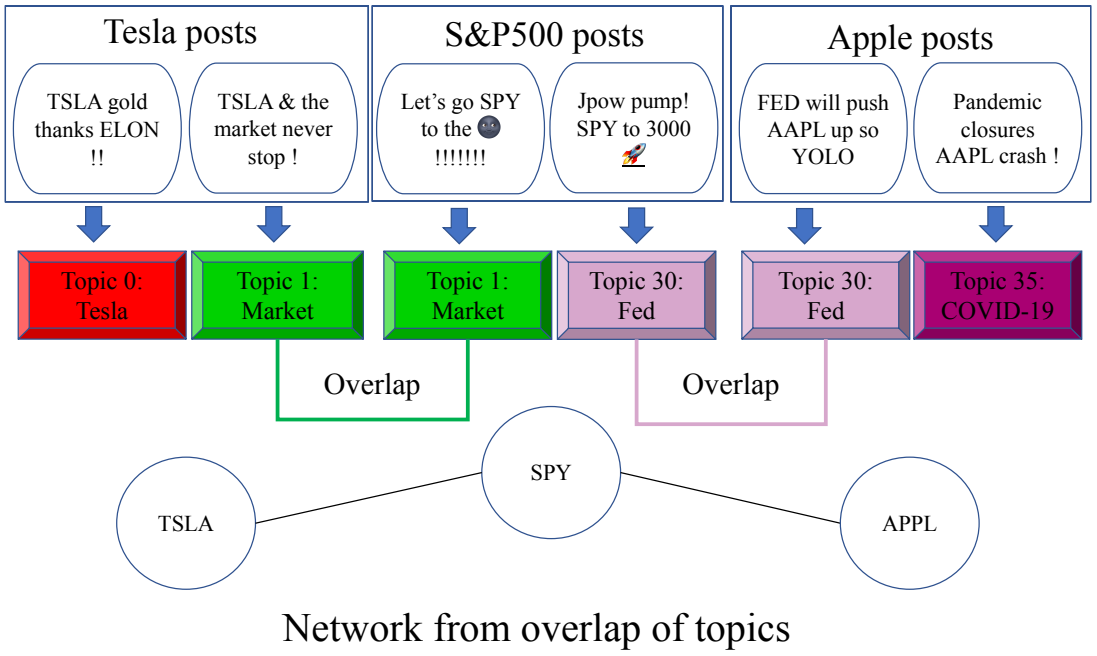
Topic Network

In the *Topic Network* approach, we link assets based on the frequency with which they are mentioned in the same set of topics. The intuition behind the approach rests on the idea that assets that are discussed within the same financial topics are likely related to each other through some underlying fundamental relationship.

We explain our intuition through a simplified example. Let us consider two assets: i) interest rate swaps and ii) bonds. We extract a topic from our topic model about the FED, and we observe that both bonds and interest rate swaps are frequently brought up in posts that are labeled with the 'FED' topic. This is unsurprising, given the fact that FED decisions would strongly affect the valuations of both bonds and interest rates swaps. In our *Topic Network* exercise we would place a connection between bonds and interest rate swaps, since they are linked by the 'FED' topic. This intuition can be extended to better understand the meaning behind the tickers linked through out topic model: they are tickers that are frequently brought up under the same topics, indicating that they are linked through some underlying economic discussion theme.

Figure 5.2a provides a toy example. TSLA and SPY are connected because they are both mentioned in the topic discussing the overall stability of the market (*market* topic); SPY and APPL are connected since they are both discussed in conjunction with the *FED* topic.

(a) *Topic Network* construction illustration



(b) *Submission Network* construction illustration

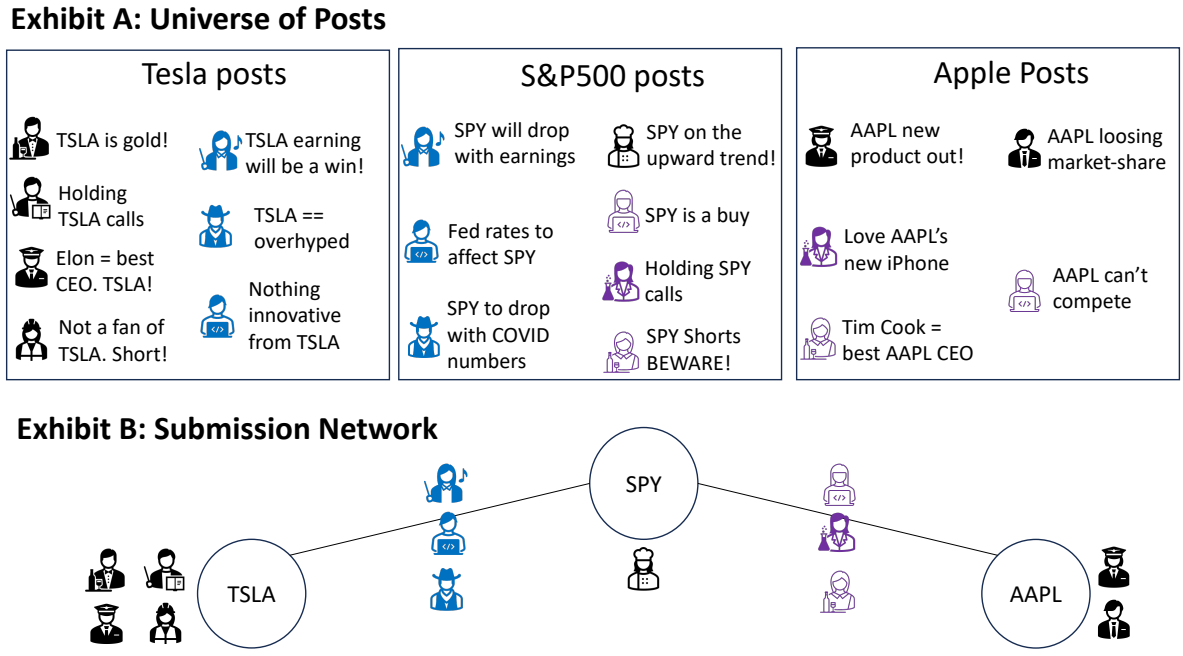


Figure 5.2: Network Construction; we demonstrate how links are identified between assets in the *Topic Network* and *Submission Network*.

In practice, we also add a thresholding step when constructing links between tickers based on overlapping topics. We select the top 50 topics, ranked by frequency of appearance in posts. This means that we only count a topic if it is represented often enough across our dataset. We then count all the topics that correspond to a given ticker and we keep the topics that are present in 20% or more posts related to that ticker – this is done to filter out less important conversations about assets. The filtering leads us to drop tickers whose topic distribution is very diffuse, meaning that discussions about it do not have consistency. As a prime example, this leads us to drop GME from our topic network as the topic distribution in GME posts is too dispersed.

Submission Network

The community structure within WSB offers a different perspective on how assets are related within the forum. To construct our submission network, we look at the overlap of users who create posts about two assets. Two assets are linked if there is a sufficiently large fraction of users discussing both assets simultaneously.

Figure 5.2b provides a simplified example. We observe a subset of users creating posts about TSLA and SPY simultaneously, and a different group of users writing about AAPL and SPY. However, we do not observe the same users creating posts about TSLA and AAPL. Therefore, we create two links: TSLA / SPY and AAPL / SPY.

In practice, our network construction exercise is slightly more complicated. We create weighted links between assets, normalized by the total number of users mentioning each asset. This means that if asset A is discussed by ten users, and all of them also post about asset B - we would create a link with weight one from asset A to B. However, if 10,000 users posts about asset B, we would not create a link back from B to A, since only a negligible fraction of posters about B care about A. Therefore, the *total weight* of the link between A and B would be 1 from the first created link. In practice, our threshold for including a link in our weight calculation is 20% – at least 20% of users posting about one ticker must be posting about the other ticker. Relating this back to Figure 5.2b, we observe that three out of seven SPY posters also post about AAPL (weight 3/7) and that three out of five AAPL posters post about SPY (weight 3/5) – the total weight for the link between AAPL and SPY would, therefore, be 3/5+3/7. We filter our submission network to contain tickers that are mentioned at least 150 times on the forum.

5.3.3 Asset network structures

Topic Network – Results

Figure 5.3a presents our *Topic Network* – links indicate which two tickers are likely to be mentioned together within the same economic discussion topics. A list of tickers and their associated clusters are presented in Table 5.1. We observe that larger and frequently

Table 5.1: **Subset of Clusters extracted from the *Topic Network***

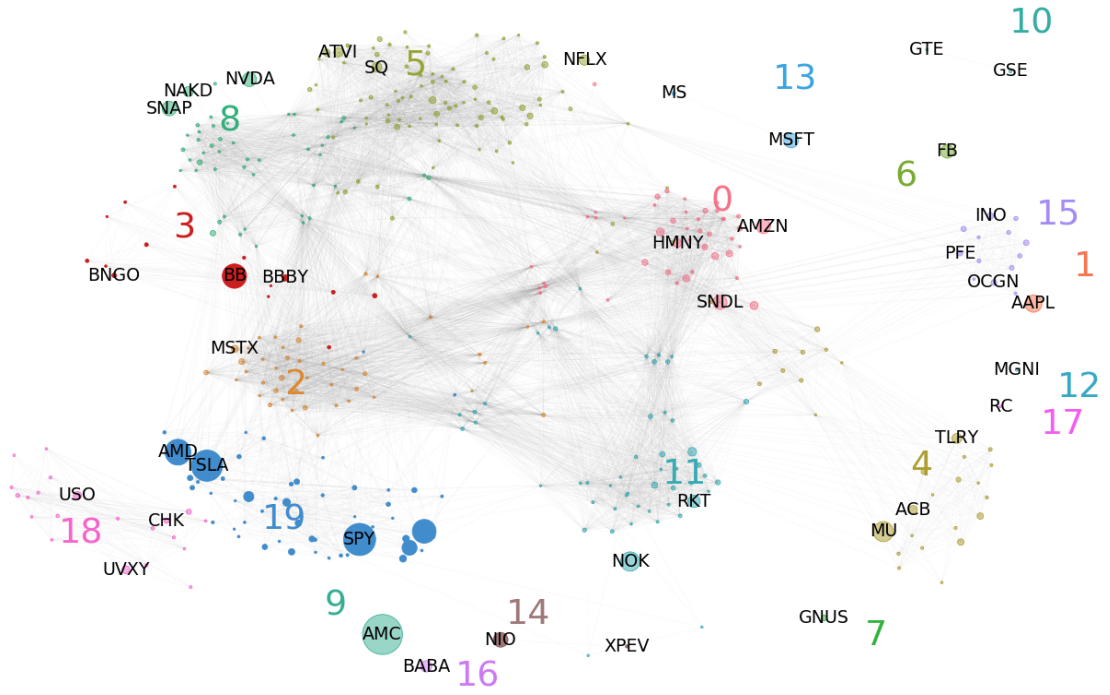
Cluster	Popular Tickers within Cluster	Posts	Total No. of Tickers
0	AMZN, SNDL, HMNY, MVIS, WISH, CLNE, CCL, DKNG, SUNE	12,373	41
1	AAPL	2,530	1
2	MSTX, AVXL, MNKD, OPK, AMDA, AUPH, SENS, AMRN, ADMP	3,312	44
3	BB, BBBY, BNGO, BRK, BBY, DB, BLNK, ABNB, BBW, BTT	9,898	12
4	MU, TLRY, ACB, CGC, APHA, CRON, AMAT, MO, OGI, HEXO	6,557	17
5	NFLX, SQ, ATVI, PLUG, WMT, SBUX, LULU, CRM, CMG, MCD	10,045	72
6	FB	1,580	1
8	SNAP, NVDA, NAKD, NAK, FSLY, TTD, DBX, KR, TWLO, GNC	5,763	35
9	AMC	19,686	1
11	NOK, RKT, WKHS, ZOM, QS, SRNE, TRCH, ASO, LFIN	11,349	33
13	MSFT, MS	1,849	2
14	NIO, XPEV	2,160	2
15	INO, OCGN, PFE, NVAX, MRNA, CVS, JNJ, AZN, VXRT, TEVA	2,474	13
16	BABA	1,175	1
18	USO, UVXY, CHK, WTI, XOM, SVXY, NAT, RIG, UCO, BP	1,001	12
19	SPY, TSLA, AMD, PLTR, SPCE, NKLA, QQQ, PRPL, GM, ROKU	28,438	30

discussed companies, such as AMC, have more cohesive topics which are not referenced in many other companies' posts. Therefore, the posts of several tickers are isolated from the main connected component of the topic network. These includes: AAPL (cluster 1), FB (cluster 6) AMC (cluster 9), and BABA (cluster 16). We interpret the isolation of these network components as the result of a systematic use of a *limited* and *distinct* set of topics when discussing these stocks. A post about AAPL will mobilise only topics similar to other AAPL discussions: focusing the attention of readers more narrowly on discussion points unique to Apple Inc as opposed to the wider market. Contrary to that, smaller companies tend to have a more central position in this network, and are often connected to stocks with higher market caps. This is visible by the more spread out cluster 19, which encompasses a variety of both smaller and larger companies as well as ETFs, likely connected through broader discussions about the economic climate.

Additionally, we find smaller clusters of assets related to a narrow set of economic themes within the market. A prime example of this is cluster 15, including INO, OCGN, PFE. These firms are likely linked by the same topics of drug discovery, FDA approvals and the COVID vaccine, in a way that is distinct to this particular group of assets.

We use our *Topic Network* as a distance mapping between firms that takes into account their 'discursive environments'. We interpret tickers belonging to the same cluster as an indication of greater similarity within their discussion topics. Consequently, belonging to the same cluster could be an indicator of a more correlated information set driving returns for all assets within the cluster.

(a) Topic network and extracted clusters



(b) Submission network and extracted clusters

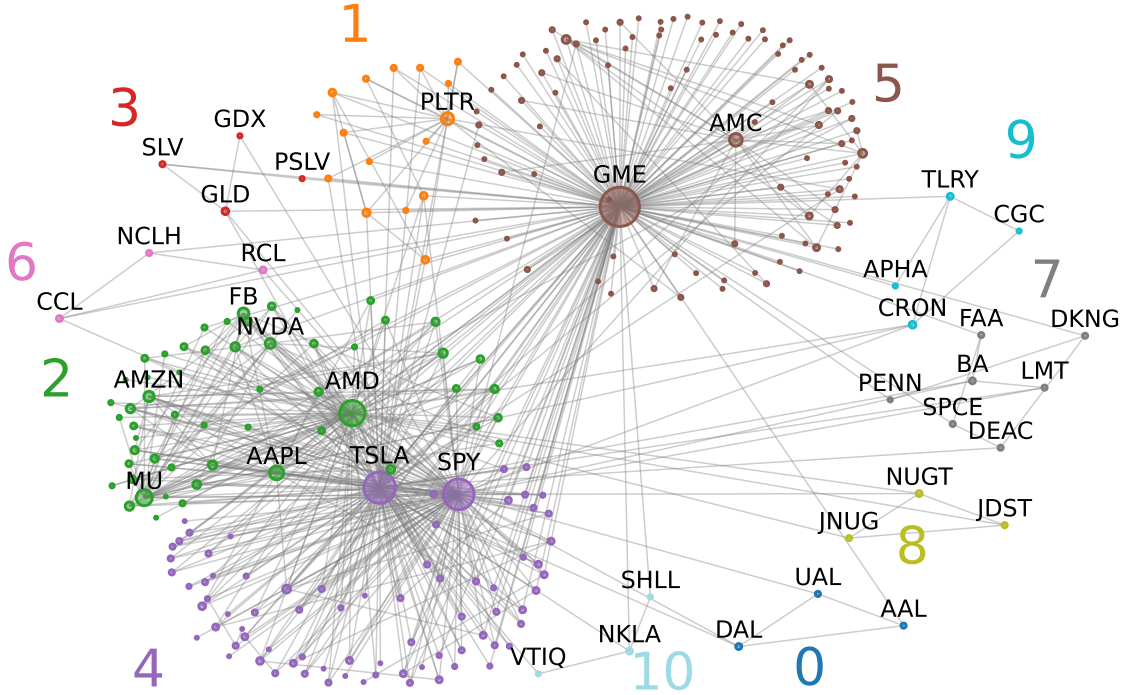


Figure 5.3: Extracted Network from WSB; we illustrate the ticker networks extracted using both the *Topic Network* and *Submission Network* approaches. Clusters of closely connected tickers are extracted using the Leiden algorithm from [224] and highlighted with the same color in both networks. In Figure 5.3a, ticker nodes are scaled by the relative size of their number of posts; only the top 50 most mentioned topics are displayed. In Figure 5.3b, nodes are scaled by the number of connections they have.

Submission Network – Results

Figure 5.3b presents the results of a clustering exercise on the *Submission Network*. We observe that similar assets appear to be mentioned within the detected clusters – implying that individuals self-select into discussions about certain asset categories. Said differently, a distinct group of investors is interested in marijuana stocks to those discussing the cruise line industry. This is perhaps most pronounced in the small cluster of gold ETFs containing JNUG, NUGT, JDST – cluster 8. Several other smaller clusters, dominated by a niche sector, are also visible in clusters 0, 3, 8, 9. We observe a large cluster around the SPY ETF and TSLA (cluster 4), as well as a large 'meme stock' clusters around GME (cluster 5). We observe a distinct cluster with some large, popular tech stocks including AAPL, FB, MU, AMZN, AMD (cluster 2). In our Supplementary Materials, we explore the correlation among assets in different clusters, as well as returns to WSB advice within the different clusters.

Returns and Correlations across Clusters

In our *Submission Network*, we observe that distinct groups of investors are interested in different types of tickers: some users come to WSB in order to discuss pharmaceutical companies, while others are interested in trading natural resource ETFs, or airline stocks. For this reason, tickers clustered through the *Submission Network* tend to have returns that are highly correlated – clustered assets appear to have an average return correlation of greater than 0.2, while assets across our entire dataset appear to exhibit an average correlation of 0.16. The most highly correlated asset returns appear to be within clusters 0, 6, and 9: the returns of the assets within these clusters exhibit correlations of 0.80, 0.84 and 0.60, respectively. The average next-day returns of investing according to the extracted sentiments within submissions is statistically significantly different from 0 at the 1% threshold according to a 2 sided t-test. The average next day returns are negative across most clusters (clusters 0,1,2,5, and 9). The most negative average returns are exhibited within clusters 0 and 5. Average returns of submissions in cluster 7, on the other hand, are statistically significant and positive. This demonstrates that most investors on WSB lack insights into future market moves. Notably, investors into the 'hype' cluster tend to lose the most. However, niche groups of investors may have worthwhile market insights – as demonstrated by returns for cluster 7. Our Supplementary Materials present returns and correlations across clusters.

A similar pattern can be observed from the *Topic Network* clusters. Investing according to the sentiments of submissions within most topic clusters results in a statistically significant, most of the strategies in a cluster exhibit average next day returns that are different from 0 at the 1% threshold according to a 2 sided t-test. However, similarly to our observations from the *Submission Network*, certain topic clusters exhibit positive daily returns, on average: cluster 1 containing AAPL, 13 containing MSFT and MS, 14

containing NIO and XPEV (two electric car makers), and 16 containing BABA. We notice that smaller clusters with fewer assets perform better than much larger clusters, again demonstrating that the forum as a whole may lack market insights, however specific topics and groups of investors may have promising insights.

5.3.4 Returns preceding and following posts

What are the characteristics of assets returns before and after they are mentioned on WSB? Is there any evidence that WSB users can predict returns, or are they trend followers just like other retail investors?

Framework

We consider how asset prices are changing shortly before and after posts on the forum. To do this, we follow the methodology of [233], which analyses market movements around news events by looking at changes in abnormal return (AR). AR is derived from the Capital Asset Pricing Model (CAPM),[16] which describes returns for company i at time t as follows:

$$r_{i,t} = \alpha_t + \beta_t r_{m,t} + \epsilon_{i,t}. \tag{5.1}$$

Here $r_{i,t}$ is the log return in the price of stock i on day t compared with the previous day, where $r_{i,t} = \log\left(\frac{p_{i,t}}{p_{i,t-1}}\right)$ and $p_{i,t}$ is the adjusted close stock price. $r_{m,t}$ is the return of the market (in our case, we use the S&P 500), hence β captures stock price moves that are driven by movements in the wider market. $\alpha_{i,t}$ captures stock over/under-performance relative to the market. $\epsilon_{i,t}$ is a stochastic error term, often referred to as abnormal return (AR).

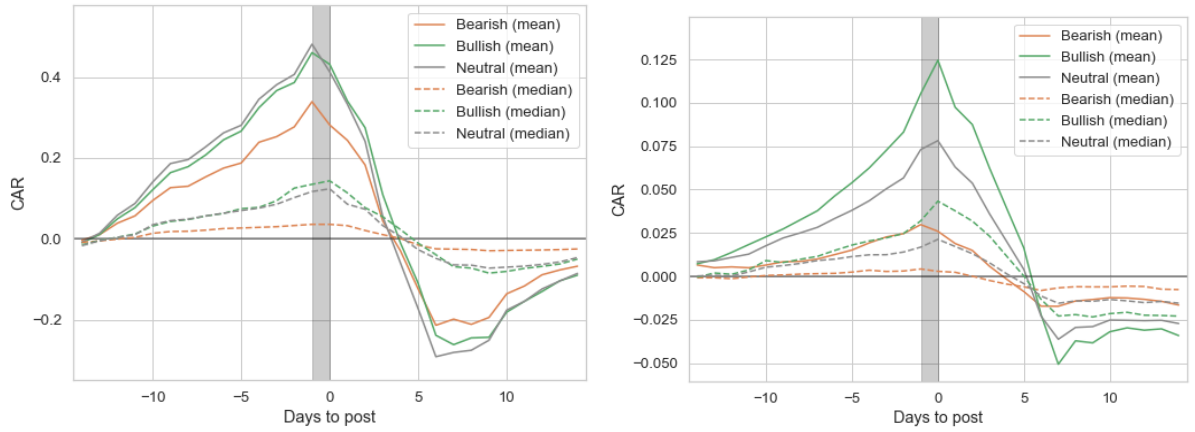
AR tends to have high magnitude in the presence of sudden price shocks (for example, a news story about a particular company), and hence it is often used for event detection in financial markets. In our case, it is useful for assessing if WSB sentiment is able to provide indication about a future price shock. We fit the CAPM model to each stock and day in our data, using a moving 180 day window. Following [233], we calculate the *seven-day cumulative abnormal return* (CAR) in stock i preceding day t:

$$CAR_{i,t} = \sum_{\tau=t-6}^{t} \epsilon_{i,\tau}. \tag{5.2}$$

[17] Once we round the timestamp of a WSB submission to the nearest future market close time, we can match the CAR for each company/time to the WSB data, along with a time series of how the CAR changed over the 14 trading days preceding and following a WSB post.

[16] See [93]

[17] where τ is a dummy variable of summation representing each time period in the event window.



(a) CAR for all posts

(b) CAR for all posts, excluding GME and AMC

Figure 5.4: Average seven-day cumulative abnormal return 14 trading days before and after post submission, grouped by post sentiment. The black bar shows the window of time when posts were written. Figure 5.4a uses all posts in the data, while Figure 5.4b excludes GME and AMC.

CAR Results

Figure 5.4 shows how the average seven-day cumulative abnormal return (CAR) changes 14 trading days before and after a submission. Figure 5.4a shows the CAR plots for all WSB posts, while Figure 5.4b excludes GME and AMC, as the short squeeze events on these stocks created exceptionally high abnormal returns. We observe the following: firstly, CAR tends to be rising up to the submission date, and rapidly declines afterwards. This suggests that WSB activity tends to follow significant price changes in the market, rather than providing a leading indicator of price changes. Secondly, the shape of the curves are quite similar, regardless of sentiment breakdown – it is primarily the magnitude of the CAR that differs. This suggests that high CAR is associated with more posts of any sentiment, although the effect is more pronounced for bullish posts.

One might attempt to infer from the sharp fall in CAR after posts that shorting stocks that are popular on WSB might be a profitable trading strategy. This inference is, however, mistaken for two reasons. Firstly, negative CAR does not necessarily imply negative returns. Secondly, and more importantly, the CAR plots suggest that WSB sentiment peaks when CAR peaks, however, predicting when sentiment will peak is challenging. If sentiment is already high, and AR increases, then sentiment is likely to continue to increase. In our Supplementary Materials, posted to our dashboard site, we consider a breakdown of the CAR plots for the twenty most popular stocks on WSB and show that the pattern presented in Figure 5.4a is most prevalent for 'meme' tickers, such as GME, AMC, BB.[18] In stocks that have a broad following outside of the WallStreetBets forum,

[18] `https://sites.google.com/view/wsbtrialsite`

such as AAPL or MSFT, the CAR appears flat before and after submissions.

5.3.5 Is information on WSB predictive of returns?

Initial Exploration

We look at the returns of several portfolios constructed from WSB data. We consider the average next day log-returns for investing into all WSB posts mentioning a single ticker according to our sentiment classifier, as well as the next day log-returns for investing into the flaired due dilligence (DD) and labeled DD posts. The results are presented in the top half of Table 5.2. We observe that investing in all submissions on WSB results in consistently loosing money, with an average next-day log-return of -0.0282 (p-value <0.001) indicating that the returns are statistically significantly different from zero at the 1% threshold according to a 2 sided t-test. The DD posts, on the other hand, which are specifically selected by moderators and our hand-labeled posts both appear to have positive returns, however the signal is weak, perhaps indicating that DD content must be further analysed for quality(a description of DD posts is available in our Data section). We perform a similar exercise after splitting the data by year and observe that the forum receives negative or nearly zero returns across all years. In the year 2021, when the infamous GameStop short squeeze occurred, the average next-day log-return for WSB posts was around -0.044 perhaps indicating that even though some forum participants made money on the incident, the subsequent hype and low-quality discussions drove a loosing strategy for the forum as a whole.

We construct several 'control' portfolios of interest, presented in the bottom half of Table 5.2. First, we look at the average, daily returns of the stocks discussed on WSB across all time – the returns appear to be very close to zero, on average, but with very heavy tails (as indicated by the kurtosis) and a positive skew. The heavy tails (high kurtosis measure) indicates that the stocks mentioned within WSB frequently experience returns that are far from the mean – extreme return values. The high positive skew, on the other hand, implies that the stocks experience more extremely high returns than low returns. This is consistent with earlier observations that WSB users are swayed by large cumulative abnormal returns of stocks, however, as noted in the previous section, the users tend to be trend followers and the forum sentiment peaks when the asset experiences a price reversion.

We consider the returns of assets on the trading day that a submission is made on WSB (or on the day before if a submission is made outside of trading hours) - labeled as 'Previous' in Table 5.2. The high, average positive log-returns indicate that the WSB forum is likely reactive to news, and is consistent with the earlier CAR analysis. We also construct a randomly selected portfolio, where we invest in a given stock proportionally to the number of times it is mentioned in WSB in non-neutral posts, however select days at

random. The results indicate a very small, positive return. We anticipate this to be driven by the large returns preceding WSB posts, rather than an inherent ability of WSB users to choose lucrative assets. The Supplementary Materials consider returns of portfolios built from the clusters within our *Topic Network* and *Submission Network*.[19]

Table 5.2: **Distribution of Log-Returns**; we present some summary statistics for the next day log-returns for a portfolio invested according to the sentiment of all submissions on WSB (All Submissions), flaired DD posts (Flaired DD) and hand-labeled DD posts (Labeled DD). We compare this to the daily log-returns of all stocks mentioned on WSB (Stock Returns), a randomly selected sample of stocks returns (Random) and the log-returns on the day submissions are made (Previous). The table presents several summary statistics, as well as the p-value for the mean of the distribution of the daily log-returns to be equal to zero (null hypothesis).

	μ	σ	skew	kurtosis	p-value	# posts
WSB Returns						
All Submissions	-0.0282	0.28	0.10	6	0.00	207,155
Flaired DD	0.0033	0.16	0.09	10	0.26	2,909
Labeled DD	0.0037	0.14	3.77	46	0.22	2,220
Control Portfolio Returns						
Stock Returns	0.0000	0.04	4.80	1,903	0.90	10,572,619
Previous	0.0242	0.34	0.50	4	0.00	207,155
Random	0.0003	0.05	1.92	109	0.01	206,631

5.3.6 Granger causal relationships on WSB

We look at whether sentiment is useful for forecasting future returns by conducting a Granger causality test.[20] A time series A is said to Granger-cause B if it can be shown that the lagged values of A provide statistically significant information about future values of B, even when lagged values of B are also included in the forecasting exercise. In other words, time series A must provide additional information about the future values of time series B, beyond B's auto-regressive terms.

Our goal is to test whether the sentiments expressed about an asset on WSB have a Granger causal relationship with the future returns of that asset. Our result is displayed as the Wald test statistic, which allows us to estimate a p-value — our confidence that WSB sentiments about an asset are related to the asset's future returns.

Sentiment Time Series

To get a time-series of sentiment, we firstly need a sentiment score for each company and date. Let $s_{i,t}^{(j)}$ represent the net sentiment of post j, which is calculated by taking the

[19] `https://sites.google.com/view/wsbtrialsite`

[20] See [102]

(softmax) output of our sentiment classifier, and calculating the difference between the bullish and bearish outputs (hence $-1 \leq s_{i,t}^{(j)} \leq 1$). We can then say the net sentiment of each stock i on day t, denoted $S_{i,t}$, is the sum of $s_{i,t}^{(j)}$ for all posts for that company and day. $S_{i,t}$, however, has a tendency to increase over time, since the forum has grown in popularity, and most posts are bullish. To resolve this, we normalise $S_{i,t}$ by dividing it by the average number of posts on the forum over the past seven days (n_t) – we denote the normalised sentiment $\hat{S}_{i,t}$. We then calculate the difference in normalised sentiment to get a value for our sentiment time series $\Delta\hat{S}_{i,t}$ where:

$$\Delta\hat{S}_{i,t} = \hat{S}_{i,t} - \hat{S}_{i,t-1} = \frac{S_{i,t}}{n_t} - \frac{S_{i,t-1}}{n_{t-1}}. \tag{5.3}$$

Granger Causality Results

We check whether our normalised sentiment measure $\Delta\hat{S}_{i,t}$ is Granger causal of daily stock price returns. We assess whether lagged values of WSB sentiment are useful for forecasting future returns by testing whether the 1,2,5 and 10 days lags of $\Delta\hat{S}_{i,t}$ help explain returns even when including 1,2,5 and 10 days lags of the stock returns themselves as described in the Appendix 11.2. Wepresent the result of our Granger causality test in Table 5.3. WSB sentiments are lagged between one and ten days with stock return. We use time-series for the top 22 most popular stocks on the forum,[21] starting from the earlier of 1 January 2016, or the first time a stock was mentioned on the forum. The removal of pre-2016 data is motivated by the small size of the forum before this point, which can result in volatile changes in our sentiment measure.

For all 22 stocks, our sentiment measure and the daily log-return time series appear stationary, which we check with an augmented Dickey-Fuller test. We run Granger causality tests at lags of 1, 2, 5, and 10 trading days. The results are shown in Table 5.3. For most stocks, WSB sentiment is not useful for forecasting returns. This is not true, however, for the meme stocks GME, AMC, NOK and BB, but this fits with the conventional wisdom that retail traders on WSB drove up the prices of these shares in early 2021.

We do get statistically significant results ($0.01 < p < 0.05$) for AMZN at a ten day lag, PLTR at a ten day lag, SNDL at a two day lag, and SPCE at a one day lag. These results are unlikely to be meaningful however – given we are running a total of 88 tests, we would expect 4 or 5 results in this range. Given that sentiment is not Granger causal of returns at other lags for these stocks, this is most likely to be a spurious result.

The most interesting statistically significant results come from SNAP, where the test statistic is strongly significant at one and two days ($p < 0.001$). This also consistent

[21] For consistency with our CAR plots, we include GE and DIS, in addition to RKT and PLTR

with the individual stock CAR results in our Supplementary Materials.[22] SNAP is not conventionally considered a meme stock, and it was considerably more discussed prior to 2021.

Table 5.3: **Results of Granger Causality Tests;** Granger causality tests for log-return for the top 22 most popular stocks on WSB. Main numbers show the Wald test statistic, with p values in parentheses.
*** Significant at 0.1% level; ** Significant at 1% level; * Significant at 5% level

	Obs	1	2	5	10
AAPL	1338	0.140	0.047	0.439	1.177
		(0.708)	(0.828)	(0.508)	(0.278)
AMC	1184	9.833	36.189	21.110	10.551
		(0.002)**	(0.000)***	(0.000)***	(0.001)**
AMD	1336	0.801	2.182	1.446	1.147
		(0.371)	(0.140)	(0.229)	(0.284)
AMZN	1334	0.224	0.183	1.735	2.033
		(0.636)	(0.669)	(0.188)	(0.154)
BABA	1339	1.055	1.329	1.260	0.887
		(0.304)	(0.249)	(0.262)	(0.346)
BB	1212	31.792	14.402	16.236	12.083
		(0.000)***	(0.000)***	(0.000)***	(0.001)***
DIS	1326	0.410	0.760	0.731	0.950
		(0.522)	(0.383)	(0.393)	(0.330)
FB	1338	0.303	0.147	0.620	0.815
		(0.582)	(0.701)	(0.431)	(0.367)
GE	1301	0.329	0.901	1.118	1.037
		(0.566)	(0.343)	(0.290)	(0.309)
GME	1284	28.820	16.922	16.206	16.749
		(0.000)***	(0.000)***	(0.000)***	(0.000)***
MSFT	1325	2.313	1.176	0.928	0.816
		(0.128)	(0.278)	(0.335)	(0.366)
MU	1295	0.120	0.128	0.098	0.920
		(0.729)	(0.721)	(0.754)	(0.337)
MVIS	261	0.306	1.658	1.322	0.632
		(0.580)	(0.198)	(0.250)	(0.427)
NIO	637	0.169	0.645	0.240	0.704
		(0.681)	(0.422)	(0.624)	(0.401)
NOK	1319	10.168	5.884	3.476	2.003
		(0.001)**	(0.015)*	(0.062)	(0.157)
NVDA	1336	0.093	0.168	0.494	0.617
		(0.760)	(0.682)	(0.482)	(0.432)
PLTR	130	0.202	0.839	1.088	2.065
		(0.653)	(0.360)	(0.297)	(0.151)
RKT	167	0.168	0.529	0.283	0.287
		(0.682)	(0.467)	(0.595)	(0.592)
SNAP	1022	13.782	6.907	2.815	1.811
		(0.000)***	(0.009)**	(0.093)	(0.178)
SNDL	327	0.312	3.500	1.810	1.212
		(0.576)	(0.061)	(0.179)	(0.271)
SPCE	391	3.901	1.169	1.298	2.017
		(0.048)*	(0.280)	(0.255)	(0.156)
TSLA	1334	0.293	0.369	0.627	0.393
		(0.588)	(0.544)	(0.428)	(0.531)

[22] `https://sites.google.com/view/wsbtrialsite`

5.3.7 Additional results

In an additional analysis, we attempt to extract long-term and short-term metrics from WSB, which we consider as potential trade signals and test their predictive power. For our short-term analysis, we focus on extracting signals from our hand-annotated sample of due dilligence (DD) posts. We fit 3,744 models predicting returns at time horizons varying from next day to 24-weeks into the future using various due dilligence post properties. Our models provide some indication of the fact that bullish sentiment in DD posts are predictive of returns at various time horizons. Among other variables that we considered, the existence of a URL (indicative of external references outside of the forum) is more likely to have a positive effect for models looking at time horizons greater than 8 weeks, and especially at time horizons greater than 15 weeks. For our long-term analysis we consider monthly and weekly changes on the WSB forum versus returns, trading volumes and volatility; we observe that changes in normalized sentiment in an asset in a given month are negatively predictive of returns in the following month. Our full results are presented in the Supplementary Materials [23] and we also display them in the Appendix 11.1. Overall, our findings suggest that WSB metrics do not have strong, predictive power for market movements, but can be used in conjunction with other factors as potential weak signals in portfolio construction.

5.4 Conclusion

What went on under the hood of the most infamous investor forum, WallStreetBets? This paper takes a data-driven look. We break down the discussion in several different ways: through topic modeling, by sentiment analysis and through a network perspective, used to map the relationships between assets. Our network approach allows us to extract groups or 'clusters' of assets which possess similar characteristics on the WSB forum. We conclude that returns for assets within large clusters (those containing many assets) are generally predicted poorly by the forum. However, niche groups of investors may have potential insights into the markets, as indicated by positive returns to WSB submissions within smaller asset clusters.

Our analysis of cumulative abnormal returns (CAR) indicates that WSB investors are, on average, reactive to market news – they follow the hype and express bullish sentiments precisely at the time when assets reach their peak and, subsequently, begin a CAR price reversal. The pattern is particularly distinct in 'meme' stocks, and less pronounced in stocks with a broad following outside of WSB. A final analysis of the average sentiment expressed within the forum indicates that a Granger-causal relationship exists between

[23] `https://sites.google.com/view/wsbtrialsite`

changes in average sentiment expressed about an asset at future asset returns in several popular assets on WSB.

The scope of our study leaves several promising directions for future research. A key area for investigation lies in the appropriate way to analyse a non-stationary dataset, such as WSB, and its predictive value across different time periods. Additionally, this analysis presents several methods for of extracting signal from text and network interactions. As unstructured datasets are becoming more and more common, it may be valuable to investigate and standardize additional methods which can be useful for forecasting.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Wisdom of the crowds or ignorance of the masses? a data-driven guide to Wallstreetbets
Publication Status	X Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Semenova, V., Gorduza, D., Wildi, W., Dong, X., and Zohren, S. (2024). Wisdom of the crowds or ignorance of the masses? a data-driven guide to wallstreetbets. Journal of Portfolio Management, 50(4).

Student Confirmation

Student Name:	Dragos Gorduza		
Contribution to the Paper	As joint first author, Dragos Gorduza contributed to the Natural language processing analysis of the topics and sentiments present in the dataset as well as clustering the topic networks of the different firms discussed by the forum participants. Additionally, he performed data collection and analysis of the returns of the clusters generated by text and network analysis. Dragos also contributed to the introduction motivation and conclusion sections.		
Signature		Date	30 August 2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Xiaowen Dong, Associate Professor		
Supervisor comments I can confirm Dragos has made significant contribution to the paper above			
Signature	Xiaowen Dong	Date	10/11/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

PRET: Domain-Adapted Transformer for Machine-Assisted Prudential Supervision

Analysis of unstructured text is a key aspect of everyday financial supervision operations run by regulators worldwide. The emergence of transformer-based language models have opened the possibility of improving the financial supervision process, by assisting supervisors with labour-intensive and time-consuming tasks of information retrieval across the wide range of complex corpora like financial rules and regulations. This paper introduces Prudential Regulation Embedding Transformer (PRET), a novel domain-adapted transformer encoder model tailored for information retrieval on topics relating to financial regulations. To train this model, we address the scarcity of high-quality training financial regulations text datasets with a dedicated pipeline to web-scrape and pre-process the Basel Framework into machine-readable format, which is then coupled with corresponding large language model (LLM) generated text as synthetic training data pairs for each rule in the Basel Framework. We evaluate the performance of our model on this domain-specific information retrieval task against commonly used state-of-the-art (SOTA) models. We show how our proposed model outperforms existing benchmarks while being substantially cheaper to train than previous methods. We discuss the implications of our findings for the design of better regulatory technology models across jurisdictions.

6.1 Introduction

Financial regulations play a crucial role in the functioning of modern economies, by establishing a structured set of rules and standards that govern the operations of financial institutions. These regulations are essential for ensuring the stability and reliability of the banking and financial sectors, fostering consumer trust in the industry and driving growth throughout the broader economy. Regulatory bodies, such as the Prudential Regulation Authority (PRA, part of the Bank of England) in the United Kingdom, are at the forefront of implementing these regulations and is responsible for supervising around 1,500 financial entities, which includes a broad spectrum of organisations such as banks, insurance companies, credit unions, and investment firms. Supervising firms requires in depth understanding of financial rules and an ability to search through a firm's disclosure requirements to check for anomalies or violations and match them to the relevant rules.

The increase in specificity and volume of financial regulations in the wake of the 2008 financial crisis is embodied in the changes of quantity and complexity of the rules in the Basel III framework, a cornerstone set of financial regulatory rules [15]. While necessary to safeguard financial market soundness and stability, this has translated to additional demands on prudential supervision operations of both regulators and regulated firms. Many institutions in this space are therefore turning to regulatory technology (RegTech) to answer this growing need for an efficient, scalable, and accurate text analysis pipeline.

In light of recent developments at the intersection of natural language processing (NLP) and deep learning, the widespread adoption and utilisation of large language models (LLMs) could radically transform the RegTech landscape as cutting-edge models are poised to redefine the way we approach text data used in support of prudential supervision. The emergence of transformer-based language models has paved the way towards the development of powerful families of LLMs—such as Llama [223], OLMo [103], Mistral [200], and GPT [181]. Many of these models have demonstrated impressive performance across a wide-range natural language tasks; in some cases, reaching performance level comparable to humans [51]. However a current model to leverage the good performance of transformer models for financial regulation information retrieval is still lacking due to a lack of training corpus.

In this paper, we make several contributions towards the application of transformer-based language models in RegTech. We introduce a novel transformer encoder model designed specifically for the financial regulation field, named Prudential Regulation Embedding Transformer (PRET), that is fine-tuned for searching and retrieving information within the realm of financial regulations. The development of PRET is supported by the creation of a novel dataset tailored for the financial regulation domain, another key contribution of this paper. This dataset includes two main parts: firstly, we have collected

rules and regulations from the Basel Framework through web scraping, which are then processed and into machine-readable corpus; secondly, we have produced a collection of artificial LLM-generated texts for each rule from the Basel Framework. These texts are crafted to correspond closely with their respective Basel rule to cover various aspects natural language tasks that might applied on those rules (e.g. questions and answers, summaries, paraphrases, etc.). By focusing on a domain-specific dataset, PRET gains a much better understanding of the specific language used in financial regulations, such as those found in the Basel Framework, which may be overlooked by more general-purpose models.

6.2 Literature review

6.2.1 Information Retrieval

Information retrieval techniques are a field of natural language processing which aim to retrieve the best match for a given query. They form the backbone of many applications in fields as diverse as healthcare [222], biochemistry [137] or search engine optimisation [10]. These techniques can be thought of as operating in two successive steps : a representation step where the documents and queries are given a vector-based representation, followed by a ranking step where the candidate vectors are retrieved based on cosine similarity to the query. This process is described in Figure 6.1. The largest diversity of methods lies in the representation step which maps text to a vector. The traditional approaches focus on named entity recognition [250], Bayesian methods [144] and term-frequency embeddings to represent text [107]. However, with the advent of deep learning methods and specifically word embeddings [169], best performing existing models for information retrieval focus on transformer-based models to generate embeddings [173]. These transformer based models are primarily encoder based models such as the Sentence-BERT family of models [197] which use an aggregation method to average the BERT embeddings of a sentence into a representation of that sentence. The increasing popularity of large language models has also led to the development of orders of magnitude larger text embedding models which use the intermediary layers of decoder language models such as the GPT [181], Mistral [200] or Llama families [223]. Yet, despite these advances, these models are trained on large text corpora that general in nature, resulting in their limited utility for RegTech use-cases.

Limits of NLP in financial regulation

Financial text-mining is the discipline within natural language processing which attempts to extract useful information from the wide variety of financial texts. This field has seen

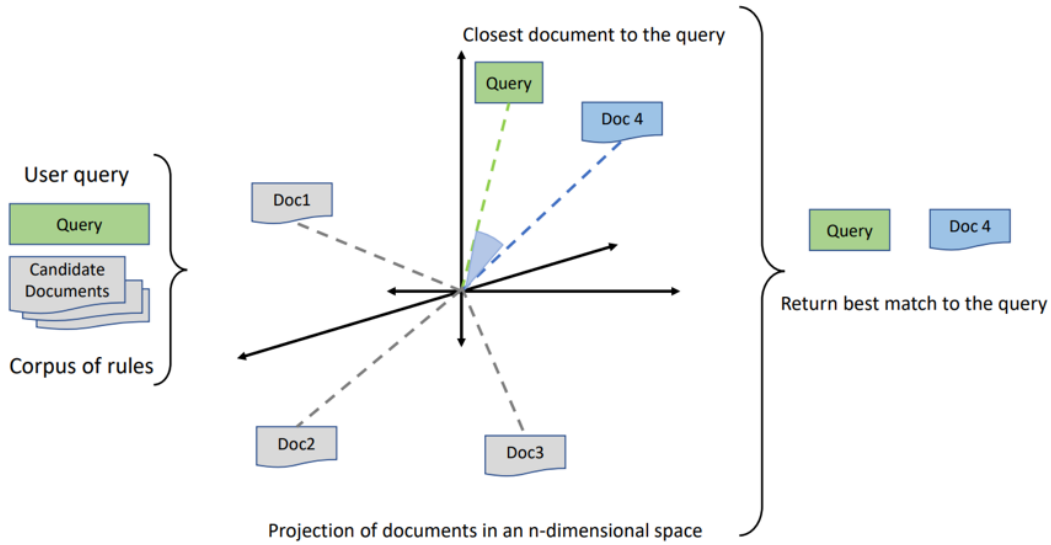


Figure 6.1: Information retrieval task

a large gain in popularity since the initial dictionary-based models for financial sentiment detection of Loughran & MacDonald [154]. Recently, advanced transformer-based models such as BERT [81], FinBERT [21] or BloombergGPT [241] have emerged and are used by researchers and practitioners for tasks as varied as topic modelling [28], text summarisation [10] and especially sentiment detection [250, 21]. These models are created by a process known as domain-adaptation : a previous model architecture such as BERT is pre-trained on a specific corpus of interest. In the case of FinBERT for instance that corpus consisted of SEC reports and news headlines. Financial regulations on the other hand, being a highly niche domain, contains vocabulary and terminologies that are different relative to standard everyday language. This is a problem when attempting to apply these models to a regulatory context as, financial regulations differ from regular financial literature. These are 'complex' texts from the standpoint of their vocabulary, their syntax and interconnected networks of citations [15]. For instance, while the terms 'basic-approach to credit value adjustment' and 'standardised-approach to credit value adjustment' may seem similar in plain English, these terminologies have very specialised and different meanings under the Basel Framework. Additionally, the risks of model hallucination, where LLMs could potentially generate factually inaccurate outputs further compounds their adoption in RegTech, a field with low tolerance for inaccuracies. Since these models rely on parametric memory for text generation, generating context-specific outputs will be difficult without access to external source data. Furthermore, the very large document size that characterises supervisory contexts and financial supervision in particular can lead to language models forgetting the middle of the text as larger large language models have been shown to do [151].

Retrieval pipelines and domain adaptation

Several methods have been proposed to circumventing the problem of domain-adaptation. Amongst the most popular approaches typically involves model fine-tuning with domain-specific corpora or implementing some form of retrieval augmented generation (RAG) pipeline, which contextualises the input prompts of LLMs with relevant information from the corpus [145]. In each of these case, the root of the problem for application in RegTech is a lack of specialised training data for the financial regulation domain. Even in the case of RAG, which typically utilise transformer encoders to generate corpus and query embeddings for a semantic search retrieval pipeline, deploying a transformer encoder that is fine-tuned on domain-specific corpus would greatly enhance the retrieval quality. Given findings from studies such as Balaguer et al. [32] point to the importance of RAG pipelines, enhancing this semantic retrieval pipeline will further improve the applicability of RAG for RegTech.

Lack of datasets and models for Reg-IR

Researchers outline another limitation with building a model for Reg-IR, the lack of high-quality retrieval data in this domain [107, 32]. Information retrieval datasets difficult to obtain as they require hand labelling of user queries and what counts as correct and incorrect answers. However these datasets are crucial for the second stage of a Reg-IR model : fine-tuning to an information retrieval task once a model has been domain adapted. This fine-tuning is crucially important in other very specific domains such as biochemistry as it allows pre-trained models to avoid performance collapses as a result of a model understanding the vocabulary but not the task after pre-training [222]. Most commonly used IR models from the Sentence-BERT family have been trained on Bing user queries from the MS-MARCO dataset [30]. In the financial domain ,the only open-source IR dataset is FiQA [1], a dataset of queries submitted on the advice website Quora included in the commonly used Benchmark for Zero-shot Evaluation of Information Retrieval (BEIR) [222]. Neither of these datasets would be useful for a regulatory context. For example, in the FiQA dataset, to the query 'should I do on the job trainning?', the dataset offers the Quora forum user's answer 'I'm not saying I don't like the idea of on-the-job training too but...'. Little work has gone into designing datasets with a specific financial regulation aspect, the most advanced of which, the Global Banking Standard (GBS), collated frequently asked questions from the Basel 3 rules [214]. However, the GBS was limited by its' small size and a coverage limited to frequently asked questions in the Basel III corpus. Existing work to design retrieval models for financial regulation is limited by the pre-existing focus of financial text mining models on tradeable signals [107].

[1] https://sites.google.com/view/fiqa/home

This is why the existing research into financial information retrieval consist of studies of legal transpositions [61] and database searches of SEC filings. Table 6.1 describes the way our PRET dataset differs from what other datasets are commonly used for training retrieval and financial NLP models by incorporating both financial retrieval-style queries and answers as well as containing financial rules and not more generic financial text.

Dataset name	Financial domain retrieval	Financial rules
SEC Filings	No	No
FinQA	Yes	No
Global Banking Standards	No	Yes
PRET-dataset (ours)	Yes	Yes

Table 6.1: How our corpus fills the research gap for retrieval data in financial regulation

The research presented above suggests a research gap between advances in NLP and the current under explored domain of financial information retrieval. Current transformers outperform existing models in other NLP tasks and are being held back in the regulatory space by a lack of suitable datasets and attention. We propose leveraging the ability of transformer models to best represent financial rules in a regulatory context and fill the research gap of data and model fit for a financial regulation retrieval. We aim to do so by presenting a comprehensive dataset built from an existing set of publicly available corpora and also present rigorous evaluation criteria for any model on this task. We test the hypothesis that applying a financial regulation domain adaptation and a fine-tuning information retrieval task to a transformer model would allow the model to perform better than existing general domain baselines on a Reg-IR task. We further detail our hypothesis and model setup in the Appendix Section 12.

6.3 Methodology

6.3.1 Data Collection

The primary aim of the pipeline is to collect a high-quality textual dataset on financial regulation, capturing relevant information such as vocabulary, terminologies, linguistic patterns and relationships within the field of interest. To this end, the Basel Framework is our primary candidate, given its central role in anchoring financial regulation standards for banks globally. The full Basel Framework is hosted on the web page of the Bank for International Settlements [2], which consists of 14 standards covering various aspects of prudential regulation of banks shown in the table 6.2.

[2]https://www.bis.org/basel_framework/

Acronym	Standard names
SCO	Scope and definitions
CAP	Definition of capital
RBC	Risk-based capital requirements
CRE	Calculation of RWA for credit risk
MAR	Calculation of RWA for market risk
OPE	Calculation of RWA for operational risk
LEV	Leverage ratio
LCR	Liquidity Coverage Ratio
NSF	Net Stable Funding Ratio
LEX	Large exposures
MGN	Margin requirements
SRP	Supervisory review process
DIS	Disclosure requirements
BCP	Core Principles for Effective Banking Supervision

Table 6.2: Standards within the Basel Framework

We then utilised web scraping frameworks Selenium [3] and BeautifulSoup [4] to parse through the standards, taking note to extract all information available within the site (including information within toggles such as footnotes and FAQs). For each rule, we assign an identifier corresponding to their number assignment within the Framework itself and extract various metadata information of interests, such as the rule's standard, chapter, section, links, and type of text (e.g. header, body, footnote, or FAQ). These extracted rule and its corresponding metadata are stored in JSON format; Figure 6.2 is an example of what an entry looks like at the end of this collection process.

6.3.2 Model setup

We build the PRET model from a set of building blocks whose performance we try to evaluate in order to measure the impact of the different training strategies on the information retrieval performance. We select two different text embedding transformer methods : BGE-Large [243], the best-performing general purpose encoder-only text embedding method and RoBERTa-large [252], a popular workhorse model for transformer text embedding. These models are domain adapted by a strategy known as masked-language modelling. This approach makes the model predict the correct word in a masked sentence by learning probabilistic mappings within words in a given domain. This step is a standard domain-adaptation step used in previous financial domain pre-training steps such as FinBERT [21]. To verify the impact of pre-training on our final performance, each of the two pre-trained models is compared to a non-pre-trained model in the evaluation.

[3] https://www.selenium.dev
[4] https://beautiful-soup-4.readthedocs.io/

```
{
    "id": "BASEL_BODY_CAP_10.24",
    "text": "Total capital instruments (ie Tier 1 and Tier 2 capital instruments)
        issued by a fully consolidated subsidiary of the bank, whether wholly or
        partly owned, to third-party investors (including amounts under CAP10.21 to
        CAP10.23) may receive recognition in Total Capital only if the instruments
        would, if issued by the bank, meet all of the criteria for classification as
         Tier 1 or Tier 2 capital.",
    "metadata": {
        "chapter": "CAP10",
        "links": [
            "CAP10.21",
            "CAP10.22",
            "CAP10.23"
        ],
        "type": "body",
        "header_text": "Minority interest (ie non-controlling interest) and other
            capital issued out of consolidated subsidiaries that is held by third
            parties",
        "rule_id": "CAP10.24"
    }
}
```


Figure 6.2: Example entry of web scraped Basel Framework rule

The summary of the different pre-training and fine-tuning combinations leading to our four tested models is described in Table 6.3.

Model	BGE	RoBERTa
pre-trained	BGE_PTFT	RoBERTa_PTFT
Non-pre-trained	BGE_FT	RoBERTa_FT

Table 6.3: Naming conventions of the different models

6.3.3 Fine-tuning using generative pseudo labelling

Generating a suitable dataset to train a language model for information retrieval of financial regulatory rules is difficult due to the lack of labelled data for what constitutes a reasonable query and answer. The exception to this are the Frequently asked Questions section which we use as gold-level data for evaluation as discussed in Section 6.4.The Basel rule set is a large unlabelled text corpus which we aim to generate labels usable for training. To generate a suitable dataset to fine tune PRET without pre-existing labelling such as the one needed to train PRET we use a widespread technique used in information retrieval: Generative Pseudo Labelling [238] whose functioning is described in Figure 6.3.

This involves instructing a language model to generate a ‘query’ that a user might ask about a given paragraph (referred to as the ‘positive’) such as paragraph MAR 50.1. An example ‘query’ would be a question such as ‘What is CVA risk?’. All the other rules in the rule set apart from the ‘positive’ referred to as the ’negatives’ are ranked by how close they are to the ‘query’. The most similar passages which are not the positive are referred

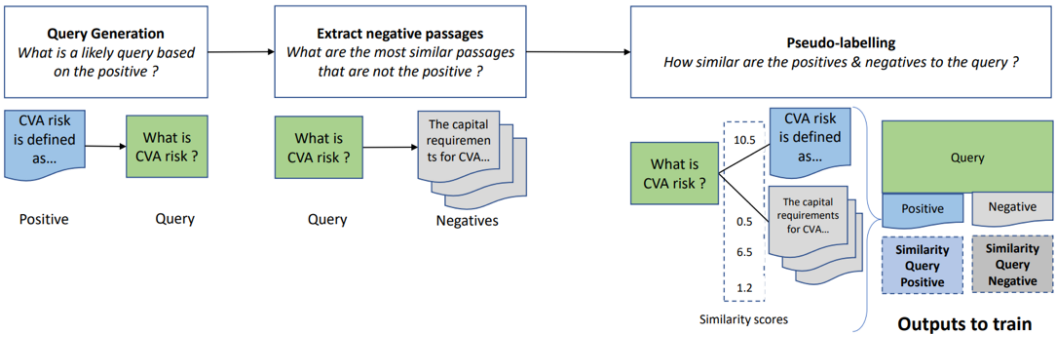


Figure 6.3: Generative Pseudo-Labelling Pipeline

to as the 'negatives'. Each triplet is scored by how close the query is to the positive relative to the 'best possible' guess – ie. the 'negative'.

GPL is a framework which generates Query, Positive and Negative triplets from initially unlabelled text. We use these triplets in a fine-tuning task where our embedding model, whether BGE or RoBERTa is tasked with learning a representation of the query, positive and negative such that the embedding of the query is closest to the embedding of the positive than to that of the negative. We give a visual representation of this scoring strategy is called the margin score in Figure 6.4.

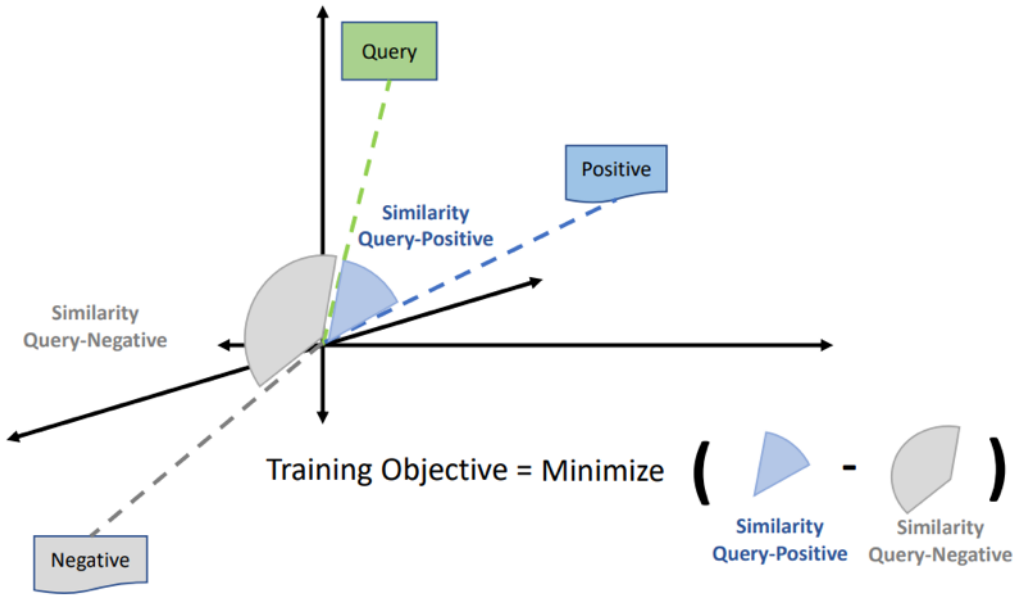


Figure 6.4: PRET fine-tuning

We report the results of the fine-tuning performed via GPL for the 2 text generation approaches which we experimented : T5 and Mistral 7B. T5 was selected as it is the default recommended approach proposed by the GPL team. Mistral 7B were selected as it is the strongest open-source generative frameworks available, which allows us to replicate our results. We report the time and estimate cost of the generation for each of the two

methods we used in the Appendix in Table 12.1

6.3.4 Other methodologies

We tested the PRET model's performance on the evaluation task against a suite of alternative models. We selected the models based on their performance on the Massive Text Embedding Benchmark (MTEB) a well-regarded benchmark for information retrieval models [173]. We selected the BGE-large [243] and all-mpnet-base-v2 [215] models as a benchmark for how well the all around best performing encoder-only methods would perform. We included these models as they show the best performance across all tasks in the MTEB. We also included fin-mpnet-base, a model similar to all-mpnet-v2 but which was specifically fine tuned on the FiQA dataset and as such purports to better represent financial text as evidenced by its' better performance on the FiQA testing task than any other model. Although we have pointed to the difference in setting between FiQA and our financial regulatory questions, fin-mpnet-base is the best performing model on the only existing public financial text retrieval benchmark and as such its inclusion is warranted as a comparison point for how useful the specific financial regulatory domain pre-training is to final performance. Lastly, we include the Salesforce Mistral-7B-SFR embeddings [5] as they represent the best performing generative model in information retrieval tasks according to the MTEB benchmarks. Moreover, it represents a good benchmark for a model that is like GPT, Llama and Falcon, primarily a text generation model used for text embedding. A good performance by this model would suggest that using generative general purpose models would be a better way to design Reg-IR pipelines.

6.4 Evaluation

We evaluate PRET by comparing its performance on a suite of information retrieval tasks. Each model is evaluated on three similar, though aligned tasks. The first task is frequently asked questions, these are questions asked by financial professionals to the BIS about the Basel rules. The task in this case is, given an FAQ, retrieve the correct answer (ie the rule the FAQ is being asked about). For instance, in Figure 6.5, the model would be presented with the question on the top and asked to return the correct answer from the bottom out of all the candidate FAQ answers.

The second task concerns footnotes, these are additional notes appended to rules in the Basel dataset. The task in this case is, given a footnote, retrieve the correct answer (ie. the rule the footnote is referring to). For example in Figure 6.6, the model would be presented with the footnote 'Unfunded credit protection [...].', and expected to return

[5] https://huggingface.co/Salesforce/SFR-Embedding-Mistral



Figure 6.5: Evaluation by Frequently Asked Questions

rule 30.7 as the best match. This task tests the ability of the model to understand which footnote contains relevant information for which rule.



Figure 6.6: Evaluation by Footnotes

The final task leverages the reference links present in the rules. For instance the MAR.50.5 rule in the Basel 3 rule set, refers to rules CRE54.14 to CRE54.16 in the same document. This task is an extension of the footnotes' task measuring the ability of the model to match relevant content to a statement. A correct match would indicate a model was able to infer the right links between the query (original rule) and the response (linked rule).

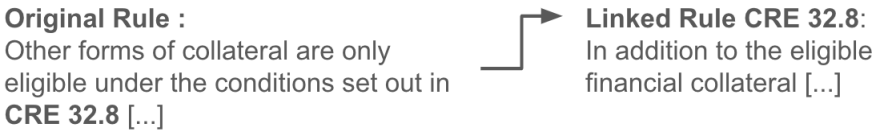


Figure 6.7: Evaluation by links

We compare PRET's performance against other existing competing approaches for information retrieval in the financial domain. In order to prevent unfair comparisons, PRET is not trained on any of the evaluation datasets. Specifically, the GPL fine-tuning strategy presented above does not include the text we use for evaluation. This allows it to be comparable in a zero-shot -ie with zero prior exposure – to all the other competitors such as all-mpnet-v2 or BGE-large. We report and compare performance on the top-1 answer for every evaluation task, in other words, we only consider the retrieval a success if the model returns the correct answer to a query as the first choice. In information retrieval, various thresholds are acceptable and used in the literature. We focus here on the top-1 response in our evaluation as in regulatory information retrieval the quality of the search is paramount given the legal and administrative sensitivity of supervision. We report the performance of the PRET model as a percentage score of the number of correctly returned answers to the total number of answers returned.

Beyond comparing to the benchmark models, we also evaluate in an ablation study, the impact of the generation pipeline used in the GPL stage on model performance.

6.5 Results

Model	Basel-FAQ	Basel-FOOTNOTE	BASEL-LINK
all-mpnet-v2	37.3	20.38	3.39
BGE-base	34.0	17.54	3.86
fin-mpnet-base	36.0	16.55	3.12
Mistral-7B-SFR	26.0	15.24	3.76
PRET_RoBERTa_FT	32.0	12.80	4.36
PRET_RoBERTa_PTFT	36.5	18.81	4.12
PRET_BGE_FT	33.0	12.80	3.30
PRET_BGE_PTFT	**37.6**	**24.05**	**5.29**

Table 6.4: Model performance on the different evaluation dimensions
*This table shows the result of a single run for each tested model with the result of the best model in **bold**.*

Table 6.4 presents the performance of the different benchmark models and PRET on our evaluation framework made of FAQs. The best model is the PRET_BGE pre-trained and fine tuned scoring 37.4% on FAQs, 23.0% on footnotes and 5% on linked passages. This means that it returns the correct response as its' top 1 choice 37.6% of the time The best performing model is the PRET_BGE across all dimensions after pre-training and fine-tuning. Models for both BGE and RoBERTa perform better when they are pre-trained before fine-tuning as opposed to the case where they are just fine tuned. The best-performing benchmark method on FAQs was all-mpnet-v2 with 37.3%. All our models perform significantly better on FAQs with an average of 34.05%, as opposed to footnotes with an average performance of 19.05% and links with an average performance of 4.45%.

Model	GPL	Basel-FAQ	Basel-FOOTNOTE	BASEL-LINK
PRET_RoBERTa_PTFT	T5	34.5	18.81	4.12
PRET_BGE_PTFT	T5	35.5	23.05	4.91
PRET_RoBERTa_PTFT	Mistral-7B	36.5	19.70	4.00
PRET_BGE_PTFT	Mistral-7B	**37.6**	**24.05**	**5.29**

Table 6.5: Model comparisons for different generation strategies
*Performance comparison for models fine tuned using different synthetic data from the GPL pipeline. Best model in **bold**.*

Table 6.5 shows the performance of the different variations of PRET. Each of them is fine tuned using the GPL procedure described in Section 6.3.3. We can observe that the PRET models trained using an information retrieval task generated using the Mistral

generation approach performs better than those generated using T5. Every method using Mistral to generate text performs better than the PRET models fine tuned on the T5, whether that is BGE or RoBERTa. Models using BGE embeddings outperform models using RoBERTa.

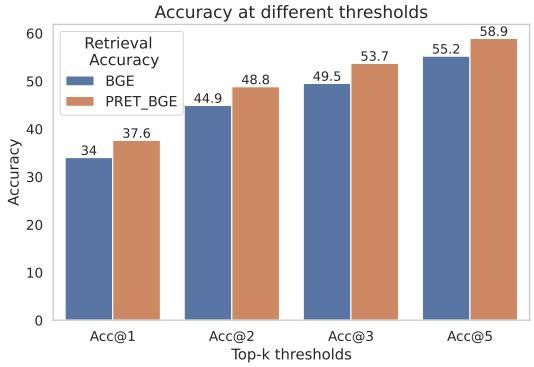


Figure 6.8: Model improvement from specific BASEL corpus at different thresholds

Figure 6.8 shows the increases in performance from training the basic BGE model with the BASEL dataset at different thresholds of accuracy. If the result is in the top 1, 2, 3 and 5 returned closest matches respectively. We can observe that there is a permanent difference of over 3 points of accuracy between the base BGE model and the PRET_BG_PTFT. This further reinforces the hypothesis that it is the specific dataset of financial rules enriched by GPL which led to this consistent increase in performance.

6.6 Discussion

The strong performance of our model on the evaluation tasks shows the the PRET approach outperforms the existing benchmarks. This confirms our initial hypothesis which suggested that domain and task adaptation can lead to a better information retrieval model for the specific topic of financial regulation. Moreover, the good performance of text embedding models from encoder-only transformers over the much larger SFR-embeddings built from decoder-only language models suggests the persistent relevance of encoder-only models for the specific task of Reg-IR. This contrast with the current popularity of decoder agents such as ChatGPT which are being used for various general-purpose applications. We argue this could be a fruitful direction of research in this field to obtain cost-effective fine-tuning and pre-training processes. For instance, our PRET model, even the one based on the slightly larger BGE architecture, is orders of magnitude smaller than existing large language models as it is only needed for this specific use case. Meanwhile, the all round better performance of BGE over the fine tuned but not pre-trained BGE model suggests that the pre-training stage allowed the model to increase it's performance up to and beyond SOTA levels. This suggests that the effort to create a comprehensive clean

training corpus from the Basel rules yielded a positive effect on performance as it has allowed us to not only fine tune but also pre-train PRET.

The better performance of PRET on the footnotes and links tasks also suggests that the model can encode complex pre-existing knowledge in its representation. By having access to financial regulatory language in its' training data, the model has been able to understand what information a rule is missing or referring to. The results suggest PRET is able to make limited inferences about financial rule's presupposed knowledge better than other competing models. This would make the model able to match user queries which may rely on pre-existing information or inferences which could only come from a good knowledge of entities in the text. The comparatively lower performance on footnotes and especially links as opposed to FAQs can be explained by the higher difficulty of the task : rules are longer than the questions asked in the FAQs, and encoding longer text is more challenging for language models. Moreover, the search universe for FAQs were all existing answers, the search space for footnotes and links was much higher. Thus a purely random model would have yielded on FAQs but only on footnotes and on links.

Our findings also show that using larger more powerful and modern text generation models for the Generative Pseudo Labelling pipeline improves performance for adapting text-embedding models to regulatory contexts. This goes against the findings in other information retrieval tasks such as general purpose question answering which argue that the generating quality of the queries does not impact the performance of the fine tuned model. We show that using a different generation strategy compared to the default recommended T5 allows our model to retrieve better. The largest model SFR performing better than T5, the smaller one and the model fine tuned using the T5-generated data does not outperform the all-mpnet-v2 baseline. Looking in depth into the data, we see that the synthetic queries created at the generative step of the GPL pipeline with the larger model (Mistral) are on average longer, more complex syntactically and grammatically. These results suggest that when attempting to design Reg-IR models using the GPL pipeline, researchers could leverage the improvements from existing innovations in the fields of large language models to produce better synthetic training data by using more recent models with more representation power.

Regulatory information retrieval has not seen much interest from the financial NLP community. Our work is the first to collate a dataset, set out an evaluation task and train a text-embedding language embedding model for the specific domain of financial regulation. We evaluate information retrieval models on the task of answering queries from the Basel III rule set. We show that our model PRET, outperforms existing baselines comprised of SOTA methods for information retrieval. We argue these results indicate that domain adaptation is useful to training an information retrieval system for financial supervision. This first step is crucial for the future design of automated financial supervision as existing

models need to learn how to correctly represent supervisor queries in a practical use case.

Our model also displays certain limitations which should also be acknowledged. Its' performance, although better on average than the competing approaches still performs markedly unevenly. It does much better on more dialogue-sounding text in FAQs as opposed to the footnote-to-rule or rule-to-rule other 2 retrieval tasks. While this can be partially explained by the size of the search space and on the difficulty of language models to keep track of the longer texts such as whole rule passages, this also points to a clear limitation of the model and suggests that in an applied setting it would need to ensure that queries to it are short and in natural language. This limitation could be mitigated by prompting the model's GPL pipeline to generate queries which sound less like natural language.

6.7 Conclusion

This paper present PRET, the first encoder model adapted to the domain of financial regulatory rules as well as task adapted to a retrieval use case. We describe the related work, data and methods to build the PRET model and discuss the results of our approach. In future work, we would also aim to explore two closely related dimensions of regulatory information retrieval. The first is building knowledge graphs for financial regulation retrieval, drawing on insights from NLP and graph machine learning where a transformer model was pre-trained using a knowledge graph to make it learn specific highly distinct vocabulary in a scientific and biology context. We would aim to extend this idea to finance. The second approach is focusing on the data streams and expanding our model's training data. We plan to include other international rule sets from other regulating institutions in the training dataset and include the FAQ, footnotes and inter rule linkages into an evaluation task. Extending our work this would allow us to get closer to facilitating the crucial demands placed on supervisory departments through better and more reliable automated regulatory NLP.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Prudential Regulation Embedding Transformer (PRET) a domain-adapted model for prudential supervision
Publication Status	x Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Dragos Gorduza, Adam Muhtar 'Prudential Regulation Embedding Transformer (PRET) a domain-adapted model for prudential supervision' (2024) Accepted at the 5th ACM International Conference on AI in Finance Workshop on LLMs and Generative AI for Finance, November 2024, New York (non archival venue)

Student Confirmation

Student Name:	Dragos Gorduza		
Contribution to the Paper	Wrote manuscript, performed data collection and analysis, interpreted results.		
Signature		Date	30 August 2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Xiaowen Dong, Associate Professor		
Supervisor comments I can confirm Dragos has made significant contribution to the paper above			
Signature	Xiaowen Dong	Date	10/11/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 7

Common Causes, Common Risk?

Price-based correlations are the standard proxy for joint risk exposure in financial markets, yet they cannot represent much of the common shocks which structure relationships between firms. We propose a text-driven alternative that embeds companies in a directed, causal-narrative network extracted from financial news and SEC 10-K filings. Using the Relatio pipeline, we parse a Reuters financial news corpus and the filings of S&P 500 firms to extract micro-narratives of the form "A acts on B". This yields a semantic graph linking macroeconomic variables, sectoral actors and individual firms. We evaluate the measure on its ability to predict out-of-sample pairwise return correlations in 2007. Our pilot results indicate that causal narratives contain latent, non-linear information about shared risk that is largely orthogonal to price history and surface-level text similarity. Incorporating narrative economics into risk models thus offers a tractable path toward finer portfolio diversification, earlier detection of contagion channels and richer interpretations of systemic vulnerability.

7.1 Introduction

In financial markets, common sources of risk are often rendered and measured in the form of correlation [159]. While there exists a substantial body of literature on the structural impact of correlation in financial markets, the fundamental intuition rests primarily on estimating the exposure to joint risk sources through the correlation matrix. However, price correlations alone fail to capture the entirety of the common risk environment within which firms operate [161]. This paper argues that a more comprehensive understanding of joint risk exposure can be achieved by incorporating a causality network extracted from narrative parsing of news and financial filings in economic models.

Traditionally, high correlation in stock prices has been regarded as a strong signal for increased market volatility. Correlations reflect the common movements of opinions within the body of investors and as such a highly concentrated market or portfolio is exposed to similar opinion shifts in reactions to the same type of events. Moreover, correlation structures can provide insights into the roles played by different companies, indicating whether they are likely to act as spreaders of financial shocks [161, 86]. It is therefore important to attempt to build a correct estimate of future correlation profiles across a market. However, it is crucial to recognise that price correlations do not encompass the full spectrum of common risk sources and that additional sources of data are needed for this. Textual resources describing the financial health and business plan of a company or commenting on the overall state of a business is a good source of such additional data [158]. It allows the revelation of supplementary data on what common sources of risks companies can be exposed to. Companies may for instance share similar correlations because they have the same supplier or are exposed to the same type of government orders. Although prices eventually incorporate this information, textual similarities can offer significant advantages in capturing new information, ranging from faster-moving sources like news and social media to more fundamental aspects such as SEC filings. Examples of potential applications include building diversified portfolios by capturing the "edges" of a textual similarity network and detecting short-term rapid sentiment events for pairs trading.

The underlying premise of this paper is that companies that exhibit greater textual proximity are more likely to be affected by the same macroeconomic shocks than their more distinct counterparts. Text similarity, primarily measured through document comparisons and context analysis, emerges as a critical metric in understanding joint risk exposure. However, this method of measurement has notable limitations, especially when it comes to identifying common sources of risk.

Firstly, if entity overlaps are used as a measure, the high dimensionality of the total number of words limits the quality of the measure. Secondly, word-based embeddings offer a range of similarity measures, each yielding different results. None of these measures

explicitly account for directionality or relationships they only consider mutual presence. Even contextualised embeddings, such as BERT, fail to capture the specifics of who is influencing whom. Consequently, encoding complex relationships found in textual statements, such as the direction of interdependence between General Electric and Ford's share prices on one hand and the price of oil on the other, becomes challenging. To capture these interlinkages we propose a novel methodology based on narrative economics and narrativity analysis. We use the Relatio Ash et al. [23] pipeline to generate a semantic graph filled with narrative statements of the type 'A does something to B'. This narrative graph captures links between entities which we use to get an inter company distance measure that is then fed as a regressor to explain the following year's correlation. Narrativity in economics involves analysing the economy through narrative statements that capture the interactions between economic entities. By incorporating narrative elements, such as relationships and causal links, we can enhance our understanding of inter-firm dynamics and measure joint risk exposure more effectively.

We present the results of a pilot study on news and financial filings data which attempts to provide an estimate to explain the out of sample forecast of correlation. We interpret the encouraging results as an indication that our semantic graph can allow us to map joint risk exposures more easily.

7.2 Literature

Historical inter-company correlation has long been recognised as a valuable metric in understanding financial markets and predicting overall market volatility. Two companies' joint stock price movements can reflect multiple features of their place within a market. Marti et al. [161] highlighted that correlated stock prices reflect the synchronised selling and buying activities of investors. But correlation can also In addition to its predictive value, the inter-company correlation provides insights into group opinion shifts and the collective behaviour of investors . When a group of investors shares similar opinions or reacts to market events in unison, it can lead to increased correlation among the stock prices of the companies they invest in. This highlights the influence of collective sentiment and the importance of understanding group dynamics in financial markets. The correlation matrix itself has become a valuable tool in quantitative finance, and as such learning good predictors for it is a useful task. Mantegna [158] found that high inter-company correlation can serve as a predictive indicator of increased market volatility. Furthermore, the correlation between stock prices of financial institutions has garnered significant attention due to its implications for financial market crises. Dungey and Gajurel [87] examined the correlation among financial institutions' stock prices and observed that high correlation can reflect contagion effects during market crises. In other words, when

financial institutions experience simultaneous declines in their stock prices, it often signifies a systemic risk spreading throughout the financial system.

Research has shown that highly central companies in correlation networks often act as propagators of disruptions. Mezei and Sarlin [168] both explored the network structure of correlations and found that certain companies play a crucial role in transmitting shocks throughout the system. These highly central companies, sometimes referred to as "high-spreaders," have the potential to amplify and propagate financial disturbances, indicating that their actions can significantly impact the stability of the overall market. Haldane and May [109] Observed in that spirit that the entire financial market due to these covariance structures turns into something akin to a complex adaptive but fragile ecosystem exposed to catastrophic collapse due to simultaneous exposure to common risk sources. Acemoglu et al. [3] also emphasised the importance of identifying structural high-spreaders within correlation networks. These companies, due to their strategic positions and extensive connections, possess the ability to transmit shocks to a wide range of other companies. Price correlation has proven useful in explaining certain dimensions of firm to firm risk. However, correlation matrices may not be sufficient to fully grasp the complexities of the joint risk exposures between firms, this is especially true in cases of high volatility when market-wide correlations sharply increase [185].

Text has emerged as an important source for capturing an added dimension of common risk exposures between firms. Researchers have explored various textual sources, such as social media, news, and SEC filings, to uncover valuable insights into the inter-dependencies among firms. Sawhney et al. [205] and Souza and Aste [216] demonstrated the importance of social media in capturing real-time sentiment and market dynamics, providing a richer understanding of the relationships between companies. Similarly, Wan et al. [233] highlighted the significance of news articles in detecting and predicting market movements, as news reports often contain crucial information that influences investor behaviour. SEC filings, which provide companies with a platform to disclose important financial and operational information, have also been utilised in textual analysis. Adosoglou et al. [6] demonstrated how inter-document similarity in SEC filings can aid in building less correlated portfolios, suggesting that textual similarities can help identify companies that are less likely to be affected by similar macroeconomic shocks. Additionally, Hoberg and Phillips [119] explored the use of textual analysis in constructing diversified portfolios by leveraging information contained within SEC filings. The frequency of co-mentions of companies within textual data has been identified as a predictive factor for short-term correlation. Özgür et al. [183] showed that a higher frequency of co-mentions between companies in news articles or other textual sources tends to correspond with increased correlation in their stock prices. This finding underscores the potential of text-based data in predicting short-term market dynamics and identifying common risk exposures.

Moreover, the similarity of embedded news articles has been found to be indicative of the exposure of companies to joint "sentiment events." Wan et al. [233] demonstrated that when news articles exhibit high similarity, it suggests that companies mentioned within those articles are likely to experience similar shifts in sentiment, potentially leading to correlated market movements. This highlights how text-based similarities can offer insights into the interconnected nature of companies and their susceptibility to shared sentiment-driven events.

Traditional text embedding approaches, which rely on measures of text similarity such as cosine similarity, have limitations when it comes to capturing common sources of risk in economic contexts. These limitations can be observed from several perspectives. Firstly, if the measurement focuses on entity overlaps, the high dimensionality of the total number of words can hinder the quality of the measure. The vast amount of words in documents makes it challenging to accurately capture the relevant information and assess the similarity between entities. Secondly, when using word-based embeddings, different similarity measures can yield varying results, and none of them explicitly account for directionality or relationship. These measures only consider the mutual presence of words or phrases, without capturing the explicit dynamics of who is doing what to whom. Even advanced contextualised embeddings like BERT do not provide explicit information on the direction of relationships, making it difficult to capture complex causal chains between entities. As a result, encoding information about common vulnerabilities or relationships between entities becomes problematic. For instance, consider the scenario where both General Electric and Ford's share prices are influenced by the price of oil. Embeddings may capture a link between these entities and oil price, but they fail to capture the direction of the relationship or the number of causal steps required to connect them.

In an attempt to address these limitations, we propose utilizing a narrative framework to capture inter company common vulnerabilities. By moving beyond traditional text similarity measures, narratives provide a promising alternative for identifying and understanding common sources of risk in economic contexts. Narratives have been employed in computational natural language processing (NLP) for economics, demonstrating their effectiveness in representing inter-entity relationships. Andre et al. [18] highlight the importance of narratives in reflecting individuals' assessments of economic variables within the macro economy. This suggests that narratives can provide insights into how individuals perceive and evaluate economic factors, offering a unique lens to analyse interdependencies. Furthermore, micro narratives have been shown to effectively represent entity-to-entity relationships by aligning political and economic narratives. Ash et al. [23] demonstrate the potential of narratives in capturing the intricate connections between entities at a micro level, enabling a more nuanced understanding of their interplay.

7.3 Methods

7.3.1 Data description

In this paper we will be utilising two datasets. The first is the Reuters News Corpus, collected in Ding et al. [83]. It consists of a comprehensive collection of financial news articles. It spans the time period from October 2006 to November 2013 and contains a total of 105,323 articles. In our pilot study, we have extracted a subset of 2.5K articles specifically from the year 2006 to test out our hypothesis. We utilise this dataset as a valuable source to examine how macroeconomic entities interact amongst each other within the context of the financial market.

The second data source that we leverage is the SEC Filings dataset, a vast collection of regulatory filings submitted to the U.S. Securities and Exchange Commission (SEC). This dataset encompasses a significant time span from 1990 to 2022, with a total of 1.4 million records. For our pilot study, we focus on a select group of 311 companies from the S&P 500 index for the year 2006 corresponding to the date range of our news data. We specifically emphasize the 10K filings, which provide detailed information about a company's management and business description, as well as the risks they face. We use this dataset to understand how macroeconomic entities impact the well-being of individual companies as reported by the companies themselves.

We make the two sources of textual information interact as filings on their own can only show how similar the list of any two companies' entities is. Following the hypothesis that more information dimensions can be mined from text when considering how it fits in a given narrative and semantic context, we want to go beyond relying purely on SEC filings as they cannot contain multiple-hop information about common risk exposures. Financial news by contrast makes narrative statements about which variables impact which other variables, allowing us to parse the exposure of firms with a given individual SEC-provided business and risk profile given the news-extracted causal environment.

7.3.2 Relatio parsing of news

We will extract the narratives present in our news dataset using the Ash et al. [23] Relatio pipeline. This methodology presented in 7.1 uses NLP tools to split the text into semantic roles and then clusters those semantic roles that do not already have an explicit named entity (such as California for instance) into coherent clusters. These clusters serve the dual purpose of reducing the noise in a dataset containing many disparate but very similar entities (such as micro-processors, microprocessors, computer chips and computer processors) as well as representing distinct named entities if such a named entity is present in the text. The original Relatio paper offers several interpretation for the

trade-off between the number of named entities to introduce which keeps the abstraction of the narrative to a minimum versus the number of narratives to cluster which allows a wider amount of narrative statements to be combines [23]. The clusters once obtained are linked to each other by the verb present in the sentence. The statements extracted through these pipelines are termed 'micro-narratives' in both [23] and [141] and we will follow that naming convention here. An example output of what an extracted micro-narrative from a sentence looks like once parsed through the Relatio process in 7.2.

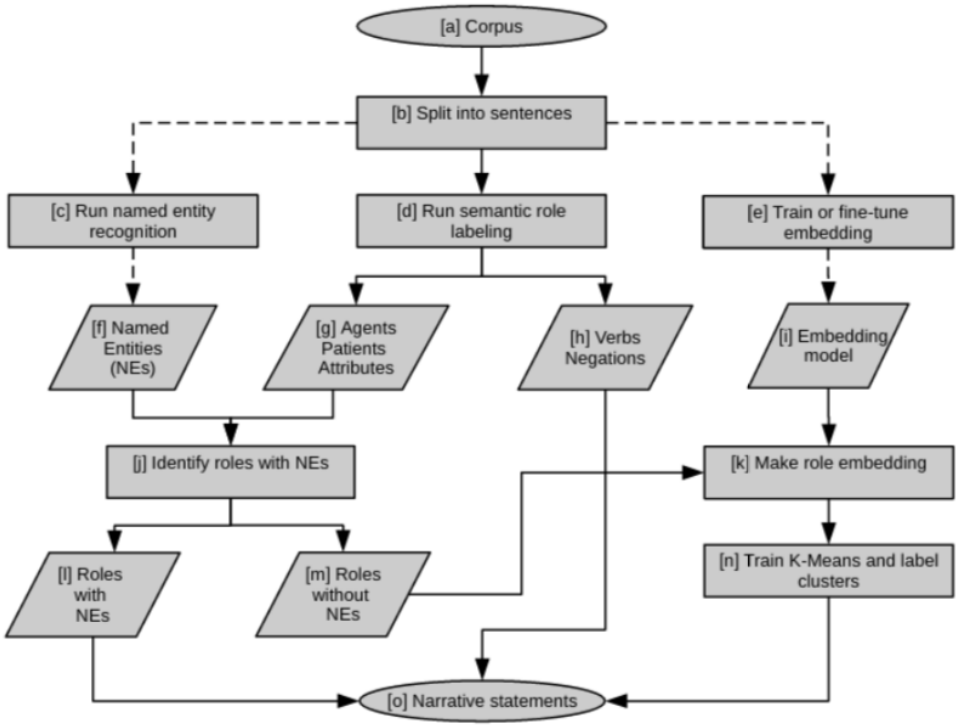


Figure 7.1: Relatio pipeline [23]

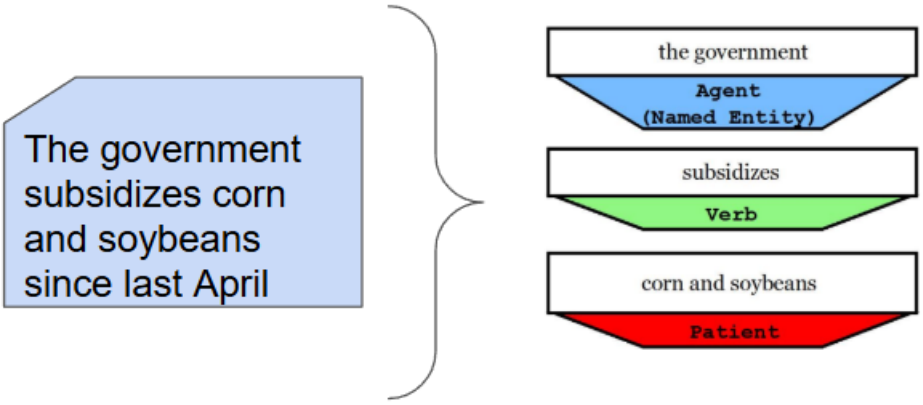


Figure 7.2: Example Relatio extraction

Of note ahead of this pipeline is an important preprocessing step present in Lange et al. [141] which led us to clean nested statements of the type 'I saw A do something to B' by removing 'I saw' and keeping only the internal part of the sentence. This preprocessing step has been shown in Lange et al. [141] to reduce noise (as the semantic role labeller has trouble parsing these statements) at very little cost in the overall interpretation power of the model.

These 'micro-narratives' can be stacked end to end as shown in 7.3 to form more complex and complete relations between entities. In the example presented here, an indirect link runs between governments who 'subsidize' microprocessors and the computers built with those microprocessors.

Stacking these statements allows us to build a latent semantic network, an example of which is presented in 7.4. As suggested in Ash et al. [23] and Shahsavari et al. [210], this graph can reflect political or economic opinions present in the discourse context of the source - in our context Reuters finance articles -. In our context, the semantic graph extracted from the news corpus allows us to map the different types of interactions that different entities such as political actors (USA, China...), economic variables (oil price, electricity price) or even direct sources of risk (fires, floods).

7.3.3 Mapping Relatio clusters to SEC Filings

In order to link what our companies' business and risk situations are with the causal risk environment they operate in, we need to first identify the entities in the SEC filings and then match them to the narratives present in the semantic graph. To extract the entities of the SEC filings

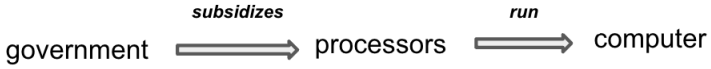


Figure 7.3: Stacking narrative statements

7.3.4 Using Node2Vec to extract common risk exposures

To construct the company-to-company distance metric given the causal structure of the narrative semantic graph, we will leverage the Node2Vec embedding methodology described in Grover and Leskovec [105]. This approach first samples nodes through the graph at random and performs random walks through the graph from those sources. These random walks output different randomly sampled traversals of the graph which can be interpreted as the 'context' of the nodes in the graph. In our case, the context of these random walks can be interpreted as randomly generated sentences which follow

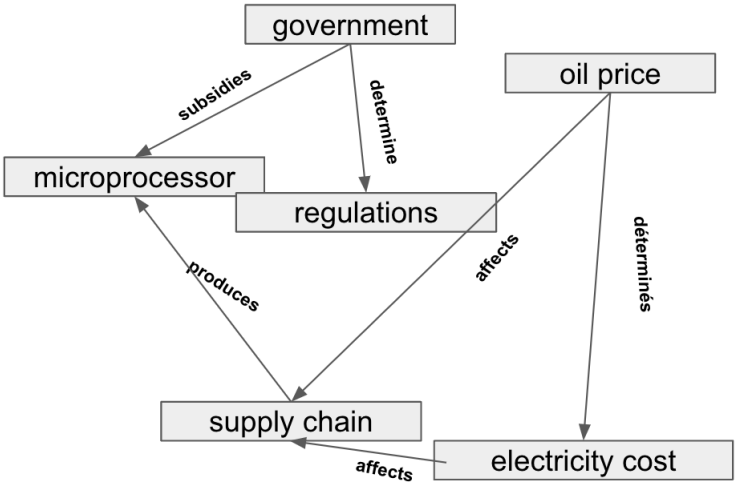


Figure 7.4: Semantic causal graph

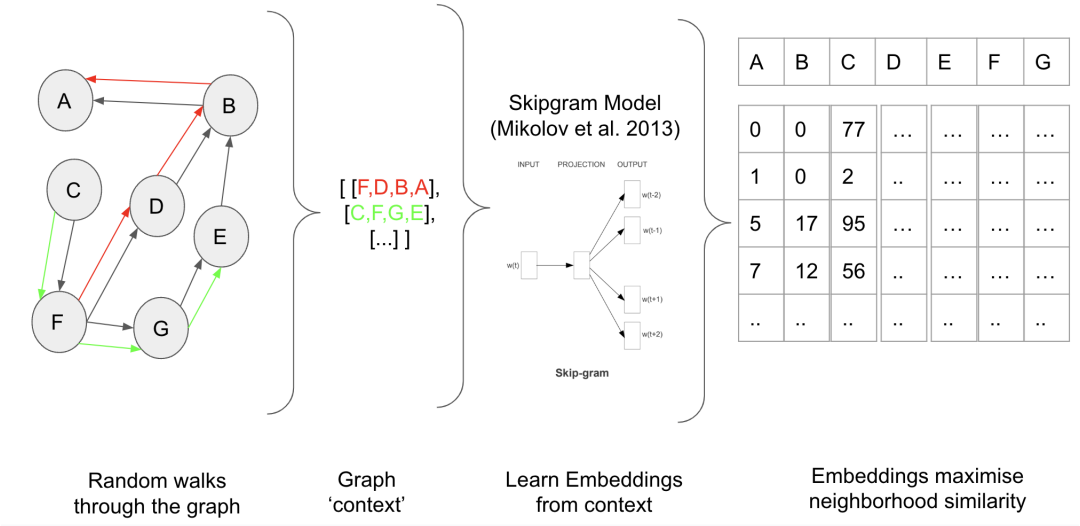


Figure 7.5: Node2Vec Pipeline Explained

the structure of our semantic graph. The individual micro-narrative triplets may exist in our news sample, but the overall stacked micro-narratives allow us to capture long-range dependencies between our entities. This context is then fed through a skip gram model which seeks to learn a representation for each node which best allows a simple feed forward neural network to predict its direct context [169]. In this way, the model will map nodes which have more similar causal neighbourhoods and are closer together in this narrative graph to more similar embeddings. We further describe how we go from the Node2Vec graph walks to embeddings in the Appendix 13.

We use the outputs of the model to generate company to company cosine similarities. These allow us to get a measure of company to company similarity that is aware of the multi-hop semantic neighbourhood. These similarities are used in our regression setup described in Section 7.3.6 as $\text{SNS}_{ij,t}^{\text{Directed}}$.

7.3.5 Building baselines of text-based common risk exposures

We also build a series of control variables to ensure that we are able to control for other dimensions of risk captured in price. Most notably, we create a TFIDF similarity of the SIC text which is a text-based embedding capturing the word level similarity of two documents. This type of measures was used in prior work Adosoglou et al. [6], to generate a mapping of company to company similarity. Alongside this, we create an entity overlap count which counts the number of similar narrative entities present within the SEC filings. This measures how similar the dimension-reduced SEC text is once passed through Relatio and serves to make sure that the eventual effect of our narrative graph based measure is not simply due to the dimension reduction of the Relatio clustering process. Through this measure, we check that it is the way these entities are arrayed within the narrative graph which helps explain common sources of risk.

We also produce an undirected version of this metric which reflects the case where our semantic graph does not care for who is doing the action in the narrative statement 'A' acts on 'B'. Through this metric we wish to control for the actual importance of directionality to our estimation of common risks. In other words, does our model account for which way the causal relation described in the news between droughts and fires as 'droughts cause fires' in its accounting of how similar two companies are or does it treat the reverse statement 'fires cause drought' as simultaneously possible. We expect that the directed version of our graph should perform better than the undirected one at explaining the correlation matrix.

7.3.6 Regression setup

The dataset consists of 311 publicly listed firms. For each year, we construct all ordered firm pairs (i, j), producing two observations per unordered pair (one with firm i as firm 1 and j as firm 2, and one with positions reversed). The dependent variable is the next-year pairwise correlation in daily equity returns. For each pair (i, j), we compute the Pearson correlation of their daily returns in year $t + 1$, giving a forward-looking design in which all predictors measured at time t serve as determinants of future return comovement.

Our key explanatory variable is the *Semantic Network Similarity (Directed)*, defined as the cosine similarity between Node2Vec embeddings of each firm's directed semantic network representation of its narrative neighbourhood. Because these networks include only edges pointing toward external entities, the measure captures asymmetric causal structure in firms' disclosures. To benchmark this measure, we also include the following controls: (i) *Same Sector*, a 4-digit SIC industry dummy; (ii) *Text Similarity*, the cosine similarity between TF–IDF bag-of-words vectors; (iii) *Entity Similarity*, the cosine similarity between entity-cluster frequency vectors; and (iv) *Semantic Network Similarity (Undirected)*,

computed analogously to the Directed measure but ignoring edge direction. We also include firm fixed effects and the lagged pairwise correlation of returns in extended specifications. All covariates are standardized to have mean zero and unit variance.

Our baseline regression (column 1 in Table 7.1) is:

$$\text{Corr}_{ij,t+1} = \beta \, \text{SNS}_{ij,t}^{\text{Directed}} + \varepsilon_{ij,t}. \tag{7.1}$$

We then add a vector of observable controls $X_{ij,t}$:

$$\text{Corr}_{ij,t+1} = \beta \, \text{SNS}_{ij,t}^{\text{Directed}} + \gamma' X_{ij,t} + \varepsilon_{ij,t}. \tag{7.2}$$

Our full specification includes firm fixed effects and lagged correlation:

$$\text{Corr}_{ij,t+1} = \beta \, \text{SNS}_{ij,t}^{\text{Directed}} + \gamma' X_{ij,t} + \rho \, \text{Corr}_{ij,t} + \alpha_i + \alpha_j + \varepsilon_{ij,t}. \tag{7.3}$$

All models are estimated via ordinary least squares (OLS). We compute two-way clustered standard errors by firm i and firm j to account for dependence in residuals across regressions involving the same firm. Coefficient significance is assessed using two-sided t-tests of $H_0 : \beta = 0$, with significance levels denoted by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). The description of the way we produced our alternative controls and variables we introduce in the regression in the Appendix 13.

7.4 Results

We present our results in the regression table below. As described in the notes underneath the regression task is regressing the next year correlation coefficient given an array of predictors including -crucially- our Node2Vec derived directed causal similarity termed 'Semantic Network Similarity (Directed)'.

Column 1 corresponds to the baseline regression we have run and shows us that a simple regression of the Semantic Network Similarity Directed against the next period correlation yields a statistically significant coefficient.

Columns 2 through 5 shows that adding our different correlated adds supplementary explanatory power. Using the same sector dummy in column 2 specifically added 1% to the R^2. In columns 3 to 4, we introduce baseline text similarity metrics used in other studies to capture dimensions of variation from text. In columns 3 we add the cosine similarity of the TFIDF embedding of the SEC filings as a predictor for correlation. In column 4 we add the entity overlap of the Relatio clusters as a predictor. In Columns 5, we added the undirected network similarity as a correlate which was statistically significant at the 0.01 threshold.

In Regressions 6 , 7 and 8, we introduce our directed network similarity metric which maintained its statistical significance even when controlling for firm fixed effects and especially in regression 8, when controlling for the lagged correlation in returns.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable: Pairwise Correlation in Daily Returns								
Semantic Network Similarity (Directed)	**0.0624*** **					**0.0515** **	**0.0276** **	**0.0275** **
	(0.0133)					**(0.0157)**	**(0.00882)**	**(0.00882)**
Same Sector		0.112***	0.108***	0.109***	0.103***	0.102***	0.103***	0.100***
		(0.0122)	(0.0121)	(0.0121)	(0.0116)	(0.0116)	(0.0104)	(0.0102)
Text Similarity			0.0376*	0.0382*	0.0376*	0.0397*	0.0703***	0.0668***
			(0.0182)	(0.0184)	(0.0184)	(0.0185)	(0.0190)	(0.0184)
Entity Similarity				-0.00757	-0.0125	-0.0109	0.0868**	0.0824**
				(0.0178)	(0.0179)	(0.0179)	(0.0290)	(0.0286)
Semantic Network Similarity (Undirected)					0.0269**	-0.00805	0.00811	0.00700
					(0.00892)	(0.00987)	(0.00545)	(0.00551)
Lagged Correlation in Returns								0.0640***
								(0.00802)
Both Firms FE							X	X
N	96410	96410	96410	96410	96410	96410	96410	96410
R-sq	0.004	0.013	0.014	0.014	0.015	0.016	0.255	0.258

Table 7.1: Regression tables with main results and controls

Note: Standard errors in parentheses, with two-way clustering by firm 1 and firm 2.
* p<0.05 ** p<0.01 *** p<0.001.

Notes: OLS regression results for main analysis. Observation is a pair of firms, firm 1 and firm 2; there are 311 firms and each pair appears twice (both permutations, with each firm in the firm 1 and firm 2 position). Outcome variable is the correlation between firm 1 and firm 2 in daily returns from 2007. Semantic Network Similarity (Directed) is the main explanatory variable, giving the cosine similarity between the Node2vec representations of the filings of firm 1 and firm 2, limiting to edges pointing to entities mentioned in the filings. Same Sector is a dummy variable equalling one if firm 1 and firm 2 share the same 4-digit SIC code. Text Similarity is the cosine similarity between the TF-IDF-weighted bag-of-words representations of the filings. Entity Similarity is the cosine similarity between the count vectors over entity clusters in the filings. Semantic Network Similarity (Undirected) is the same as the Directed measure but including all edges (not just those directed toward the firm's entities). Both Firms FE means the full set of indicator variables for firm 1 and for firm 2. Lagged Correlation in Returns means the correlation in returns between firm 1 and firm 2 from 2006. All variables are standardised to mean zero and variance one.

7.5 Discussion

We interpret these findings as a encouraging preliminary confirmation of our hypothesis that looking at a narrative graph of causal statements would help in the forecasting of the out of sample correlation coefficient. In our regressions, the closer two companies are on the semantic similarity graph, the closer their out of sample similarity is the following year. This effect is significant even when accounting for the common risk exposures in by industry, for the past common risk exposures represented in the lagged correlations. This effect remains significant at the p-value threshold of 0.01 even when introducing the other textual dimensions of similarity which represent the directly shared risks of two firms as described in their SEC filings or news environment.

The results of our preliminary analysis for this pilot study provide support for the hypothesis that the company-to-company distance extracted from the narrative links captures additional correlations in returns that are not captured by other textual similarity measures. This finding suggests that the narratives, specifically the micro-narratives, represent distinct and latent non-linear links between the firms that go beyond the surface-level similarities captured by traditional similarity measures. This is an idea that underpins other causal-graph approaches to risk detection which attempts to leverage causal Directed Acyclic Graphs (DAGs) to map out the direction causality runs in economic variables [89]. However, it must be reasserted that these results are only preliminary due to the size of our data sample.

The existence of these unique narrative links between firms can be interpreted as explained by various mechanisms. One possible explanation is that narratives serve as sense-making operations under conditions of uncertainty, allowing market participants to interpret and make sense of complex economic relationships [89]. Additionally, narratives could also serve as common organisational tools that shape the interactions and perceptions of financial entities, and therefore having a narrative-aware approach to represent this decision making process will give us the ability to capture a distinct dimension of variation of correlation [71].

Understanding how these narratives emerge, evolve, and integrate is of paramount importance as argued within the broader context of narrative economics [212]. The dynamics of these narratives provide insights into the underlying interdependencies and risk propagation mechanisms in the financial system. By examining the way in which financial entities are processed and embedded within these narratives, we can gain a deeper understanding of how common vulnerabilities and systemic risks emerge and spread.

7.6 Conclusion

Overall, our findings underscore the significance of narrative analysis in capturing and explaining inter-firm relationships and common sources of risk in the financial markets. Further research in this area is crucial for unravelling the complex dynamics of narratives and their implications for financial stability and risk management. We wish to pursue these results by expanding the date range of our analysis as well as by providing more interpretable visualisations of the results.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Common causes, Common Risk
Publication Status	x Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Dragos Gorduza, Elliott Ash, 'Common Causes, Common Risk' (2023) Accepted at the 5th Comptext Conference held in 2023 in the University of Strathclyde, Glasgow (non archival venue)

Student Confirmation

Student Name:	Dragos Gorduza		
Contribution to the Paper	Wrote manuscript, collected data, trained the narrative extraction model and extracted networks, interpreted results.		
Signature		Date	30 August 2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Xiaowen Dong, Associate Professor		
Supervisor comments I can confirm Dragos has made significant contribution to the paper above			
Signature Xiaowen Dong		Date	10/11/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 8

Discussion and Conclusion

8.1 Discussion

In this thesis we propose new ways to conceptualize financial markets as interconnected systems in line with the views developed by policymakers Haldane [108], Haldane and May [109], Bardoscia et al. [34] and researchers Acemoglu et al. [2] in response to the increasing complexity of modern economies. We leverage both network-based and text-driven approaches to extract a more nuanced view of the relationships among market actors than can be captured by traditional network-free approaches. This expanded perspective leads us to novel insights into risk detection, trading signal generation, and regulatory applications. We now first summarise the contributions of the chapters presented in this thesis and discuss the key achievements of this work in light of its broader implications for financial research and practice, we then present several limitations, and finally, we outline future research directions this thesis proposes.

Overview of contributions and applications

We first turn to the two chapters which focus on network processing. Chapter 3 introduces a volatility indicator, derived from a graph auto-encoder applied to correlation networks, that can be incorporated into standard autoregressive models. Thus, it offers regulators tasked with monitoring financial stability earlier warnings of turbulence. Chapter 4 shows that the pattern of sell-side analyst coverage, when fed to a graph attention network, improves return signal generation. This has applications for financial practitioners interested in incorporating a novel trading feature into a forecasting task.

Chapter 5 combines both network and text processing techniques to extract topic, sentiment and interaction graphs. These are used to gauge the informational value of Wall Street Bets discussions and show that the forum is mostly reactive save for small isolated clusters of users which may form the basis of a trading strategy.

We also highlight useful applications of text-only processing in financial markets. Chapter 6 presents PRET, a domain-tuned retriever trained on Basel III regulations. This model solves a difficult but crucial task for financial supervisors : understanding and retrieving financial regulatory text. This offers a blueprint for scaleable supervisory technology using NLP tools. Lastly, Chapter 7 extracts networks between firms by chaining narratives triplets from news data. This allows us to link firms by the number of shared stories involving risks they are both exposed to, rather than co-mentions or correlations. We show that using these linkages it is possible to explain out of sample correlations setting the stage for use cases for financial practitioners. interested in measuring firm co-movement.

After summarising the contributions and applications of each of our findings, we will now aim to thematically summarise the contributions to the thesis to the literature.

Novel tools for financial relationship extraction

A central contribution of this thesis is proposing tools to explicitly represent relationships between assets in markets from relational and textual data. In Chapter 4 and 5, we extract networks from relational datasets, IBES analyst ratings and WSB forum posts respectively. These expand on previous literature turning financial and economic relational databases into networks [164, 125]. Alongside these relational databases, in chapters 5 and 7, we extract relationships between entities contained in texts, therefore building an 'ad-hoc' relational database from text. We achieve this by using natural language processing algorithms to build features such as topics and capture syntactic roles in order to measure entity-similarity. These approaches complement existing text-based distances between firms such as those found in Wan et al. [233] where firms are related if they co-appear in the same window of text in the same article.

New tools for text mining applications in financial markets

Financial texts have surged in volume and diversity in recent years [178]. This evolution makes automated text processing one of the key areas for research into the evolution of financial markets with implications for investors and regulators. We propose a text-based appraisal to diagnose the informational content of an infamous online trading discussion platform in Chapter 5 : Wall Street Bets. To do so, we propose to train a novel WSB-specific FinBERT [21] model to capture sentiment as well as to use a topic model (BertTopic) [104]. Furthermore, in Chapter 6, we highlight an important mechanism in the application of NLP techniques to markets : solving for a pressing need, usually by enhancing the productivity of labour in a resource constrained setting. We show how supportive supervisory technology (SupTech) can be used to process vast quantities

of regulatory complex text better than baselines using a limited retraining procedure and discuss the applications to regulators [61, 145]. lastly, in chapter 7, we highlight how a more in-depth look at syntax through tools such as Relatio [23] can enhance the applicability of semantic-only approaches such as FinBERT embeddings in a financial task. Furthermore, by quantitatively analysing the narrative structures of news, we highlight the potential applications of narrative economics [212], thus expanding the literature.

Exploiting novel sources of financial information

We collect, clean and parse varied under leveraged data sources, thus contributing to existing financial literature by highlighting the potential information content present in novel sources of information. This is in keeping with the view highlighted in our Motivation section, where we point to going beyond price and similar quantitative metrics to represent markets. In Chapter 5 we investigate the forum Wall Street Bets which despite its' 'Meme stock' [206] episode at the start of 2021, has only received limited attention. The systematic financial benchmarking perspective we propose allows us to capture the on average poor performance of the website, which contrasts with other investigations of the information content on the website [52]. The datasets of financial rules we introduce in Chapter 6 to train the PRET model also leverages less investigated datasets : regulatory data, specifically the BASEL III dataset. While previous work has investigated the potential of fine-tuned models on financial regulation text retrieval [5], we present the first work on an open-sourced dataset, going beyond the existing Basel III datasets such as [214]. Lastly, in Chapter 4, we also present work on underutilised relational datasets in the IBES analyst network. This data has received a lot of attention from traditional price-first views that focus on how analysts vote 'in aggregate' (ie what are the moments of the distribution of analyst estimates). However, we show the important informational roles of analysts orienting investor attention to the firms they cover by using the analyst network as an additional feature in a node-level stock return forecasting task.

Developing graph machine learning methods to extract novel signals from financial networks

This thesis develops new graph machine learning approaches in financial markets. Multiple previous approaches using financial networks as input data have focused on topological information such as centrality, clustering coefficients, diameter as explanatory variables [180, 138]. These approaches are limited by an inability to represent node level and edge level information jointly. Chapters 3 and 4 fill a gap in the literature by demonstrating that graph machine learning can jointly learn node (asset level) information and structural (edge level) information for multiple types of financial networks. The approaches we

present draw from the large literature on graph machine learning [113, 184] and apply it to a financial setting through a node-level forecasting task (Chapter 4) and a graph-level unsupervised learning task (Chapter 3). Furthermore, in Chapter 4 we present a benchmarking exercise of multiple adjacency matrices and multiple graph machine learning approaches on the same node forecasting task. This also serves to enhance the existing understanding on the sources of performance in graph machine learning based financial signal extraction which has functioned as a black box from the perspective of why a strategy based on a Graph Neural Network functions well. We interpret the good performance to the representative power of the attention operator combined with an informative prior : the analyst matrix, thus going beyond initial usage of that matrix [13, 126, 246].

Comparative perspectives in text and network information

Finally, the thesis highlights the potential for convergence between relationships found in text (such as the topic graphs in Chapter 5 or the syntactic graphs in Chapter 7) and those in relational databases (like the analyst coverage networks in Chapter 4). We present evidence on the potential of explicit statements about firms made by investors or news to serve–if properly mined by the right text processing algorithm–as a good basis for an asset-to-asset distance measure with more distant assets being involved in fewer similar narratives. This expands the existing literature on benchmarking different network extraction strategies and their usefulness to market practitioners [193]. This convergence between text and relational databases underscores the potential for next-generation market analytics to leverage NLP and graph theory in order to profit from the interplay between the two. Chapters 5 and 7, for instance, suggest that textual relationships can complement or even replace explicit information about links from relational databases simply by reading the relevant texts.

8.2 Ethical considerations

Financial markets form a major part of life in modern societies. Whether one is taking out a loan, depositing money in a bank or putting money in a pension fund, financial decisions directly impact our present and future. As such a careful ethical consideration of the impacts of the tasks and methodologies propose in this work is warranted.

Data privacy

This thesis leverages various information sources to build the network representations that can help represent financial interactions. While most of the data used in this thesis such

as news and price data is anonymous, data sources such as IBES and Reddit social media posts requires careful privacy-protecting measures to protect. To ensure this, we adhered to GDPR data-protection norms. This includes implementing de-identification protocols on the name of IBES analysts and the usernames of WSB forum members. However, we should recognize the potential risks of advanced language modelling tools to de-anonymise user data by identifier posters by learning idiosyncratic personal writing styles from external data [179]. This is especially likely given the likelihood forum users may use their public identity to post on other social media on Reddit or even other platforms like Facebook, Instagram or X. Future work could address these risks by for example not releasing the direct forum posts wherever possible, or by displaying compressed information representations of posts through topics instead of the whole text.

Explainability and fairness

Developing machine learning based methodologies as in Chapters 3, 4 and 5 by fine tuning models allows us to automate a difficult time and labour-intensive task by learning signals directly from the data at the cost of a loss of explainability. Indeed, machine learning algorithms often function as black boxes: even if the learning function if useful for an outside application, it is difficult to audit *why* a given input is related to a given output. This strongly contrasts with the importance in a production setting of explainability and -especially in a regulatory setting- of transparency and auditability of a decision system. Various tools such as shapely values [167], visualising attention in Transformers [64] and more recently mechanistic interpretability in the context of LLMs [40] highlight pathways towards ensuring that the machine learning tools presented in this thesis can be applied in a fair and effective way.

Adaptations of markets to network and text monitoring

Financial market actors attempt -given some behavioural biases [221]- to maximise expected profits. With this in mind, we can reasonably expect that the techniques presented in this thesis encouraging more efficient and more diverse data collection by making use of relational and textual data in novel ways could have adverse effects by disrupting the information production incentives of market actors. Indeed, this could lead to the additional information captured by the network such as the analyst firm coverage network from Chapter 4 becoming part of the standard evaluation toolkit and therefore eliminating the very signal it was advertising. This risk of arbitrage is lessened somewhat in the case of the more macro-level monitoring presented in Chapters 3, 5, 6 and 7. However, it is not unreasonable to expect, as pointed out in Cao et al. [57], that managers, news writers, forum participants and more broadly all the individuals

whose information production labour we relied on, will adapt in an adversarial fashion as they become more aware that NLP algorithms mine their speech for information. This may have adverse implications for the transparency and quality of available data for future research as market actors may increasingly guard what they say when they know 'a machine is listening'.

8.3 Limitations and future work

Through the work presented here, we demonstrate the possibilities of combining network analysis and NLP for financial research. However, several limitations and avenues for future exploration remain and we propose to describe them in this section.

Going beyond the currently explored sources of networks

The analysis presented in this thesis primarily focused on traditional sources of financial data, such as market prices, analyst reports, and regulatory filings. This limits the range of relationships that can be explored. Future work should explore additional data sources that can provide a more nuanced view of financial dynamics. For example, communication networks derived from earnings call transcripts [116], corporate insider networks [125, 100] or speeches by influential market figures [67] represent a useful resource for the goal of building firm-to-firm representations as they give information about individual firms do and how they are related to other firms. These sources can help uncover latent sentiment and shifts in market expectations that are not immediately visible in numerical data. Extending network and text analysis methods to these domains could yield innovative insights into how information disseminates among market participants and how it influences investment decisions.

Looking at more complex networks

The 'complex adaptive systems' [108] making up modern finance are composed of very diverse stakeholders ranging from small retail investors to large market makers regulated by jurisdictions of different sizes answering to governments of varying economic and geopolitical influence. This thesis has made a first step in utilising simple networks with limited numbers of relationship types such as correlation, analyst coverage and narrative similarity matrices to represent asset to asset links. However, future research could explore more generalised settings with multiple node types representing more than assets, for instance assets, investors and news channels. They could also expand in the number of relationship types described by looking at multiplex networks as in Bardoscia et al. [34], de la Concha et al. [79], Brummitt and Kobayashi [50], Jeude et al. [125]. Hypergraphs

could also form part of the solution by creating representations of interactions that involve multiple nodes and edges together. These have seen some initial exploration in the financial realm [156, 184] but the field merits further exploration and extension. Furthermore, future research can expand the dual text and network processing angle we take here by using the network position of social media users to identify central 'influencers' and assess their ability to influence market movements based on the sentiment in their posts. This would offer a strong use case for the dual approach defended here.

Introducing LLM-powered network extraction from text

Large language models (LLMs) have emerged as one of the most popular machine learning tools for tasks such as text generation, summarisation, and information extraction in general applications as well as in financial markets [147]. These models are increasingly being deployed to extract structured knowledge from unstructured texts, such as news articles, financial reports, and social media feeds [58]. A promising direction for future work is to leverage LLMs for extracting network structures directly from text [251, 130].In Chapter 7, we showed how a text processing approach based on syntactic parsing could extract meaningful relationships between financial entities. More complex tools such as LLMs can provide more flexibility in parsing narratives that do not fit a user-defined pattern but instead leverage the generalisation capability of LLMs to adapt to different relational information. By parsing narratives and identifying relationships among entities (e.g., companies, market events, or regulatory actions), LLMs can help build dynamic graphs that capture evolving market interdependencies.

8.4 Conclusion

By systematically examining how network- and text-analysis techniques can be applied to financial markets, this thesis opens new avenues for understanding market behaviour, shaping investment strategies, and improving regulation. We show that interaction effects can be detected, encoded as networks, and mined for actionable insights. Additionally, the flexible language-processing tools we use both reveal the relationships needed for network construction and generate complementary measures —such as sentiment and topical focus— from financial documents. Section 8.1 summarises the insights these methods yield about the inner workings of modern markets. We hope the study serves as a springboard toward a richer, more transparent view of financial systems, ultimately strengthening their resilience by making the signals embedded in relationships and texts explicit.

Although limitations remain, from data quality to model interpretability, these contributions highlight the possibility of increasingly sophisticated and complete market

analytics. It is important to keep this in mind as the discussions of this works' ethical implications highlight. It is particularly important as new techniques such as LLMs promise to develop the range of datasets and questions we can incorporate into this flexible representation of the world.

Recent shocks, from Covid-19 to escalating geopolitical tensions, have thrown a spotlight on issues that are naturally captured by network representations and whose sheer informational load demands automated pipelines. Localised disruptions reverberating through supply chains have laid bare the interconnected structures underpinning financial markets, while real-time social-media coverage has multiplied the data available to interpret every new development. These shifts make the tools presented here all the more relevant for market participants and the wider public, enabling them to identify the actors occupying central network positions and to gauge more accurately the risks inherent in the economy's intricate, dynamic system.

Chapter 9

Appendix A Supplementary results and material for Chapter 3

Hyperparameter tuning

We describe the different hyperparameters we used in tuning our Graph Convolutional Network (GCN) used in Chapter 3 in Table 9.1. This shows us the results of a hyperparameter tuning run where we assess the performance of the different combinations against their performance on the held out test set as assessed by their ROC-AUC.

w_{hid}	w_{out}	Mean Dev ROC-AUC
64	8	**0.8904**
64	32	0.8835
64	18	0.8810
32	8	0.8715
32	32	0.8687
32	18	0.8643

Table 9.1: Validation ROC-AUC performance across $(w_{\text{hid}}, w_{\text{out}})$ hyperparameter configurations for $W_0 \in \mathbb{R}^{w_{\text{in}} \times w_{\text{hid}}}$ and $W_1 \in \mathbb{R}^{w_{\text{hid}} \times w_{\text{out}}}$. The input dimension $w_{\text{in}} = T$ corresponds to the feature vector length per firm. Ordered by descending ROC-AUC.

Proportion of edges and 0.7 threshold additional discussion

We display in Figure 9.1 the proportion of edges out of the total number of connections of the graph at each time point. We have a total of 401 companies, therefore the theoretical maximum connections excluding self-edges is 160400. The proportion of edges

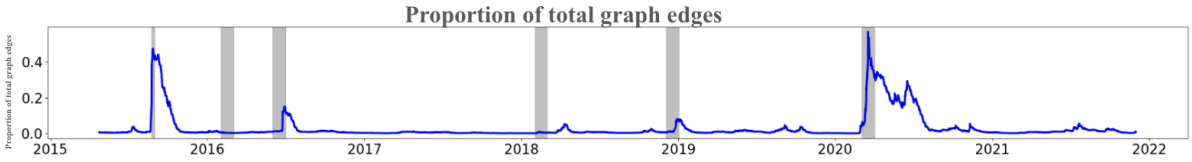


Figure 9.1: Proportion of edges in 0.7 correlation graph overtime

is calculated as the number of total edges in the network at time t out of that total $Proportion_t = |E_t|/160400$.

We can see that the proportion of edges in the network spikes with the periods of financial crisis in the grey bands which were highlighted in the Results section 3.3.1. This suggests that the proportion of edges could drive some of the results observed in that section as a higher proportion of edges .

As observed in Section 3.2.2, the 0.7 threshold is selected to only keep strongly connected pairs and discount correlations of insufficient strengths [53][138]. This approach was selected as it is straightforward to interpret the resulting network as the subset of edges above 0.7 correlation. However, it is important to note this edge threshold method suffers from significant drawbacks which we present here and use to highlight and discuss our results and their limitations. Firstly, this approach has a tendency to produce either overly connected or sparse graphs, especially in financial networks as the correlation coefficient is highly fragile to large volatility data points, making the total number of connections potentially affected by the volatility we are attempting to measure. Secondly, threshold-based filters of correlation matrices do not preserve other network properties than direct one-to-one linkages which ignores weakly correlated but topologically significant structures.

Chapter 10

Appendix B Supplementary results and material for Chapter 4

10.1 Hyperparameter tuning

This table presents the results of a hyperparameter tuning for the GAT model.

Learning Rate	Hidden Size	Weight Decay	Attention Heads	Attention Layers	Mean Dev F1
0.010	128	0.00010	8	2	**0.7340**
0.010	64	0.00010	8	2	0.7217
0.010	64	0.00001	8	2	0.7121
0.010	128	0.00001	8	2	0.7078
0.010	128	0.00100	8	2	0.6733
0.010	128	0.00001	8	1	0.6688
0.010	64	0.00001	8	1	0.6687
0.010	128	0.00010	8	1	0.6631
0.010	64	0.00001	2	2	0.6613
0.001	64	0.00001	8	2	0.6604
0.010	64	0.00010	8	1	0.6602
0.010	64	0.00010	2	2	0.6599
0.001	64	0.00010	8	2	0.6499
0.010	64	0.00100	2	2	0.6483
0.010	64	0.00100	8	2	0.6458
0.010	128	0.00001	2	2	0.6371
0.010	128	0.00010	2	2	0.6358
0.001	64	0.00010	8	1	0.6354
0.010	64	0.00100	8	1	0.6341
0.010	128	0.00100	8	1	0.6276
0.010	128	0.00100	2	2	0.6220
0.001	64	0.00001	8	1	0.6166
0.010	128	0.00001	2	1	0.6128
0.010	64	0.00010	2	1	0.6063
0.010	128	0.00010	2	1	0.6053
0.010	64	0.00001	2	1	0.6040
0.001	64	0.00001	2	2	0.5971
0.010	128	0.00100	2	1	0.5948
0.001	64	0.00010	2	2	0.5932
0.001	64	0.00010	2	1	0.5866
0.010	64	0.00100	2	1	0.5865
0.001	64	0.00001	2	1	0.5853

Table 10.1: Validation performance across GAT hyperparameter configurations.

Chapter 11

Appendix C Supplementary results and material for Chapter 5

11.1 Detailed Data Appendix

11.1.1 Data scraping

Submission and comment data from WSB is collected through the Pushshift Reddit database [37], which provides an archive of all posts and comments on Reddit. We collect data from the inception of the forum in April 2012 up to 24 June 2022. Posts on WSB typically refer to companies by their stock symbol, and we use the work of [208] to obtain a list of stock symbols (from Yahoo Finance and Compustat) and extract these from the submission text (with upper-case matching). For symbols that form common words/acronyms (e.g. *$WISH*), we only match a post with that particular ticker if the symbol is preceded by a dollar sign $. We inspected several thousand posts to ignore symbols referring to companies that are virtually never discussed on the forum, but whose symbol is used for other references. For example, the appearance of the letters SI almost always refers to short-interest, and never to the associated company: Silvergate Capital Corp. We also ignore any reference to SPY (S&P 500), since we want to use this for baseline.

For simplicity, we only keep companies traded on the NYSE or NASDAQ exchanges, since this allows us to operate in a consistent timezone. This is also a fairly innocuous step, since the vast majority of WSB activity is focused on the US and companies that trade in US markets. After all these steps are taken, we filter our data to only include posts that mention exactly one stock.

Stock price data comes from AlphaVantage. AlphaVantage helpfully provides an adjusted close price to take into account events like dividends and stock splits, which we use to calculate returns. We collect a daily time series for each stock in our data set, and

we remove posts when AlphaVantage did not have data on a particular stock. At the end of this process, we ended up with 202,681 submissions, with 6,320,968 corresponding comments. As forum activity has increased by several orders of magnitude since inception (Figure 11.1), our data is concentrated in more recent months, with the 2021 data accounting for over 60 percent of posts, despite only accounting for around five percent of the forum's lifetime.

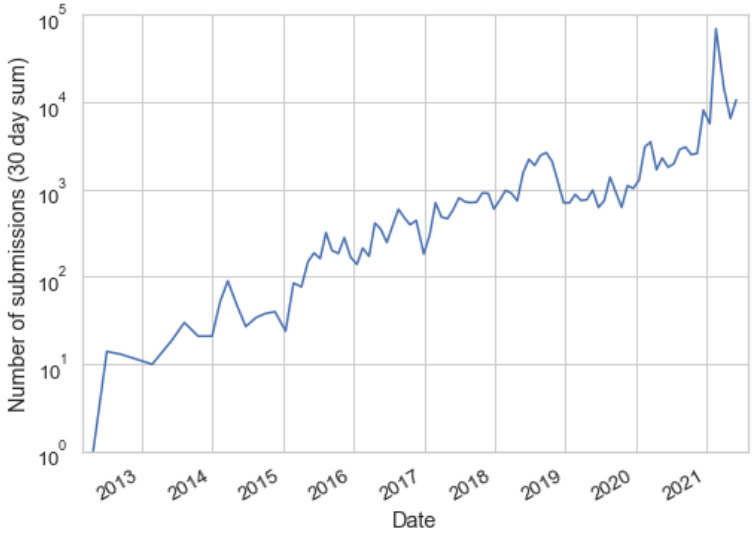


Figure 11.1: Total number of new posts on WSB every 30 days

11.1.2 Due Diligence (DD) posts

To obtain the subset of due diligence (DD) reports in the forum, we firstly filter to posts that have the DD flair. Flairs are used on subreddits to categorise posts, and posts that are flaired as DD are supposed to be moderated by WSB admins. This is the same method used by [48]. We note, however, two issues with this approach: firstly, the DD flair does not appear to be used in any significant way until July 2018, so this excludes any older posts. Secondly, we noticed upon reading the DD posts that quality control is limited in our dataset. The flair guidelines state that DD posts must be a high-effort, longer, researched posts, but many posts in the data set fail to meet this criteria. This is partially because Pushshift retains many posts that are later removed by moderators – these posts are straightforward to remove by looking up each DD submission on the Reddit API – but many posts that contain the DD flair still do not appear to meet the criteria. For example:

> *Very short DD , but potentially bigly (not investment advice) will be going in today myself on this. NOK [Nokia] literally going to the mooooooonnnnnnn*

To resolve these issues, we take two additional filtering steps beyond [48]. Firstly, we supplement the posts containing the DD flair with posts that contain the DD acronym. Secondly, we manually review every single potential DD post (i.e. flaired posts that are not removed by moderators, or contain the text 'DD'), and we remove any post that does not appear to be a valid DD. For us to consider a post a valid DD, we require:

1. The post is at least 20 words long
2. The post is primarily about one company
3. The post has a positive or negative sentiment on the company's future price
4. The post attempts to provide a reason for the future price to change

Inevitably, there is some subjectivity in this filtering, especially for the third criteria. We attempt to be quite permissive in assessing these criteria; for example, we would allow news about a new product release to be used as a reason (even though this is very likely to already be incorporated into the price). We do not, however, permit reasons that are excessively vague; for example, saying a company's price will increase because 'fundamentals are good' is insufficient without some added specificity of what is good about the fundamentals. These filtering criteria remove around 40 percent of our candidate posts, leaving us with 3,650 DD posts.

The DD posts are sufficiently small in number that we can manually label each post as bullish/positive or bearish/negative manually (neutral posts are removed). We find 77 percent of posts are bullish, and 23 percent are bearish.

11.1.3 Sentiment detection

For the forum as a whole there are far too many posts for manual classification. Previous papers have either relied on counting keywords (for example, 'buy', 'sell', 'call', 'put' etc. [52, 48]), or they have used off-the-shelf rule based sentiment classifiers [234], such as VADER [121].

The advantage of such approaches is their simplicity, but they have several problems. Choosing keywords/phrases is inherently arbitrary – it could easily miss posts that are clearly positive/negative because they don't contain the required keywords, and it is vulnerable to mistakes when authors negate phrases in the dictionary (e.g. 'calls are going to be worthless'). Algorithms like VADER, by contrast, lack the specific domain vocabulary of the WSB forum, hence their performance can be quite poor [234].

Another issue with using VADER is that we are not interested in the general sentiment of WSB posts *per se*, instead, we are interested in whether the poster has a bullish, bearish, or neutral attitude towards the future price of an asset. To clarify, suppose a trader talks

about how much money they made from a recent trade on Tesla; the post is likely to contain positive sentiment towards Tesla, but it is backwards looking, and the author may take no position on the future value of Tesla shares. We want to categorise such a post as neutral, despite it having an overall positive valence.

To resolve these issues, we firstly hand-code a random sample of 4,000 WSB posts, categorising them as bullish, bearish, or neutral depending on what the author's attitude is on future price increases. Of the posts in the sample 41 percent are bullish, 37 percent are neutral and 22 percent are bearish. [1] We then take a pre-trained BERT [80] model called FinBERT [21], which we fine-tune using our manually labelled sample. FinBERT has been trained to perform sentiment detection on financial news, which is helpful as it can pick up on financial jargon used in the forum.

The sample was split, with 75 percent of data used for training, 10 percent for validation and 15 percent for testing. The maximum text length was 256 words (shorter posts were padded, longer posts were truncated), and we ran batch sizes of 8[2]. Validation accuracy was assessed every 100 batches, and training was stopped when cross-validation accuracy failed to increase after 10 evaluations.

Table 11.1: Confusion matrix showing the results of the fine-tuned FinBERT classifier on the test data ($n = 604$)

		Prediction		
		Bullish	Bearish	Neutral
Label	Bullish	173 (28.6%)	30 (5.0%)	49 (8.1%)
	Bearish	14 (2.3%)	95 (15.7%)	21 (3.5%)
	Neutral	42 (7.0%)	30 (5.0%)	150 (24.8%)

Our final model had a test accuracy of 69.2 percent, which we consider sufficient due to the challenging task of predicting future-facing sentiment. When a sample of the manually annotated data was checked by a second human, we found labelling disagreement in around 10 percent of cases. Around half the time this was due to human error, but in around 5 percent of cases, it was genuinely ambiguous. Furthermore, the performance is substantially better than a majority class baseline (40.7 percent). The confusion matrix is shown in Table 11.1.

[1] Since bearish posts tend to be a minority on the forum, bearish posts were up-sampled to reduce class imbalance.

[2] Fine tuning was performed on a Nvidia RTX 3070 GPU, which ran into memory issues when larger batch sizes were attempted

11.2 Granger causality test details

Granger Causality Setup

We test whether our normalised sentiment measure $\Delta\hat{S}_{i,t}$ is Granger causal for daily log-returns. In contrast to a joint lag test, we test each lag *individually*. That is, for lag $\ell \in \{1, 2, 5, 10\}$ we run a separate regression and test only the ℓ-day sentiment lag, rather than all lags up to ℓ.

For each stock i and each lag ℓ, we estimate the following predictive regression:

$$r_{i,t} = \alpha_i + \sum_{j=1}^{\ell} \beta_{i,j}\, r_{i,t-j} + \gamma_{i,\ell}\, \Delta\hat{S}_{i,t-\ell} + u_{i,t}, \tag{11.1}$$

where $r_{i,t}$ is the daily log-return and $\Delta\hat{S}_{i,t-\ell}$ is the sentiment measure lagged by ℓ days. We include ℓ own-lags of returns to control for return dynamics, but only a *single* sentiment lag in each specification.

Hypothesis.

For each lag ℓ, we test whether sentiment at that lag helps forecast future returns:

$$H_{0,\ell} :\ \gamma_{i,\ell} = 0 \qquad \text{vs.} \qquad H_{1,\ell} :\ \gamma_{i,\ell} \neq 0. \tag{11.2}$$

We compute a Wald statistic for $\gamma_{i,\ell}$; because this test imposes only one restriction, the Wald statistic follows

$$W_\ell \sim \chi^2(1) \quad \text{under } H_{0,\ell}. \tag{11.3}$$

Decision rule.

We reject $H_{0,\ell}$ if the Wald statistic exceeds the corresponding $\chi^2(1)$ critical value. The thresholds for our significance levels are:

$$\chi^2_{1,0.95} = 3.841, \qquad \chi^2_{1,0.99} = 6.635, \qquad \chi^2_{1,0.999} = 10.828.$$

Thus, if $W_\ell > 3.841$ (resp. 6.635, 10.828), we conclude that $\Delta\hat{S}_{i,t-\ell}$ is Granger causal for $r_{i,t}$ at the 5% (resp. 1%, 0.1%) level.

11.3 Topic Model

We present a more detailed description of the tickers present in different topics from our topic model in Table 11.2. For each cluster, we report the top most mentioned tickers, by number of posts associated with it. We assign up to 10 tickers per cluster for readability, but we also report the number of posts in total accross all tickers as well as the total number of tickers in that cluster.

11.4 CAR for Popular Stocks

Figure 11.2 shows a breakdown of the CAR plots for the twenty most popular stocks on the forum. [3] We see similar patterns to the results within the main paper for smaller 'memey' stocks like GME, AMC, BB, SPCE, and NOK. This pattern can also be observed for TSLA and AMD. Other large companies like AAPL, MSFT, AMZN, NVDA, GE and BABA are fairly flat in their pattern, suggesting a weaker relationship between sentiment and returns. The only two cases where sentiment does seem to meaningfully precede peaks in CAR are SNDL and SNAP. The performance of SNDL (a cannabis company) appears by a spike in popularity in late January 2021, which preceded a tripling in the share price over the next few weeks. The performance of SNAP also appears to be driven by outliers, as its median CAR chart is quite flat compared to its mean. Overall, the CAR analysis suggests that besides a few outliers, WSB sentiment appears to be a lagging indicator of stock price changes, not a leading one. Our interactive online dashboard allows one to individually inspect stock CAR plots, as well as returns, leading up to and following posts on WSB.

11.5 Discussion of Clusters and Returns

Author Clusters

We observe eleven investor clusters, where assets are connected to other assets based on whether a similar subset of authors interacts in discussions about both assets simultaneously. In this section, we investigate the profitability of investors on WSB within these different clusters, as well as the correlation of asset returns within these clusters. We remind the reader of the most central assets within each of the discussion clusters: 0) DAL, UAL, 1) PLTR, 2) AMD, AAPL, MU, 3) SLV, GLD, 4) TSLA, SPY, 5) GME, AMC, 6) CCL, RCL, 7) LMT, BA, 8) JNUG, NUGT, 9) TLRY, CGC, 10) NKLA.

[3] Palantir Technologies (PLTR) and Rocket Companies (RKT) are more popular stocks on WSB than General Electric (GE) or Disney (DIS), however, both stocks had their IPOs in 2020, so there is not sufficient data to fit the CAPM model for these stocks.

Table 11.2: **Sizes and contents of topic clusters**

Cluster	Tickers	No. of Posts	No. of Companies in Cluster
0	AMZN, SNDL, HMNY, MVIS, WISH, CLNE, CCL, DKNG, SUNE, CTRM	12,373	41
1	AAPL	2,530	1
2	MSTX, AVXL, MNKD, OPK, AMDA, AUPH, SENS, AMRN, ADMP, CARA	3,312	44
3	BB, BBBY, BNGO, BRK, BBY, DB, BLNK, ABNB, BBW, BTT	9,898	12
4	MU, TLRY, ACB, CGC, APHA, CRON, AMAT, MO, OGI, HEXO	6,557	17
5	NFLX, SQ, ATVI, PLUG, WMT, SBUX, LULU, CRM, CMG, MCD	10,045	72
6	FB	1,580	1
8	SNAP, NVDA, NAKD, NAK, FSLY, TTD, DBX, KR, TWLO, GNC	5,763	35
9	AMC	19686	1
11	NOK, RKT, WKHS, SOS, ZOM, QS, SRNE, TRCH, ASO, LFIN	11,349	33
13	MSFT, MS	1,849	2
14	NIO, XPEV	2,160	2
15	INO, OCGN, PFE, NVAX, MRNA, CVS, JNJ, AZN, VXRT, TEVA	2,474	13
16	BABA	1,175	1
18	USO, UVXY, CHK, WTI, XOM, SVXY, NAT, RIG, UCO, BP	1,001	12
19	SPY, TSLA, AMD, PLTR, SPCE, NKLA, QQQ, PRPL, GM, ROKU	137 28,438	30

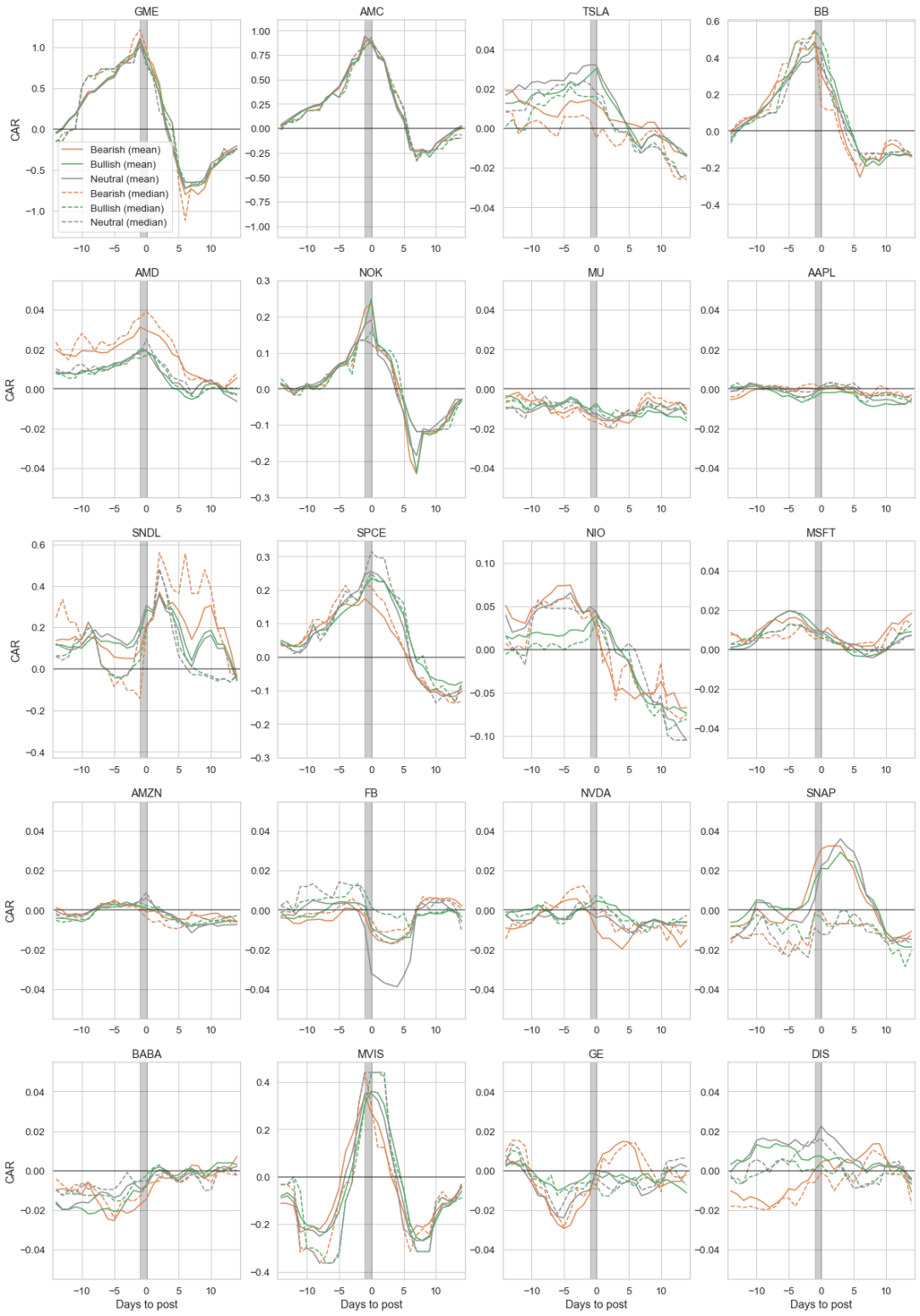


Figure 11.2: Average cumulative abnormal return 14 trading days before and after post submission, grouped by post sentiment for each of the 20 most popular stocks on WSB.

Table 11.3 presents our key result. We observe that portfolios built using opinions from

Table 11.3: **Distribution of Log-Returns for Investor Clusters**; we present some summary statistics for the next day log-returns for a portfolio invested according to the sentiment of submissions in different asset clusters on WSB as presented in Figure 5.3b. The table presents several summary statistics, as well as the p-value for the mean of the distribution to be equal to zero. Cluster **8** has insufficient observations and, therefore, is not considered.

	μ	σ	kurtosis	p-value	# posts	*Within Cluster Asset Correlation*
Cluster						
0	-0.0211	0.09	1.99	0.00	593	0.80
1	-0.0101	0.09	1.58	0.00	5,567	0.36
2	-0.0016	0.04	18.23	0.00	14,089	0.22
4	-0.0012	0.07	44.03	0.04	13,006	0.23
5	-0.0791	0.44	1.56	0.00	61,787	0.17
6	0.0062	0.10	5.99	0.21	398	0.84
7	0.0077	0.09	2.32	0.00	2,324	0.35
9	-0.0241	0.21	3.81	0.00	867	0.60
10	-0.0095	0.12	1.34	0.11	430	0.10

WSB fairly consistently lose money. The tendency is particularly stark in the "meme" stock cluster – cluster **5**. This perhaps highlights the tendency for less sophisticated investors to participate in conversations surrounding these assets. A distinct, more profitable cluster appears to be cluster **7**. Unlike several of the other asset clusters, the assets appear to be more loosely connected, even though a potential unifying theme of large-scale engineering / space exploration is noticeable, since the cluster contains BA (Boeing), LMT (Lockheed Martin) and SPCE (Virgin Galactic Holdings). This analysis, in combination with our portfolio-building exercise with DD posts, highlights that overall investors on WSB appear to be trend-followers and fail to predict the market, however, small pockets of investors may have meaningful insights and outperform the market.

Topic Clusters

Asset clusters identified by our topic graph isolate either individual tickers with highly distinct topics in their discussions, or broader groups of clusters with less distinct and more widely shared discursive environments; ticker topic clusters are displayed in Table 11.2. Table 11.4 assesses the statistical properties and behaviour of these asset clusters.

We observe, that most of the identified clusters with statistically significant p-values at the 1% threshold lose money. This suggests that on the whole, as reported above, strategies identified by mining the topic graph lose money. However, similarly to Table 11.3, certain clusters topic clusters exhibit positive daily returns, on average: cluster 1

Table 11.4: **Distribution of Log-Returns for Topic Clusters**; we present some summary statistics for the next day log-returns for a portfolio invested according to the sentiment of submissions in different asset clusters on WSB. The Table 11.4 below presents several summary statistics, as well as the p-value of the t-test for the mean of the distribution to be statistically significantly different to zero. The H_0 hypothesis is that the mean return is 0 and the alternative hypothesis is that the mean is different from 0. The clusters of GNUS (cluster 7), GTE (cluster 10), MGNI (cluster 12) and RC (cluster 17) are not considered as they do not meet a minimal threshold of 1000 posts.

	μ	σ	kurtosis	p-value	# posts	# *tickers in cluster*
cluster						
0	-0.0135	0.14	10.60	0.00	12,373	41
1	0.0027	0.03	4.33	0.00	2,530	1
2	-0.0211	0.13	9.57	0.00	3,312	44
3	-0.0301	0.18	2.71	0.00	9,898	12
4	-0.0168	0.13	21.04	0.00	6,557	17
5	-0.0005	0.06	30.90	0.41	10,045	72
6	-0.0018	0.03	17.60	0.03	1,580	1
8	-0.0029	0.06	58.04	0.00	5,763	35
9	-0.0639	0.37	0.57	0.00	19,686	1
11	-0.0742	0.16	3.63	0.00	11,349	33
13	0.0023	0.02	4.06	0.00	1,849	2
14	0.0130	0.09	1.60	0.00	2,160	2
15	-0.0095	0.13	10.90	0.00	2,474	13
16	0.0026	0.03	2.69	0.00	1,175	1
18	-0.0016	0.06	3.13	0.40	1,001	12
19	-0.0009	0.07	3.41	0.03	28,438	30

(AAPL), 13 containing Microsoft and Morgan-Stanley, 14 containing NIO and XPEV two electric carmakers, and cluster 16 containing Alibaba. We notice that the smaller clusters with fewer assets perform better than much larger clusters, as they are the only ones displaying positive returns on average. On one hand, this may be driven by the systematic over-performance of tech stocks such as AMC, APPL and MSFT over the period. However, not all clusters with positive returns identified by the topic graph are composed of tech stocks, and conversely, not all the small clusters have positive returns. This allows us to suggest that tighter, more focused topic environments may indicate the presence of specialized 'pockets' of investors with better understanding and insights into the market.

11.6 Short and long term trade signals

11.6.1 Due diligence reports

Signal Extraction

Since DD reports only form a small fraction of total WSB activity, the methods used for investigating the relationship between average sentiment and future returns are not appropriate, as these relied on averaging over thousands of data points per stock, and constructing time series for the Granger causality tests. Instead, it is better to consider whether the event of each DD post precedes a change in return. We largely follow the methods of [48] in our analysis of the DD reports.

Our main variable of interest is the sentiment of posts on day t for company i, which we will denote $d_{i,t}$. This is very similar to our definition of $S_{i,t}$ from the previous section, however in this case we separate bullish posts from bearish posts and fit models on each of them instead of computing a net score. The justification for this choice is that we wanted to consider the effect of variables beyond sentiment, and this is much more straightforward when the posts are treated separately.

Beyond sentiment, we defined a series of additional variables that might be a relevant indicator of the quality of a post:

1. The number of comments a post receives. We consider the submission to be a comment so each submission has at least one comment.

2. The number of words the post contains.

3. The maximum depth of a submission comment. We consider the submission itself to have a depth of zero, comments replying to the submission have a depth of one, comments replying to comments replying to the submission have a depth of two and

so on. This value is divided by the log number of comments in the post, else it is highly correlated with our first variable.

4. A binary indicator whether a post contains a URL. URLs could be to another Reddit post or an external source, but we did not include urls to images.

5. A binary indicator whether a post was 'proactive'. We consider a post to be proactive if the post occurred when the CAR of the stock was not statistically significantly different from zero (using a t-test). Contrast this to 'reactive' posts, which are talking about companies whose price is substantially different from what we would expect under the CAPM.

All variables were aggregated by summing across each unique day/company combination, and the first two variables were also log transformed. Since two of our variables depend on actions taken after the post is submitted, we only include comments that occur within 24 hours of the original post.

The choice of each of these features was motivated by various considerations. Number of comments and maximum depth can both be considered as graph features when the Reddit submission and successive comments are represented as a directed acyclic graph, also known as an information cascade. Previous research on Twitter has found that graph features can be predictive of the content of Tweets, which has been a particularly active area of research in fake news detection [232], [199],[171]. If Reddit users react differently to posts with more predictive power (perhaps because they are better analysed), then such graph features could be useful for our models.

The other three features serve as proxies for the quality of a post. Longer posts may have more detailed robust analysis, and therefore may be more predictive of price changes. Posts that contain URLs are providing sources, which again could be an indicator of a more reliable signal. The proactive feature is inspired by one of the previous WSB papers [52], which suggested that when a submission is created when the price of an asset has not changed recently, then it might be more likely to contain a useful signal compared to posts about assets whose price is already very far from its typical value (and therefore likely to experience mean-reversion).

For each day and company in our dataset, we calculate a net DD sentiment score. We then fit the following ordinary least squares (OLS) regression with fixed effects :

$$\begin{aligned} r_{i,[t,t+m]} = \beta_1 d_{i,t} + \beta_2 x_{i,t,k} + \beta_3 r_{i,[t,t-1]} \\ + \beta_4 CAR_{i,t} + \gamma_t + \epsilon_{i,t}. \end{aligned} \tag{11.4}$$

$r_{i,[t,t+m]}$ is the (log) return between day t and day $t+m$, $d_{i,t}$ represents our sentiment variable, and $x_{i,t,k}$ is one of our additional variables, where k denotes which variable is being used. $r_{i,[t-1,t]}$ is the return of the stock when comparing the previous day price to

day t, and γ_t is our fixed time effects respectively. The coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ and the time fixed effects γ_t are estimated using ordinary least squares (OLS). The goal of OLS is to find the regression parameters that make the model's predictions as close as possible to the observed returns by choosing the parameters that minimise the squared error between the predicted value of the dependent variable and its realised value. Time fixed effects ensure that we control for overall movements of the markets, but including fixed effects for firms is inappropriate for two reasons. Firstly, it would mean that our model was including future information about a company's performance, and secondly, if the forum is good at picking stocks that generally outperform the market (even if they do not get the timing especially precise), then we want to be able to observe that signal in our model. To estimate our fixed effects and control variables accurately, we include closing price data for all stocks mentioned in the forum at least once, and include prices from 2017 July 17 up to 21 June 2021, which is the date range of our DD sample. This gives each model around 1.5 million data points to estimate our time effects and control variables.

Note that the comparison date t is defined as being the next open or close price of the asset 24 hours after the post was published. This is done to block reverse causality, since it is possible that number of comments or maximum comment depth increase after price changes, which would show as a misleading association in our model.

Having previous day return and CAR included in our model has a couple of benefits. Firstly, depending on the size of m, the control variables can help to control for price changes resulting from momentum or mean reversion [194]. Second, as we saw in our previous section, forum activity tends to be highest when CAR is very high, so including these controls ensures that any effect measured is not simply a proxy for prices already being high. To ensure our control variables are estimated precisely, our model includes all companies and trading days from the creation date of the earliest DD post, up to the creation date of the last DD post.

We fit 3,744 models, the number resulting from the multiplication of the following configurations:

- Bullish and bearish posts are considered in separate models – this makes aggregation of the remaining independent variables considerably more straightforward (2 configurations).
- Values of m are considered from one to 26 weeks in one week intervals (26 configurations).
- Every model has sentiment as an independent variable, and zero or one of our other independent variables (6 configurations).
- Whether the model considers all posts, or just flaired posts (2 configurations).

- Whether the model considers all stocks, or excludes meme stocks: GME, AMC, BB, and NOK (2 configurations).

- Whether the model considers all posts, posts before 2021, and posts from 2021 (3 configurations).

Having multiple configurations allows us to discern whether any effects we observe are robust to alternative configurations. We therefore not only consider whether a parameter is statistically significant in a single model, but whether it is statistically significant under other considerations.

Due diligence and market returns

The results of the fixed effects models for the DD posts are shown in the Supplementary Materials, Tables 11.6, 11.7, 11.8, and 11.9. Since we fitted over three thousand models, it is not practical to show all results, so instead presentation is limited to just the models looking at price changes over four weeks and twelve weeks. The results for the models that only include the flaired DD posts are not included, since it was of little practical importance.

Note that the sign for the sentiment variable is flipped for the bearish models, so it is always the case that a positive coefficient for the sentiment variable implies that prices are increasing in the same direction the post creator predicted, while negative coefficients imply that prices moved in the opposite direction to the prediction. The signs for other variables is not reversed, so positive coefficients imply increasing asset prices in all model specifications, and vice-versa for negative coefficients.

Our results indicate that the number of comments and the maximum comment depth are not associated with future price changes. Across all our configurations, number of comments and maximum comment depth were respectively statistically significant in 0.8 and 2.2 percent of the models they were included. The coefficient for sentiment was statistically significant in 19 percent of our models. There are distinct patterns to the model configurations where this occurs. Firstly, the parameter is rarely statistically insignificant in the bearish models. The only time this occurs in the displayed results is in Table 11.7, for the post-2021 model with meme stocks included and no other independent variables (effect size -0.93, $p < 0.05$). Secondly, when sentiment is statistically significant, the coefficient is negative in about three models for every one model where the coefficient is positive. In the bullish four week models shown in Table 11.6, the sentiment parameter is never positive and statistically significant. Thirdly, in the 543 configurations where sentiment is statistically significant and the coefficient is negative, 63 percent of these models include only the posts published in 2021. This matches a finding from [48], that asset prices tend to move in the opposite direction of DD predictions made in 2021. Figure 11.3 shows how the sentiment parameter varies when altering the time window of our

model. Interestingly, as Figure 11.3 demonstrates, this finding only holds in the models where meme stocks are excluded. If meme stocks are included, returns are negative in the short-run but positive (albeit with high uncertainty) at time horizons longer than 6 weeks. Finally, in the 174 configurations where sentiment is statistically significant and the coefficient is positive, 172 of those configurations contain an additional independent variable. Examples of this phenomenon are apparent in Table 11.8 in some of the models where post word count is included as an additional feature.

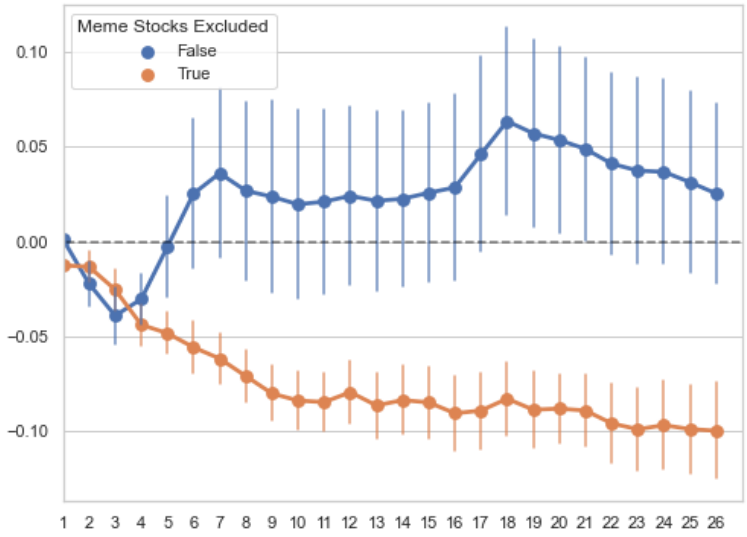


Figure 11.3: The coefficient for the sentiment variable for post-2021 bullish models with no additional features at different time horizons (1-26 weeks). Split by whether meme stocks (GME, AMC, BB, and NOK) are included. Bars indicate 95 percent confidence intervals.

This leads naturally to a discussion of our remaining independent variables. Post word count is statistically significant in around 21 percent of models where it was included, but when it is statistically significant, the coefficient is negative 97 percent of the time. For example, in Table 11.8, in the model including all bullish posts, the word count variable has an effect size of -0.0135 ($p < 0.001$). Notice, however, that in the same model, the sentiment parameter is statistically significant with an effect size of 0.0672 ($p < 0.001$), roughly five times the magnitude and the opposite sign. Since the number of comments variable was logged, it turns out that the average post has a logged word count of approximately five, meaning for the typical post these effects washout, and the apparent effects are primarily a result of multicollinearity.

The URL feature is statistically significant in 18 percent of models, and always positive when significant. This feature tends to only be significant in the bearish models, for example, in the pre-2021 bearish model including meme stocks, the URL coefficient is 0.0597 ($p < 0.05$). This effect cannot be explained by multicollinearity, as the sentiment

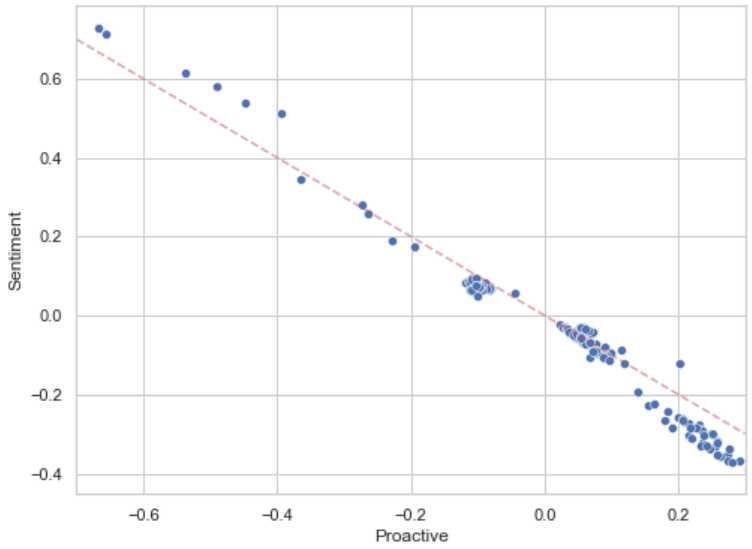


Figure 11.4: Scatterplot showing the parameter values of proactive and sentiment in all models where proactive is statistically significant. The red dashed line shows $y = -x$

parameter for this model is negative with effect size of -0.0242 (not statistically significant), but since this is a bearish model, that implies that bearish posts with URLs are more likely to precede an asset price increase. The URL parameter is also more likely to show up as positive for models looking at time horizons greater than 8 weeks, and especially at time horizons greater than 15 weeks.

Finally, the proactive variable is statistically significant in 27 percent of models. In bullish models where it is a statistically significant feature, it is always positive, and conversely in bearish models where the proactive variable is statistically significant, it is always negative. While this is promising, the actual effect size of the proactive variable is typically very close in magnitude to the sentiment variable with the opposite sign. For example, in the four-week model with all bullish posts, sentiment has a effect size of -0.0404 ($p < 0.001$), while proactive has an effect size of 0.0443 ($p < 0.001$).Since 80 percent of posts are labelled as proactive under our definition [4], these effects are largely cancelling out, and once again indicate multicollinearity.

11.6.2 Longer-term trade signals

In addition to short-term signals, which look at the sentiments on the forum on a given day, we propose that longer-term trends within the forum may be predictive of market returns. Specifically, we look at the relationship between several signals on the forum on a weekly and monthly timescale, and their relation to returns, volatility and volumes

[4]While this may seem like a surprisingly high proportion, recall our definition of a 'proactive' post was one with a CAR that was not statistically significantly different from zero, which is going to be true 95 percent of the time when selecting stocks and days at random.

(also at weekly and monthly time periods). The signals we develop for this purpose are : Momentum $M_{i,t}$, Normalized sentiment $S^N_{i,t}$ and Normalized sentiment change $\Delta S^N_{i,t}$.

Momentum: We propose a metric for momentum on the forum:

$$M_{i,t} = \log\left(\frac{n_{i,t}/n_t}{n_{i,t-1}/n_{t-1}}\right) \tag{11.5}$$

which captures the log of the fraction of posts discussing asset i in week t versus $t-1$.

Normalized sentiment

$$S^N_{i,t} = \frac{S_{i,t}}{n_{i,t}} \tag{11.6}$$

which captures the extent to which all the posts on the forum about asset i at time t, $n_{i,t}$, agree on the sentiment.

Normalized sentiment change

$$\Delta S^N_{i,t} = S^N_{i,t} - S^N_{i,t-1} \tag{11.7}$$

which captures the extent of agreement on the forum around a specific sentiment.

The 6 ordinary least squares regressions presented in Table 11.5 below are the following.

$$r^{(W)}_{i,t+1} = \alpha_1 + \beta_{1M}M_{i,t} + \beta_{1\Delta S}\Delta S^N_{i,t} + \beta_{1S}S^N_{i,t} + \gamma_{1r}r_{i,t} + \gamma_{1\sigma}\sigma_{i,t} + \varepsilon^{(1)}_{i,t+1} \tag{11.8}$$
$$\sigma^{(W)}_{i,t+1} = \alpha_2 + \beta_{2M}M_{i,t} + \beta_{2\Delta S}\Delta S^N_{i,t} + \beta_{2S}S^N_{i,t} + \gamma_{2r}r_{i,t} + \gamma_{2\sigma}\sigma_{i,t} + \varepsilon^{(2)}_{i,t+1} \tag{11.9}$$
$$v^{(W)}_{i,t+1} = \alpha_3 + \beta_{3M}M_{i,t} + \beta_{3\Delta S}\Delta S^N_{i,t} + \beta_{3S}S^N_{i,t} + \gamma_{3r}r_{i,t} + \gamma_{3\sigma}\sigma_{i,t} + \varepsilon^{(3)}_{i,t+1} \tag{11.10}$$

$$r^{(M)}_{i,t+1} = \alpha_4 + \beta_{4M}M_{i,t} + \beta_{4\Delta S}\Delta S^N_{i,t} + \beta_{4S}S^N_{i,t} + \gamma_{4r}r_{i,t} + \gamma_{4\sigma}\sigma_{i,t} + \varepsilon^{(4)}_{i,t+1} \tag{11.11}$$
$$\sigma^{(M)}_{i,t+1} = \alpha_5 + \beta_{5M}M_{i,t} + \beta_{5\Delta S}\Delta S^N_{i,t} + \beta_{5S}S^N_{i,t} + \gamma_{5r}r_{i,t} + \gamma_{5\sigma}\sigma_{i,t} + \varepsilon^{(5)}_{i,t+1} \tag{11.12}$$
$$v^{(M)}_{i,t+1} = \alpha_6 + \beta_{6M}M_{i,t} + \beta_{6\Delta S}\Delta S^N_{i,t} + \beta_{6S}S^N_{i,t} + \gamma_{6r}r_{i,t} + \gamma_{6\sigma}\sigma_{i,t} + \varepsilon^{(6)}_{i,t+1} \tag{11.13}$$

These regressions test whether longer-term discussion dynamics on WallStreetBets predict future market outcomes at weekly and monthly horizons. For each frequency, we estimate three specifications, using as dependent variables next-period returns, volatility, and trading volume. The key variables of interest are forum momentum ($M_{i,t}$), the level of normalized sentiment ($S^N_{i,t}$), and changes in normalized sentiment ($\Delta S^N_{i,t}$). The null hypotheses associated with these coefficients are $H_0 : \beta_{\cdot M} = 0$, $H_0 : \beta_{\cdot \Delta S} = 0$, and $H_0 : \beta_{\cdot S} = 0$, implying that WSB signals contain no predictive information for the respective market outcomes. Lagged returns and lagged volatility are included as controls.

Each parameter estimated using OLS is tested using a two-sided t-test where the null hypothesis is that the parameter is 0. Rejecting any of these null hypotheses indicates that the corresponding variable has statistically significant predictive power at that horizon.

Table 11.5: Stock returns versus granular social shocks

	Weekly Dependent Variables			Monthly Dependent Variables		
	$r_{i,t+1}$	$\sigma_{i,t+1}$	$v_{i,t+1}$	$r_{i,t+1}$	$\sigma_{i,t+1}$	$v_{i,t+1}$
	(1)	(2)	(3)	(4)	(5)	(6)
WSB Variables						
$M_{i,t}$	-0.002*	-0.001***	-0.020***	0.014	-0.001***	-0.026***
	(0.001)	(0.000)	(0.005)	(0.017)	(0.000)	(0.005)
$\Delta S^N_{i,t}$	-0.000	-0.001	0.007	0.030	-0.001**	0.022**
	(0.002)	(0.001)	(0.008)	(0.036)	(0.000)	(0.009)
$S^N_{i,t}$	0.004	0.001*	0.015	-0.103**	0.002***	-0.018
	(0.003)	(0.001)	0.010	0.047	(0.001)	(0.012)
Controls						
$r_{i,t}$	-0.042***	0.022***	-0.030	-0.026***	-0.006***	0.045**
	(0.008)	(0.002)	(0.032)	(0.009)	(0.001)	(0.023)
$\sigma_{i,t}$	-0.138***	0.387***	-3.355***	-2.002	0.435***	-4.693***
	(0.022)	(0.006)	(0.090)	(0.548)	(0.007)	(0.143)
Observations	13,119	13,119	13,119	9,528	9,528	9.528
R^2_{adj}	0.007	0.234	0.105	0.003	0.317	0.114

Notes: this table presents the relationship between longer-term signals from WSB and the markets. Columns 1-3 examine the relationship on a weekly timescale, while columns 4-6 examine the relationship on a monthly timescale. The test is a 2 sided t-test with the null hypothesis H_0 for each variable is that it is not different from 0 at the respective significance level.
*** Significant at 1% level ** Significant at 5% level * Significant at 10% level

Results

Table 11.5 presents the results for the relationship between various long-term signals and market variables. We select stocks that have a non-zero number of posts both within the time period t and $t+1$ for our sample. We observe that, consistently with our previous analysis - returns are particularly hard to predict using signals from WSB, but also that stocks within our sample have a tendency to revert in returns, as per the negative coefficient on $r_{i,t}$. The change in momentum has a consistent and negative effect, potentially capturing the reversion in hype, which is known to follow patterns similar to the hump-shaped patterns in infectious disease.

11.6.3 Due dilligence tables

We present the comprehensive results of our exploration of due diligence posts and their link to returns.

Time Filter	Meme Stocks Excluded	Other Variable Name	Sentiment	Other Variable	CAR	$r_{[t,t-1]}$
All posts	False		-0.0102		-0.0356	-0.0048
			(0.0072)		(0.011)**	(0.0123)
		Num comments	-0.0345	0.0113	-0.0356	-0.0048
			(0.0134)**	(0.0079)	(0.011)**	(0.0123)
		Post word count	-0.0301	0.0052	-0.0356	-0.0048
			(0.0099)**	(0.0033)	(0.011)**	(0.0123)
		Max comment depth	-0.0316	0.0159	-0.0356	-0.0048
			(0.0232)	(0.0183)	(0.011)**	(0.0123)
		Proactive	-0.0404	0.0443	-0.0356	-0.0048
			(0.0121)***	(0.0119)***	(0.011)**	(0.0123)
		Includes URL	-0.0059	-0.0116	-0.0356	-0.0048
			(0.0089)	(0.0112)	(0.011)**	(0.0123)
	True		-0.0112		-0.0368	-0.0049
			(0.006)		(0.0111)***	(0.0123)
		Num comments	-0.0106	-0.0002	-0.0368	-0.0049
			(0.0109)	(0.0038)	(0.0111)***	(0.0123)
		Post word count	0.0153	-0.0057	-0.0368	-0.0049
			(0.0158)	(0.003)	(0.0111)***	(0.0123)
		Max comment depth	-0.0063	-0.0036	-0.0368	-0.0049
			(0.0101)	(0.0057)	(0.0111)***	(0.0123)
		Proactive	-0.0389	0.0326	-0.0368	-0.0049
			(0.018)*	(0.0172)	(0.0111)***	(0.0123)
		Includes URL	-0.0094	-0.0049	-0.0368	-0.0049
			(0.0066)	(0.0094)	(0.0111)***	(0.0123)
Post-2021	False		-0.0304		-0.0358	-0.005
			(0.0139)*		(0.011)**	(0.0123)
		Num comments	-0.0518	0.0144	-0.0358	-0.005
			(0.0107)***	(0.0116)	(0.011)**	(0.0123)
		Post word count	-0.0414	0.0041	-0.0358	-0.005
			(0.0055)***	(0.0045)	(0.011)**	(0.0123)
		Max comment depth	-0.0926	0.0495	-0.0358	-0.005
			(0.0319)**	(0.0297)	(0.011)**	(0.0123)
		Proactive	-0.0502	0.0463	-0.0358	-0.005
			(0.015)***	(0.0192)*	(0.011)**	(0.0123)
		Includes URL	-0.0205	-0.025	-0.0358	-0.005
			(0.0224)	(0.0294)	(0.011)**	(0.0123)
	True		-0.0439		-0.0373	-0.005
			(0.0112)***		(0.0111)***	(0.0123)
		Num comments	-0.0456	0.0007	-0.0373	-0.005
			(0.0184)*	(0.0068)	(0.0111)***	(0.0123)
		Post word count	-0.0257	-0.0039	-0.0373	-0.005
			(0.0351)	(0.0061)	(0.0111)***	(0.0123)
		Max comment depth	-0.0515	0.0061	-0.0373	-0.005
			(0.0173)**	(0.0107)	(0.0111)***	(0.0123)
		Proactive	-0.0807	0.0428	-0.0373	-0.005
			(0.0316)*	(0.0297)	(0.0111)***	(0.0123)
		Includes URL	-0.0477	0.0085	-0.0373	-0.005
			(0.0135)***	(0.017)	(0.0111)***	(0.0123)
Pre-2021	False		0.0032		-0.0373	-0.005
			(0.009)		(0.0112)***	(0.0123)
		Num comments	-0.0019	0.002	-0.0373	-0.005
			(0.013)	(0.0063)	(0.0112)***	(0.0123)
		Post word count	0.0304	-0.0059	-0.0373	-0.005
			(0.0171)	(0.0035)	(0.0112)***	(0.0123)
		Max comment depth	0.0109	-0.0055	-0.0373	-0.005
			(0.0118)	(0.0082)	(0.0112)***	(0.0123)
		Proactive	-0.0158	0.0224	-0.0373	-0.005
			(0.023)	(0.0202)	(0.0112)***	(0.0123)
		Includes URL	0.0046	-0.0041	-0.0373	-0.005
			(0.0092)	(0.0107)	(0.0112)***	(0.0123)
	True		-0.003		-0.037	-0.0052
			(0.0068)		(0.0111)***	(0.0123)
		Num comments	0.0045	-0.0029	-0.037	-0.0052
			(0.0121)	(0.0043)	(0.0111)***	(0.0123)
		Post word count	0.0251	-0.0061	-0.037	-0.0052
			(0.0181)	(0.0036)	(0.0111)***	(0.0123)
		Max comment depth	0.0091	-0.0087	-0.037	-0.0052
			(0.0117)	(0.0067)	(0.0111)***	(0.0123)
		Proactive	-0.0292	0.0309	-0.037	-0.0052
			(0.0205)	(0.0197)	(0.0111)***	(0.0123)
		Includes URL	-0.0012	-0.005	-0.037	-0.0052
			(0.0073)	(0.0111)	(0.0111)***	(0.0123)

Table 11.6: Panel regression results for the bullish post model where the dependent variable is the asset price change four weeks after the post. Standard errors are in parentheses, with * indicating levels of statistical significance.

Time Filter	Meme Stocks Excluded	Other Variable Name	Sentiment	Other Variable	CAR	$r_{[t,t-1]}$
All posts	False		-0.0037		-0.0363	-0.0029
			(0.0139)		(0.011)***	(0.012)
		Num comments	-0.0039	-0.0001	-0.0363	-0.0029
			(0.0204)	(0.0094)	(0.011)***	(0.012)
		Post word count	0.0234	-0.0063	-0.0363	-0.0029
			(0.0247)	(0.0063)	(0.011)***	(0.012)
		Max comment depth	-0.0008	-0.002	-0.0363	-0.0029
			(0.0245)	(0.0141)	(0.011)***	(0.012)
		Proactive	0.0211	-0.0331	-0.0363	-0.0029
			(0.0334)	(0.0363)	(0.011)***	(0.012)
		Includes URL	-0.012	0.0228	-0.0363	-0.0029
			(0.0162)	(0.0164)	(0.011)***	(0.012)
	True		0.0051		-0.037	-0.003
			(0.0125)		(0.0111)***	(0.0121)
		Num comments	-0.0076	-0.0046	-0.037	-0.003
			(0.0198)	(0.0081)	(0.0111)***	(0.0121)
		Post word count	0.0248	-0.0045	-0.037	-0.003
			(0.0228)	(0.0047)	(0.0111)***	(0.0121)
		Max comment depth	0.0049	0.0001	-0.037	-0.003
			(0.022)	(0.0136)	(0.0111)***	(0.0121)
		Proactive	0.0245	-0.025	-0.037	-0.003
			(0.0383)	(0.0378)	(0.0111)***	(0.0121)
		Includes URL	0.0004	0.0124	-0.037	-0.003
			(0.0145)	(0.0168)	(0.0111)***	(0.0121)
Post-2021	False		-0.0451		-0.0367	-0.0055
			(0.0354)		(0.011)***	(0.0123)
		Num comments	-0.0588	-0.0055	-0.0367	-0.0055
			(0.1576)	(0.0596)	(0.011)***	(0.0123)
		Post word count	0.029	-0.0197	-0.0367	-0.0055
			(0.1315)	(0.0342)	(0.011)***	(0.0123)
		Max comment depth	-0.2937	0.1583	-0.0367	-0.0055
			(0.2155)	(0.1456)	(0.011)***	(0.0123)
		Proactive	-0.0293	-0.0385	-0.0367	-0.0055
			(0.0721)	(0.1319)	(0.011)***	(0.0123)
		Includes URL	-0.0395	-0.0171	-0.0367	-0.0055
			(0.0494)	(0.105)	(0.011)***	(0.0123)
	True		-0.0389		-0.0375	-0.0056
			(0.0569)		(0.0111)***	(0.0123)
		Num comments	-0.391	-0.1153	-0.0375	-0.0056
			(0.1963)*	(0.0553)*	(0.0111)***	(0.0123)
		Post word count	-0.5337	0.0934	-0.0375	-0.0056
			(0.4621)	(0.0824)	(0.0111)***	(0.0123)
		Max comment depth	-0.1592	0.0861	-0.0375	-0.0056
			(0.1234)	(0.0839)	(0.0111)***	(0.0123)
		Proactive	-0.2465	0.2475	-0.0375	-0.0056
			(0.2097)	(0.2032)	(0.0111)***	(0.0123)
		Includes URL	-0.0532	0.0317	-0.0375	-0.0056
			(0.0712)	(0.111)	(0.0111)***	(0.0123)
Pre-2021	False		0.0001		-0.0371	-0.0027
			(0.0146)		(0.0111)***	(0.0121)
		Num comments	0.0031	0.0011	-0.0371	-0.0027
			(0.0194)	(0.0081)	(0.0111)***	(0.0121)
		Post word count	0.0226	-0.0052	-0.0371	-0.0027
			(0.0229)	(0.0046)	(0.0111)***	(0.0121)
		Max comment depth	0.0124	-0.0088	-0.0371	-0.0027
			(0.0223)	(0.0146)	(0.0111)***	(0.0121)
		Proactive	0.0334	-0.0427	-0.0371	-0.0027
			(0.0387)	(0.0392)	(0.0111)***	(0.0121)
		Includes URL	-0.0094	0.0258	-0.0371	-0.0027
			(0.0188)	(0.0212)	(0.0111)***	(0.0121)
	True		0.0069		-0.037	-0.0027
			(0.013)		(0.0111)***	(0.0121)
		Num comments	0.0058	-0.0004	-0.037	-0.0027
			(0.0192)	(0.0082)	(0.0111)***	(0.0121)
		Post word count	0.0298	-0.0053	-0.037	-0.0027
			(0.0221)	(0.0047)	(0.0111)***	(0.0121)
		Max comment depth	0.01	-0.0023	-0.037	-0.0027
			(0.0225)	(0.0139)	(0.0111)***	(0.0121)
		Proactive	0.0325	-0.033	-0.037	-0.0027
			(0.0392)	(0.0388)	(0.0111)***	(0.0121)
		Includes URL	0.0024	0.0122	-0.037	-0.0027
			(0.0151)	(0.0172)	(0.0111)***	(0.0121)

Table 11.7: Panel regression results for the bearish post model where the dependent variable is the asset price change four weeks after the post. Standard errors are in parentheses, with * indicating levels of statistical significance.

Time Filter	Meme Stocks Excluded	Other Variable Name	Sentiment	Other Variable	CAR	$r_{[t,t-1]}$
All posts	False		0.0158		-0.0512	-0.011
			(0.0312)		(0.0148)***	(0.0162)
		Num comments	0.0067	0.0042	-0.0512	-0.011
			(0.0265)	(0.0076)	(0.0148)***	(0.0162)
		Post word count	0.0672	-0.0135	-0.0512	-0.011
			(0.0191)***	(0.0038)***	(0.0148)***	(0.0162)
		Max comment depth	-0.0377	0.0398	-0.0512	-0.011
			(0.0181)*	(0.0275)	(0.0148)***	(0.0162)
		Proactive	0.0033	0.0183	-0.0512	-0.011
			(0.0445)	(0.0218)	(0.0148)***	(0.0162)
		Includes URL	0.0174	-0.0043	-0.0512	-0.011
			(0.0339)	(0.0155)	(0.0148)***	(0.0162)
	True		-0.0224		-0.0513	-0.0109
			(0.0101)*		(0.0148)***	(0.0162)
		Num comments	-0.0319	0.0038	-0.0513	-0.0109
			(0.0189)	(0.0074)	(0.0148)***	(0.0162)
		Post word count	0.0497	-0.0156	-0.0513	-0.0109
			(0.0278)	(0.0058)**	(0.0148)***	(0.0162)
		Max comment depth	-0.0383	0.0117	-0.0513	-0.0109
			(0.0146)**	(0.0085)	(0.0148)***	(0.0162)
		Proactive	-0.09	0.0793	-0.0513	-0.0109
			(0.0288)**	(0.0268)**	(0.0148)***	(0.0162)
		Includes URL	-0.025	0.0071	-0.0513	-0.0109
			(0.0115)*	(0.0139)	(0.0148)***	(0.0162)
Post-2021	False		0.0241		-0.0512	-0.0109
			(0.0481)		(0.0148)***	(0.0162)
		Num comments	0.0105	0.0092	-0.0511	-0.0109
			(0.0297)	(0.0223)	(0.0148)***	(0.0162)
		Post word count	0.0599	-0.0133	-0.0512	-0.0109
			(0.0129)***	(0.0113)	(0.0148)***	(0.0162)
		Max comment depth	-0.0925	0.0927	-0.0511	-0.0109
			(0.0383)*	(0.0511)	(0.0148)***	(0.0162)
		Proactive	0.0131	0.0257	-0.0512	-0.0109
			(0.0478)	(0.0286)	(0.0148)***	(0.0162)
		Includes URL	0.0354	-0.0284	-0.0511	-0.0109
			(0.0624)	(0.0492)	(0.0148)***	(0.0162)
	True		-0.0795		-0.0515	-0.0104
			(0.017)***		(0.0148)***	(0.0161)
		Num comments	-0.0852	0.0026	-0.0515	-0.0104
			(0.0314)**	(0.0127)	(0.0148)***	(0.0161)
		Post word count	-0.0112	-0.0147	-0.0515	-0.0104
			(0.0539)	(0.0111)	(0.0148)***	(0.0161)
		Max comment depth	-0.1014	0.0175	-0.0515	-0.0104
			(0.0289)***	(0.0189)	(0.0148)***	(0.0161)
		Proactive	-0.2773	0.2298	-0.0515	-0.0104
			(0.0591)***	(0.0607)***	(0.0148)***	(0.0161)
		Includes URL	-0.1023	0.0509	-0.0515	-0.0104
			(0.0218)***	(0.0291)	(0.0148)***	(0.0161)
Pre-2021	False		0.0102		-0.0522	-0.0107
			(0.02)		(0.0148)***	(0.0162)
		Num comments	-0.0057	0.0061	-0.0522	-0.0107
			(0.0206)	(0.0096)	(0.0148)***	(0.0162)
		Post word count	0.1025	-0.0202	-0.0522	-0.0107
			(0.0447)*	(0.0072)**	(0.0148)***	(0.0162)
		Max comment depth	-0.011	0.0151	-0.0522	-0.0107
			(0.0158)	(0.0132)	(0.0148)***	(0.0162)
		Proactive	-0.0212	0.037	-0.0522	-0.0107
			(0.0332)	(0.0248)	(0.0148)***	(0.0162)
		Includes URL	0.009	0.0034	-0.0522	-0.0107
			(0.0194)	(0.0153)	(0.0148)***	(0.0162)
	True		-0.0081		-0.0519	-0.0111
			(0.0116)		(0.0148)***	(0.0162)
		Num comments	-0.009	0.0004	-0.0519	-0.0111
			(0.0209)	(0.0083)	(0.0148)***	(0.0162)
		Post word count	0.0629	-0.0154	-0.0519	-0.0111
			(0.0299)*	(0.0061)*	(0.0148)***	(0.0162)
		Max comment depth	-0.0163	0.0058	-0.0519	-0.0111
			(0.0156)	(0.0096)	(0.0148)***	(0.0162)
		Proactive	-0.0465	0.0452	-0.0519	-0.0111
			(0.0281)	(0.0264)	(0.0148)***	(0.0162)
		Includes URL	-0.0086	0.0014	-0.0519	-0.0111
			(0.0125)	(0.0155)	(0.0148)***	(0.0162)

Table 11.8: Panel regression results for the bullish post model where the dependent variable is the asset price change twelve weeks after the post. Standard errors are in parentheses, with * indicating levels of statistical significance.

Time Filter	Meme Stocks Excluded	Other Variable Name	Sentiment	Other Variable	CAR	$r_{[t,t-1]}$
All posts	False		-0.0098		-0.0513	-0.0084
			(0.0201)		(0.0148)***	(0.0159)
		Num comments	-0.018	-0.003	-0.0513	-0.0084
			(0.0226)	(0.0111)	(0.0148)***	(0.0159)
		Post word count	0.0488	-0.0137	-0.0513	-0.0084
			(0.0298)	(0.0076)	(0.0148)***	(0.0159)
		Max comment depth	-0.0219	0.0086	-0.0513	-0.0084
			(0.0279)	(0.0139)	(0.0148)***	(0.0159)
		Proactive	0.0165	-0.0351	-0.0513	-0.0084
			(0.0344)	(0.0318)	(0.0148)***	(0.0159)
		Includes URL	-0.0309	0.058	-0.0513	-0.0084
			(0.0234)	(0.0234)*	(0.0148)***	(0.0159)
	True		0.0012		-0.0516	-0.0087
			(0.0195)		(0.0148)***	(0.0159)
		Num comments	-0.0343	-0.0129	-0.0516	-0.0087
			(0.0217)	(0.0093)	(0.0148)***	(0.0159)
		Post word count	0.0414	-0.0093	-0.0516	-0.0087
			(0.0338)	(0.0073)	(0.0148)***	(0.0159)
		Max comment depth	-0.0172	0.0132	-0.0516	-0.0087
			(0.0276)	(0.0141)	(0.0148)***	(0.0159)
		Proactive	0.0354	-0.0439	-0.0516	-0.0087
			(0.0383)	(0.0363)	(0.0148)***	(0.0159)
		Includes URL	-0.0151	0.0438	-0.0516	-0.0087
			(0.0222)	(0.0227)	(0.0148)***	(0.0159)
Post-2021	False		-0.093		-0.0518	-0.0106
			(0.0445)*		(0.0148)***	(0.0161)
		Num comments	0.1031	0.0793	-0.0518	-0.0105
			(0.1535)	(0.0858)	(0.0148)***	(0.0161)
		Post word count	0.141	-0.0622	-0.0518	-0.0106
			(0.0991)	(0.0416)	(0.0148)***	(0.0161)
		Max comment depth	-0.1093	0.0104	-0.0518	-0.0106
			(0.2298)	(0.1442)	(0.0148)***	(0.0161)
		Proactive	-0.0423	-0.1243	-0.0518	-0.0106
			(0.0322)	(0.0988)	(0.0148)***	(0.0161)
	True		-0.0005		-0.0521	-0.0108
			(0.0729)		(0.0148)***	(0.0162)
		Num comments	-0.2574	-0.0841	-0.0521	-0.0108
			(0.1374)	(0.0542)	(0.0148)***	(0.0162)
		Post word count	0.1286	-0.0244	-0.0521	-0.0108
			(0.2413)	(0.0478)	(0.0148)***	(0.0162)
		Max comment depth	0.0101	-0.0076	-0.0521	-0.0108
			(0.1246)	(0.1226)	(0.0148)***	(0.0162)
		Proactive	0.0893	-0.107	-0.0521	-0.0108
			(0.0894)	(0.063)	(0.0148)***	(0.0162)
		Includes URL	-0.0493	0.1081	-0.0521	-0.0108
			(0.0492)	(0.1519)	(0.0148)***	(0.0162)
Pre-2021	False		-0.0021		-0.0517	-0.0084
			(0.0201)		(0.0148)***	(0.0159)
		Num comments	-0.0282	-0.0095	-0.0517	-0.0084
			(0.0215)	(0.0092)	(0.0148)***	(0.0159)
		Post word count	0.0313	-0.0078	-0.0517	-0.0084
			(0.0348)	(0.0074)	(0.0148)***	(0.0159)
		Max comment depth	-0.0187	0.0119	-0.0517	-0.0085
			(0.0274)	(0.014)	(0.0148)***	(0.0159)
		Proactive	0.0309	-0.0424	-0.0517	-0.0084
			(0.0391)	(0.0377)	(0.0148)***	(0.0159)
		Includes URL	-0.0242	0.0597	-0.0517	-0.0085
			(0.0239)	(0.0254)*	(0.0148)***	(0.0159)
	True		0.0013		-0.0516	-0.0085
			(0.0202)		(0.0148)***	(0.0159)
		Num comments	-0.026	-0.01	-0.0516	-0.0085
			(0.0216)	(0.0094)	(0.0148)***	(0.0159)
		Post word count	0.0407	-0.0092	-0.0516	-0.0085
			(0.0335)	(0.0074)	(0.0148)***	(0.0159)
		Max comment depth	-0.018	0.0138	-0.0516	-0.0085
			(0.0284)	(0.0139)	(0.0148)***	(0.0159)
		Proactive	0.0338	-0.0419	-0.0516	-0.0085
			(0.0394)	(0.0375)	(0.0148)***	(0.0159)
		Includes URL	-0.0138	0.0409	-0.0516	-0.0086
			(0.0231)	(0.0225)	(0.0148)***	(0.0159)

Table 11.9: Panel regression results for the bearish post model where the dependent variable is the asset price change twelve weeks after the post. Standard errors are in parentheses, with * indicating levels of statistical significance.

Chapter 12

Appendix D Supplementary results and material for Chapter 6

Model costs Chapter 6

Model	inference time (minutes)	inference cost (dollars)
T5	480	3.56
Mistral	720	5.35

Table 12.1: Cost comparison for generation pipelines using a Google Colaboratory compute price per hour estimates

Model hypothesis

12.0.1 Working Hypothesis, Problem Setup, and Embedding Generation

Working hypothesis.

Our approach is based on two hypotheses. First, we expect that *domain pre-training and fine-tuning on Basel regulatory text* aligns token semantics with the specialised vocabulary and hierarchical structure of prudential regulation. This allows the encoder to discriminate between expressions that appear similar in everyday language but have distinct meanings

in the Basel Framework. Second, we hypothesise that *fine-tuning on synthetic queries generated by stronger LLMs* in the Generative-Pseudo Labelling step described below leads to a more discriminative embedding space. Larger generator models (such as Mistral-7B) produce 'harder' (ie. more challenging to distinguish from the correct answer) and more realistic queries, which in turn yield harder negatives during training. Together, domain adaptation and hard-negative fine-tuning should improve the model's ability to retrieve the correct regulatory passage for a given supervisory query. We test this hypothesis by comparing the performance in a retrieval task of an embedding model fine-tuned and pre-trained on Basel regulatory text against the performance of open-source top-performing models.

Algorithm 1 PRET Training via GPL

```
for each Basel passage P do
    Generate query Q using T5 or Mistral-7B
    Retrieve candidate negatives and select hard negative P⁻
    Compute teacher-encoder margin δ
    Compute student-retriever margin δ̂
    Update encoder using MarginMSE loss
end for
```

Problem setup: generative pseudo-labelling.

We adopt the Generative Pseudo-Labelling (GPL) [238] framework to construct training triplets (Q, P^+, P^-) without human annotation. For each Basel passage P, we generate a query Q using two language models: **T5**, the reference model used in the original GPL paper, and **Mistral-7B**, a strong open-source model available during our research period on the device sizes we had at our disposal. In practice, Mistral-7B produces longer, more specific, and more realistic supervisory-style queries, resulting in better hard negatives for dense retriever training. These generative models have been used in an out of the box setting and required no tuning. To compute passage similarity in order to identify suitable negative P^- candidate passages, all queries Q, answers P^+ and negative P^- passages are embedded using the strongest open source sentence-Bert embedding model available at time of writing : msmarco-distilbert-base-v3.

Problem setup: training and testing the encoding model

To train the student embedding models (whose parameters we represent as θ, we use the **MarginMSE loss**, which teaches the encoder to mimic the score difference between positive and negative passages as computed by a teacher embedding model for which we choose msmarco-distilbert-base-v3 with parameters *teacher*. Let δ_i denote the cross-encoder score margin and $\hat{\delta}_i$ the margin produced by the dense retriever we whish to train.

The loss is:

$$\mathcal{L}_{\text{MarginMSE}}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\delta}_i - \delta_i\right)^2$$

where similarity between any two embedding vectors a and b is measured as cosine similarity.

$$\text{sim}(a,b) = \frac{a^T b}{\|a\|_2 \, \|b\|_2}$$

therefore we have the margin for each triplet $\hat{\delta}_i$

$$\hat{\delta}_i = \text{sim}(f_\theta(Q_i), f_\theta(P_i^+)) - \text{sim}(f_\theta(Q_i), f_\theta(P_i^-)).$$

and the target margin

$$\delta_i = \text{sim}(f_{teacher}(Q_i), f_{teacher}(P_i^+)) - \text{sim}(f_{teacher}(Q_i), f_{teacher}(P_i^-)).$$

This loss measures the distance between the positive and negative pairs as measured by the original embedding model and the learnt embedding model. The purpose is to fine tune the student embedding model f_θ such that it learns to differentiate between financial specific vocabulary terms contained in the queries Q, answers P^+ and negative P^- passages in a way where the query is closer to the positive than to the negative passage.

Embedding model and representation.

As described above, we use an encoder architecture, where the same transformer

$$q = f_\theta(Q), \qquad p = f_\theta(P)$$

maps both the query and passage into a shared vector space. We evaluate two encoders—**BGE-Large** and **RoBERTa-Large**—both producing 1024-dimensional embeddings.

Accuracy@k Retrieval Performance Evaluation

Retrieval setup.

At inference time, both supervisory queries and Basel regulatory passages are mapped into the same dense embedding space by the encoder introduced above. Given a query Q

and a collection of candidate passages $\{P_1, \ldots, P_N\}$, the encoder produces embeddings

$$q = f_\theta(Q), \qquad p_j = f_\theta(P_j).$$

Similarity between q and each p_j is computed as cosine similarity.

For each query, passages are ranked in descending order of similarity. The top-k passages form the retrieved set $L_Q = (P_{(1)}, \ldots, P_{(k)})$.

Evaluation metric: Accuracy@k.

We evaluate retrieval performance using the `accuracy_at_k` measure in the sentence-transformers framework [197]. Each supervisory query Q is associated with one gold-standard passage R_Q which we create for each of the retrieval tasks Basel-FAQ, Basel-FOOTNOTE and BASEL-LINK as described in 6.4. Retrieval for Q is counted as correct if at least one gold passage appears in the top-k retrieved set:

$$\text{Correct}(Q, k) = \begin{cases} 1, & \text{if } L_Q \cap R_Q \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases}$$

The overall retrieval accuracy is computed for each of the retrieval tasks as:

$$\text{Accuracy@}k = \frac{1}{|Q|} \sum_{Q} \text{Correct}(Q, k).$$

Accuracy@k measures the proportion of supervisory queries for which the retriever successfully surfaces at least one correct regulatory passage within the top k ranked results. This metric is aligned with real supervisory workflows, where finding one correct regulatory reference quickly is typically more valuable than producing a complete ranking of all relevant text segments.

In summary, we first encode both the supervisory questions and the candidate regulatory passages into numerical vector embeddings. We then compare each query with every passage using cosine similarity and rank the passages from most to least semantically similar. If any of the passages identified in the top k positions correspond to a correct regulatory reference, we count the retrieval as successful. Accuracy@k is the percentage of times the system succeeds across all test queries. This evaluation therefore reflects how reliably the model can retrieve a usable regulatory grounding passage for real-world supervisory analysis.

Chapter 13

Appendix E Supplementary results and material for Chapter 7

Relatio Modelling description

Using the Relatio model as described in the Relatio pipeline in Figure 7.1 is a largely tuning-free process. We will describe each step of the diagram highlighting the tuning steps necessary where relevant. Once the text is sent through the pipeline, each piece of text is deterministically split sentence by sentence in step b). Each sentence is passed through the Spacy python package's named entity recognition algorithm in step c) which identifies named entities in text such as 'Microsoft', 'Tesla' or 'America'. In parallel, each sentence is passed through the SRL semantic role labelling in d) which extracts the actor - verb - passive triplets as described in Rapp [196] such as 'The government' (Actor) - 'funds' (Verb) - 'microprocessors' (Patient). SRL is a deterministic step with no tuning. We assign an actor or patient such as 'Microsoft' or 'government' to the relevant named entity if that named entity is contained in the actor / patient text segment in step i). For the actor / passives which do not contain a named entity, we embed them using the strongest sentence embedding model at the time of writing : all-mpnet-v2 by Song et al. [215].

Step e) and i) is omitted in this application as we use an open source sentence embedding tool (all-mpnet-v2 [215]) from the Sentence BERT library [197] as our goal is not to fine-tune financial specific text embedding tools. Similarly step h) verb negation is omitted as we only found 5% of verb captured by the SRL step had a negative attribute, so they were instead dropped.

Therefore the only tuning step is finding the optimal number of K-Means clusters to group text passages that do not contain any named entities in step n). This process is done by fixing a set of candidate cluster numbers from deciles of the total entity semantic roles that are not associated with a named entity. To measure how good a cluster number

is, we look at the silhouette score. The silhouette score measures how well each data point fits its assigned cluster. It calculates the mean intra-cluster and mean nearest cluster distance for each point. Then we calculate as in 13.1 the normalized difference between how well a point fits in its assigned cluster vs how well the point would fit in the nearest other non assigned cluster. The choice of clusters is performed by looking at the highest average silhouette score calculated in Equation13.4 of the different k-means options.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \tag{13.1}$$

where:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\substack{j \in C_i \\ j \neq i}} d(i,j), \qquad \text{(mean intra-cluster distance)} \tag{13.2}$$

$$b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j), \qquad \text{(mean nearest-cluster distance)} \tag{13.3}$$

and the overall average silhouette score is:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{13.4}$$

Decile (d)	Number of Clusters ($k = d$)	Mean Silhouette Score (S_d)
1	9171	0.31
2	18342	0.32
3	**27513**	**0.43**
4	36684	0.38
5	45855	0.36
6	55026	0.35
7	64197	0.32
8	73368	0.24
9	82539	0.19
10	91711	0.22

Table 13.1: Average silhouette scores for K-Means clustering by decile-based grouping over 91,711 samples. The 3rd decile yields the highest score, indicating the most coherent clustering.

This is an example of what this process look like :

Raw sentence: “Microsoft develops cloud computing services for government agencies.”

- Named Entities (Step c): `Microsoft`,

- SRL Triplet (Step d): Actor = "Microsoft", Verb = "develops", Patient = "cloud computing services"
- Because no named entities are present for the patient, cloud computing services are embedded using all-mpnet-v2 and then clustered (Step n) yielding the closest cluster 'government'.
- Resulting Relatio Triple: `Microsoft` — develops —> `cloud computing`

Graph construction

Once we obtain the series of narrative statements, we combine them to yield a directed network of statements. The network has the following properties. In order to specifically collect causal narratives, we only conserve the narratives present in sentences with a causal statement ['because','hence','therefore','due to','since','as a result of','causes','caused','result in','results in']. This strongly downsamples the total number of relationships and therefore makes the resulting graph much sparser, but gives us a network which represents causal relationships as stated in the news article.

Fig 13.2 describes our dataset's total size in number of articles and number of sentences.

# articles	2500
# sentences	110000
# sentences with semantic role	131800
# Named entities	7603
# Semantic Roles	121392
# Semantic Roles without named entity	91711
# Semantic Roles post clustering	27513
# Unique narratives	10964
# Unique narratives in causal sentence	3494
# Nodes in narrative graph	2830

Table 13.2: Dataset description

Node2Vec embedding

Step 1: Random Walk Generation

Consider the directed narrative graph $G = (V, E)$, where each node $v \in V$ corresponds to a narrative entity and each directed edge $(u \to v) \in E$ represents a causal statement

extracted from the text corpus. To capture the multi-hop relational structure of the graph, Node2Vec performs multiple random walks from each node. Each walk is a sequence

$$W = (v_0, v_1, \ldots, v_L),$$

where v_0 is sampled uniformly from V, and each subsequent node v_{t+1} is sampled from the out-neighbours of v_t. Using the following Node2Vec parameters ($p = 1q = 0.5$). We choose these parameters as setting a lower q than p makes (according to the original Node2Vec paper Grover and Leskovec [106]) the walks from each node reflect the communities they share. The resulting set of walks can be viewed as a corpus of "sentences" describing how causal influence flows across the network. Two nodes will be located closer if their causal neighbourhood is more similar.

Step 2: Skip-Gram Representation Learning

Once the random walks have been generated, we treat each walk as if it were a "sentence" and each node as if it were a word appearing in that sentence. The core idea is that nodes that frequently appear in similar walk-contexts should receive similar embeddings. This is operationalized by training a skip-gram model, originally introduced in the Word2Vec framework [169].

For a node v, and for every random walk in which it appears, we consider the nodes that occur within a fixed window length w around it in the walk:

$$\mathcal{N}_w(v) = \{u \mid u \text{ occurs within } w \text{ positions of } v \text{ in any random walk}\}.$$

The skip-gram objective then seeks to learn a function

$$f : V \to \mathbb{R}^d$$

that assigns each node a d-dimensional embedding vector $f(v)$, such that the vector for v is predictive of the vectors of its context nodes $c \in \mathcal{N}_w(v)$. We arbitrarily define d equal to 128.

Formally, the embedding vectors are chosen to maximize the probability that nodes are more similar to nodes in their walk-contexts.

$$\max_f \sum_{v \in V} \sum_{c \in \mathcal{N}_w(v)} \log \Pr(c \mid v).$$

We parametrise the conditional probability $\Pr(c \mid v)$ using a softmax function:

$$\Pr(c \mid v) = \frac{\exp(f(c)^\top f(v))}{\sum_{u \in V} \exp(f(u)^\top f(v))}.$$

This objective encourages similarity in the dot-product space between nodes that share many walk-contexts. Indeed, the embedding function is created to

Step 3: Firm Embedding Extraction

From the full set of nodes V, we identify the subset $F \subseteq V$ corresponding to the firms in our sample. The learned embedding for each firm $i \in F$ is given by

$$e_i = f(i).$$

These vectors capture the firm's position within the broader causal narrative environment, reflecting both direct and multi-hop relationships.

Step 4: Computing Inter-Firm Similarity

To quantify narrative proximity between any two firms i and j, we compute the cosine similarity between their embeddings:

$$\text{Similarity}(i, j) = \frac{e_i^\top e_j}{\|e_i\|_2 \, \|e_j\|_2}.$$

This measure forms the *Semantic Network Similarity (Directed)* variable used in the empirical analysis. The *Semantic Network Similarity (Undirected)* variable used in the regression is built using the undirected version of the network G.

Control Variable Construction

To benchmark the explanatory power of our narrative similarity measure, we construct a set of standard control variables capturing alternative channels through which firms may exhibit common risk exposure.

Same Sector.

We first include a categorical industry proximity measure based on firms' primary 4-digit SIC codes. For a firm pair (i, j), the *Same Sector* indicator takes value 1 if firms i and j share the same 4-digit SIC classification, and 0 otherwise. This control captures

similarities in production technology, regulatory exposure, and market structure that are typically associated with industry grouping.

Text Similarity.

Second, to account for surface-level similarity in firms' disclosures, we compute a cosine similarity score over TF–IDF representations of each firm's annual 10-K filing. Specifically, each filing is converted into a bag-of-words document vector weighted by term frequency–inverse document frequency, and similarity between two firms is computed as

$$\text{TextSim}(i,j) = \frac{doc_i^\top doc_j}{\|doc_i\|_2 \, \|doc_j\|_2},$$

where doc_i and doc_j denote the TF–IDF vectors of firms i and j respectively. This measure reflects the extent to which firms use similar language and terminology in their SEC filings disclosures.

Entity Similarity.

Third, to ensure that the effect of our narrative similarity measure is not solely driven by shared mention of similar economic entities, we construct an entity-overlap control. Each filing is represented as a frequency vector over the Relatio-derived entity clusters appearing in the document. We then compute cosine similarity between these entity-cluster count vectors:

$$\text{EntitySim}(i,j) = \frac{e_i^\top e_j}{\|e_i\|_2 \, \|e_j\|_2},$$

where e_i and e_j is the vector of entity-cluster frequencies for firm i and j respectively. This captures the extent to which firms reference the same economic actors, resources, and risk factors, independent of how these entities are narratively positioned in the causal graph.

Together, these controls allow us to isolate the additional explanatory power of the narrative-graph-based similarity measure beyond shared industry classification, lexical similarity in filings, and overlap in mentioned entities.

Bibliography

[1] Accornero, M. and Moscatelli, M. (2018). Listening to the buzz: social media sentiment and retail depositors' trust. *Bank of Italy Temi di Discussione (Working Paper) No*, 1165.

[2] Acemoglu, D., Akcigit, U., and Kerr, W. (2016). Networks and the macroeconomy: An empirical exploration. *NBER Macroeconomics Annual*, 30(1):273–335.

[3] Acemoglu, D., Ozdaglar, A., and Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.

[4] Acemoglu, D., Ozdaglar, A. E., and Tahbaz-Salehi, A. (2013). The Network Origins of Large Economic Downturns. *SSRN Electronic Journal*.

[5] Achitouv, I., Gorduza, D., and Jacquier, A. (2023). Natural language processing for financial regulation. *ArXiv preprint arXiv:2311.08533*.

[6] Adosoglou, G., Park, S., Lombardo, G., Cagnoni, S., and Pardalos, P. M. (2022). Lazy Network: A Word Embedding-Based Temporal Financial Network to Avoid Economic Shocks in Asset Pricing Models. *Complexity*, 2022.

[7] Aggarwal, C. C. (2018). Information Retrieval and Search Engines. *Machine Learning for Text*, pages 259–304.

[8] Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, 53:16–38.

[9] Agrawal, S., Azar, P. D., Lo, A. W., and Singh, T. (2018). Momentum, mean-reversion, and social media: Evidence from stocktwits and Twitter. *The Journal of Portfolio Management*, 44(7):85–95.

[10] Agrawal, Y., Anand, V., Gupta, M., Arunachalam, S., and Varma, V. (2021). Goal-Directed Extractive Summarization of Financial Reports. *International Conference on Information and Knowledge Management, Proceedings*, pages 2817–2821.

[11] Ahrens, M. and McMahon, M. (2021). Extracting economic signals from central bank speeches. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics.

[12] Al-Moslmi, T., Gallofre Ocana, M., Opdahl, A. L., and Veres, C. (2020). Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8:32862–32881.

[13] Ali, U. and Hirshleifer, D. (2020). Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics*, 136(3):649–675.

[14] Allen, F. and Gale, D. (2000). Financial contagion. *Journal of Political Economy*, 108(1):1–33.

[15] Amadxarif, Z., Brookes, J., Garbarino, N., Patel, R., Walczak, E., Bholat, D., Black, J., Colliard, J.-E., Danielsson, J., Dickinson, S., Francis, B., Grout, P., Haldane, A., Lodge, M., Moscatelli, M., Saporta, V., and Tanaka and, M. (2021). Staff Working Paper No. 834 The language of rules: textual complexity in banking reforms The language of rules: textual complexity in banking reforms. *Bank of England Staff Working Papers Series*.

[16] Anand, K., van Lelyveld, I., Banai, Á., Friedrich, S., Garratt, R. J., Halaj, G., Fique, J., Hansen, I., Jaramillo, S. M., Lee, H., Molina-Borboa, J. L., Nobili, S., Rajan, S. S., Salakhova, D., Silva, T. C., Silvestri, L., and de Souza, S. R. S. (2017). The missing links: A global study on uncovering financial network structures from partial data. *ERN: Central Banks - Policies (Topic)*.

[17] Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2006). Chapter 15 Volatility and Correlation Forecasting. *Handbook of Economic Forecasting*, 1:777–878.

[18] Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2024). Narratives about the macroeconomy. *Leibniz Institute for Financial Research SAFE*, (426).

[19] Angelov, D. (2020). Top2vec: Distributed representations of topics. *ArXiv*, abs/2008.09470. preprint.

[20] Antweiler, W. and Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *Journal of Finance*, 59(3):1259–1294.

[21] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. https://arxiv.org/abs/1908.10063.

[22] Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

[23] Ash, E., Gauthier, G., and Widmer, P. (2024). Relatio : Text Semantics Capture Political and Economic Narratives . *Cambridge University Press*, 32(1):115–132.

[24] Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15:659–688.

[25] Aste, T. (2025). Information filtering networks: Theoretical foundations, generative methodologies, and real-world applications. *ArXiv*, abs/2505.03812.

[26] Aste, T., di Matteo, T., and Hyde, S. T. (2004). Complex networks on hyperbolic surfaces. *Physica A-statistical Mechanics and Its Applications*, 346:20–26.

[27] Azar, P. D. and Lo, A. W. (2016). The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds. *The Journal of Portfolio Management*, 42(5):123–134.

[28] Aziz, S., Dowling, M., Hammami, H., and Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach. *European Financial Management*, 28(3):744–770.

[29] Baid, P., Gupta, A., and Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179(7):45–49.

[30] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CEUR Workshop Proceedings*, 1773.

[31] Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636.

[32] Balaguer, A., Benara, V., Cunha, R. L. d. F., Filho, R. d. M. E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., and Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture.

[33] Balash, V., Chekmareva, A., Faizliev, A., Sidorov, S., Mironov, S., and Volkov, D. (2019). Analysis of news flow dynamics based on the company co-mention network characteristics. *Studies in Computational Intelligence*, 813:521–533.

[34] Bardoscia, M., Barucca, P., Battiston, S., Caccioli, F., Cimini, G., Garlaschelli, D., Saracco, F., Squartini, T., and Caldarelli, G. (2021). The physics of financial networks. *Nature Reviews Physics*, 3(7):490–507.

[35] Battiston, S., Gatti, D. D., Gallegati, M., Greenwald, B., and Stiglitz, J. E. (2012a). Default cascades: When does risk diversification increase stability? *Journal of Financial Stability*, 8(3):138–149.

[36] Battiston, S., Puliga, M., Kaushik, R., Tasca, P., and Caldarelli, G. (2012b). DebtRank: Too central to fail? Financial networks, the FED and systemic risk. *Scientific Reports*, 2(1):1–6.

[37] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

[38] Baz, J., Granger, N., Harvey, C. R., Roux, N. L., and Rattray, S. (2015). Dissecting investment strategies in the cross section and time series. *Risk Management eJournal*.

[39] Bekkerman, R., Fich, E. M., and Khimich, N. V. (2023). The effect of innovation similarity on asset prices: Evidence from patents' big data. *The Review of Asset Pricing Studies*, 13(1):99–145.

[40] Bereska, L. and Gavves, E. (2024). Mechanistic interpretability for ai safety–a review. *ArXiv preprint arXiv:2404.14082*.

[41] Bernanke, B. (2004). What Policymakers Can Learn from Asset Prices. *Remarks before The Investment Analyst Society of Chicago*. Accessed: 2025-10-20.

[42] Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559.

[43] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003a). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

[44] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

[45] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1–8.

[46] Boss, M., Elsinger, H., Summer, M., and Thurner, S. (2004). Network topology of the interbank market. *Quantitative Finance*, 4(6):677–684.

[47] Boylston, C., Palacios, B., Tassev, P., and Bruckman, A. (2021). Wallstreetbets: Positions or ban. *ArXiv*, abs/2101.12110.

[48] Bradley, D., Hanousek Jr, J., Jame, R., and Xiao, Z. (2021). Place your bets? the market consequences of investment advice on reddit's wallstreetbets. *The Market Consequences of Investment Advice on Reddit's Wallstreetbets (March 15, 2021)*.

[49] Brownlees, C. T. and Engle, R. F. (2012). Volatility, Correlation and Tails for Systemic Risk Measurement. *SSRN Electronic Journal*.

[50] Brummitt, C. D. and Kobayashi, T. (2015). Cascades in multiplex financial networks with debts of different seniority. *Physical Review E*, 91(6):062813.

[51] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.

[52] Buz, T. and de Melo, G. (2021). Should you take investment advice from wallstreetbets? a data-driven approach. preprint.

[53] Caccioli, F., Barucca, P., and Kobayashi, T. (2018). Network models of financial systemic risk: a review. *Journal of Computational Social Science*, 1(1):81–114.

[54] Campbell, R., Koedijk, K., and Kofman, P. (2002). Increased Correlation in Bear Markets. *Financial Analysts Journal*, 58(1):87–94.

[55] Cao, B., Wang, S., Lin, X., Wu, X., Zhang, H., Ni, L. M., and Guo, J. (2025). From deep learning to LLMs: a survey of AI in quantitative investment. *arXiv preprint arXiv:2503.21422*.

[56] Cao, J., Chordia, T., and Lin, C. (2016). Alliances and return predictability. *Journal of Financial and Quantitative Analysis*, 51(5):1689–1717.

[57] Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies*, 36(9):3603–3642.

[58] Carriero, V. A., Azzini, A., Baroni, I., Scrocca, M., and Celino, I. (2024). Human evaluation of & nbsp;procedural knowledge graph extraction from & nbsp;text with & nbsp;large language models. In *Knowledge Engineering and Knowledge Management: 24th International Conference, EKAW 2024, Amsterdam, The Netherlands, November 26–28, 2024, Proceedings*, page 434–452, Berlin, Heidelberg. Springer-Verlag.

[59] Casanova, P., Lio, A. R. P., and Bengio, Y. (2018). Graph attention networks. *ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio*.

[60] Chacon, R. G., Morillon, T. G., and Wang, R. (2022). Will the reddit rebellion take you to the moon? evidence from wallstreetbets. *Financial Markets and Portfolio Management*, pages 1–25.

[61] Chalkidis, I., Fergadiotis, M., Manginas, N., Katakalou, E., and Malakasiotis, P. (2021). Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations.

[62] Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., and Murphy, K. (2022). Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *Journal of Machine Learning Research*, 23(89):1–64.

[63] Chandra, M. (2003). Multivariate GARCH modelling of volatility and comovements in Multivariate GARCH modelling of volatility and comovements in Asia Pacific markets Asia Pacific markets. *Edith Cowan University*.

[64] Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

[65] Chen, Z., Chu, L., Liang, D., and Tu, J. (2022). Far away from home: Investors' underreaction to geographically dispersed information. *Journal of Economic Dynamics and Control*, 136:104325.

[66] Cho, K., van Merrienboer, B., Çaglar Gülçehre, Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.

[67] Cinelli, M., Ficcadenti, V., and Riccioni, J. (2021). The interconnectedness of the economic content in the speeches of the us presidents. *Annals of Operations Research*, 299:593–615.

[68] Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 56(4):837–877.

[69] Cohen, L. and Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011.

[70] Cohen, L. and Lou, D. (2012). Complicated firms. *Journal of financial economics*, 104(2):383–400.

[71] Collier, P. and Tuckett, D. (2021). Narratives as a coordinating device for reversing regional disequilibrium. *Oxford Review of Economic Policy*, 37(1):97–112.

[72] Cont, R., Moussa, A., and Santos, E. B. e. (2013). Network structure and systemic risk in banking systems. In Fouque, J.-P. and Langsam, J. A., editors, *Handbook of Systemic Risk*, pages 327–367. Cambridge University Press.

[73] Correa, R., Garud, K., Londoño, J. M., and Mislang, N. (2017). Sentiment in Central Banks' Financial Stability Reports. *Monetary Economics: Central Banks - Policies & Impacts eJournal*.

[74] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

[75] Cortez, P., Moro, S., Rita, P., King, D., and Hall, J. (2018). Insights from a text mining survey on Expert Systems research from 2000 to 2016. *Expert Systems*, 35(3):e12280.

[76] Cortis, K., Freitas, A., Daudert, T., Hürlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of SemEval*, pages 519–535.

[77] Cumming, G. S. and Peterson, G. D. (2017). Unifying Research on Social–Ecological Resilience and Collapse. *Trends in Ecology and Evolution*, 32(9):695–713.

[78] Das, S. R., Kim, S., and N., O. D. (2019). Dynamic Systemic Risk:. *Journal of Financial Data Science*, pages 141–158.

[79] de la Concha, A., Martinez-Jaramillo, S., and Carmona, C. (2018). Multiplex financial networks: Revealing the level of interconnectedness in the banking system. In *Complex Networks & Their Applications VI: Proceedings of Complex Networks 2017 (The Sixth International Conference on Complex Networks and Their Applications)*, pages 1135–1148. Springer.

[80] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. preprint.

[81] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

[82] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[83] Ding, X., Zhang, Y., Liu, T., and Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1415–1425.

[84] Ding, X., Zhang, Y., Liu, T., and Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. In *Proceedings of IJCAI*.

[85] Du, K., Zhao, Y., Mao, R., Xing, F., and Cambria, E. (2024). Natural language processing in finance: A survey. *Inf. Fusion*, 115:102755.

[86] Dungey, M., Fry, R. A., Gonzalez-Hermosillo, B., and Martin, V. L. (2011). *Transmission of Financial Crises and Contagion: A Latent Factor Approach*. Number 9780199739837 in OUP Catalogue. Oxford University Press.

[87] Dungey, M. and Gajurel, D. (2014). Equity market contagion during the global financial crisis: Evidence from the world's eight largest economies. *Economic Systems*, 38(2):161–177.

[88] Duz Tan, S. and Tas, O. (2021). Social media sentiment in international stock returns and trading activity. *Journal of Behavioral Finance*, 22(2):221–234.

[89] Eliaz, K., Spiegler, R., Alexander, Y., Cassela, A., Helpman, E., Rubinstein, A., Thysen, H., and Wolton, S. (2020). A Model of Competing Narratives. *American Economic Review*, 110(12):3786–3816.

[90] Emmert-Streib, F., Tripathi, S., Yli-Harja, O., and Dehmer, M. (2018). Understanding the World Economy in Terms of Networks: A Survey of Data-Based Network Science Approaches on Economic Networks. *Frontiers in Applied Mathematics and Statistics*, 4:37.

[91] Faizliev, A., Balash, V., Petrov, V., Grigoriev, A., Melnichuk, D., and Sidorov, S. (2019). Stability analysis of company co-mention network and market graph over time using graph similarity measures. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(3):55.

[92] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

[93] Fama, E. F. and French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, 18(3):25–46.

[94] Fatemi, S. and Hu, Y. (2023). A Comparative Analysis of Fine-Tuned LLMs and Few-Shot Learning of LLMs for Financial Sentiment Analysis. *ArXiv*, abs/2312.08725.

[95] Fiedor, P. (2014). Networks in financial markets based on the mutual information rate. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(5):052801.

[96] Gai, P. (2013). 2 the robust-yet-fragile nature of financial systems. *Systemic Risk*, pages 8–27.

[97] Gao, G. P., Moulton, P. C., and Ng, D. T. (2017). Institutional ownership and return predictability across economically unrelated stocks. *Journal of Financial Intermediation*, 31:45–63.

[98] Gao, R., Zhang, Z., Shi, Z., Xu, D., Zhang, W., and Zhu, D. (2021). A review of natural language processing for financial technology. In *International Symposium on Artificial Intelligence and Robotics 2021*, volume 11884, pages 262–277. SPIE.

[99] Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

[100] Goergen, M., Renneboog, L., and Zhao, Y. (2019). Insider trading and networked directors. *Journal of Corporate Finance*, 56:152–175.

[101] Gomes, A., Gopalan, R., Leary, M. T., and Marcet, F. (2023). Analyst coverage networks and corporate financial policies. *Management Science*.

[102] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

[103] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. (2024). OLMo: Accelerating the Science of Language Models.

[104] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

[105] Grover, A. and Leskovec, J. (2016a). Node2vec: Scalable feature learning for networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:855–864.

[106] Grover, A. and Leskovec, J. (2016b). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA. Association for Computing Machinery.

[107] Gupta, A., Dengre, V., Kheruwala, H. A., and Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation 2020 6:1*, 6(1):1–25.

[108] Haldane, A. (2009). Rethinking the Financial Network, Speech by Andrew Haldane, Executive Director, Financial Stability. *Bank of England Speech Repository*, pages 1–41.

[109] Haldane, A. G. and May, R. M. (2011). Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355.

[110] Hameed, A., Morck, R., Shen, J., and Yeung, B. (2015). Information, analysts, and stock return comovement. *The Review of Financial Studies*, 28(11):3153–3187.

[111] Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

[112] Hamilton, W. L. (2020). Graph Representation Learning Hamilton. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

[113] Hamilton, W. L., Ying, R., and Leskovec, J. (2018). Representation learning on graphs: Methods and applications.

[114] Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. *ArXiv*, abs/2503.04761.

[115] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

[116] Hassan, T. A., Hollander, S., Kalyani, A., van Lent, L., Schwedeler, M., and Tahoun, A. (2024). Economic surveillance using corporate text. NBER Working Paper 33158, National Bureau of Economic Research.

[117] Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

[118] Heiberger, R. H. (2014). Stock network stability in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 393:376–381.

[119] Hoberg, G. and Phillips, G. M. (2018). Text-based industry momentum. *Journal of Financial and Quantitative Analysis*, 53(6):2355–2388.

[120] Hou, K. (2007). Industry information diffusion and the lead-lag effect in stock returns. *The review of financial studies*, 20(4):1113–1138.

[121] Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

[122] Inoue, H. and Todo, Y. (2019). Firm-level propagation of shocks through supply-chain networks. *Nature Sustainability*, pages 1–7.

[123] Iori, G., De Masi, G., Precup, O. V., Gabbi, G., and Caldarelli, G. (2008). A network analysis of the italian overnight money market. *Journal of Economic Dynamics and Control*, 32(1):259–278.

[124] Israelsen, R. D. (2016). Does common analyst coverage explain excess comovement? *Journal of Financial and Quantitative Analysis*, 51(4):1193–1229.

[125] Jeude, J. A. v. L. d., Aste, T., and Caldarelli, G. (2019). The multilayer structure of corporate networks. *New Journal of Physics*, 21(2):025002.

[126] Jiang, D., Peng, X., and Zhu, H. (2023). Shared analyst coverage and connected-firm momentum spillover in china. *Systems Engineering Theory & Practice, Forthcoming*.

[127] Juola, P. (2013). Using the google n-gram corpus to measure cultural complexity. *Literary and linguistic computing*, 28(4):668–675.

[128] Kaminski, J. (2014). Nowcasting the bitcoin market with Twitter signals. *ArXiv preprint arXiv:1406.7577*.

[129] Kardulias, P. N. (1989). The Collapse of Complex Societies . By Joseph A. Tainter. . *American Journal of Archaeology*, 93(4):599–601.

[130] Kellermeier, T., Repke, T., and Krestel, R. (2020). Mining business relationships from stocks and news. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11985 LNAI:70–84.

[131] Kelly, B., Xiu, D., et al. (2023). Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4):205–363.

[132] Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. (2014). Review: Text mining for market prediction: A systematic review. *Expert Syst. Appl.*, 41(16):7653–7670.

[133] Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

[134] Kipf, T. N. and Welling, M. (2016). Variational Graph Auto-Encoders. *ArXiv*, 1611.07308v1.

[135] Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

[136] Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression. In *Proceedings of NAACL*, pages 272–280.

[137] Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., and Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, 117(12):7673–7761.

[138] Kukreti, V., Pharasi, H. K., Gupta, P., and Kumar, S. (2020). A Perspective on Correlation-Based Financial Networks and Entropy Measures. *Frontiers in Physics*, 8:323.

[139] Kumar, B. S. and Ravi, V. (2016). A Survey of the Applications of Text Mining in the Financial Domain. *Knowledge-Based Systems*, 114:128–147.

[140] Laloux, L., Cizeau, P., Potters, M., and Bouchaud, J.-P. (2000). Random Matrix Theory and Financial Correlations. *International Journal of Theoretical and Applied Finance*, 03(03):391–397.

[141] Lange, K.-R., Reccius, M., Schmidt, T., Müller, H., Roos, M. W. M., and Jentsch, C. (2022). Towards extracting collective economic narratives from texts. *Ruhr Economic Papers*.

[142] Lee, C. M. and So, E. C. (2017). Uncovering expected returns: Information in analyst coverage proxies. *Journal of financial economics*, 124(2):331–348.

[143] Lee, C. M., Sun, S. T., Wang, R., and Zhang, R. (2019). Technological links and predictable returns. *Journal of Financial Economics*, 132(3):76–96.

[144] Lewis, C. and Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5):587–615.

[145] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

[146] Li, W., Bao, R., Harimoto, K., Chen, D., Xu, J., and Su, Q. (2020). Modeling the stock relation with graph network for overnight stock movement prediction. *IJCAI International Joint Conference on Artificial Intelligence*, 2021-Janua:4541–4547.

[147] Li, Y., Wang, S., Ding, H., and Chen, H. (2023). Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA. Association for Computing Machinery.

[148] Lim, B., Zohren, S., and Roberts, S. (2020). Detecting Changes in Asset Co-Movement Using the Autoencoder Reconstruction Ratio. *SSRN Electronic Journal*.

[149] Lim, B., Zohren, S., and Roberts, S. J. (2019). Enhancing time-series momentum strategies using deep neural networks. In *The Journal of Financial Data Science*.

[150] Lin, H. and Sun, Q. (2021). Financial volatility forecasting: A sparse multi-head attention neural network. *Information (Switzerland)*, 12(10):419.

[151] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

[152] Lo, A. (2004). The adaptive market hypothesis: market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5):15–29.

[153] Loretan, M. and English, W. B. (2005). Evaluating Correlation Breakdowns During Periods of Market Volatility. *SSRN Electronic Journal*.

[154] Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

[155] Loughran, T. and McDonald, B. (2020). Textual Analysis in Finance. *Annual Review of Financial Economics*, 12:357–375.

[156] Ma, X., Zhao, T., Guo, Q., Li, X., and Zhang, C. (2022). Fuzzy hypergraph network for recommending top-k profitable stocks. *Information Sciences*, 613:239–255.

[157] Mahdavi, S., Khoshraftar, S., and An, A. (2020). Dynamic joint variational graph autoencoders. *Communications in Computer and Information Science*, 1167 CCIS:385–401.

[158] Mantegna, R. N. (1999). Information and hierarchical structure in financial markets. *Computer Physics Communications*, 121:153–156.

[159] Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

[160] Martens, T. and Sextroh, C. J. (2021). Analyst coverage overlaps and interfirm information spillovers. *Journal of Accounting Research*, 59(4):1425–1480.

[161] Marti, G., Nielsen, F., Bińkowski, M., and Donnat, P. (2021). A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *Progress in information geometry: Theory and applications*, pages 245–274.

[162] Massara, G. P. and Aste, T. (2019). Learning clique forests. *ArXiv*, abs/1905.02266.

[163] Massara, G. P., di Matteo, T., and Aste, T. (2015). Network filtering for big data: Triangulated maximally filtered graph. *J. Complex Networks*, 5:161–178.

[164] Matsunaga, D., Suzumura, T., and Takahashi, T. (2019). Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis. ** Equal contribution. NeurIPS*.

[165] Mellace, S., Vani, K., and Antonucci, A. (2020). Relation clustering in narrative knowledge graphs. *CEUR Workshop Proceedings*, 2794:23–27.

[166] Menzly, L. and Ozbas, O. (2010). Market segmentation and cross-predictability of returns. *The Journal of Finance*, 65(4):1555–1580.

[167] Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). Model-agnostic interpretability with shapley values. In *2019 10th international conference on information, intelligence, systems and applications (IISA)*, pages 1–7. IEEE.

[168] Mezei, J. and Sarlin, P. (2018). RiskRank: Measuring interconnected risk. *Economic Modelling*, 68:41–50.

[169] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

[170] Mittermayer, M.-A. (2004). Forecasting Intraday Stock Price Trends with Text Mining Techniques. In *Proceedings of the Proceedings of the 37th Annual Hawaii*

International Conference on System Sciences (HICSS'04) - Track 3 - Volume 3, HICSS '04, page 30064.2, USA. IEEE Computer Society.

[171] Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *ArXiv preprint arXiv:1902.06673*.

[172] Moskowitz, T. J. and Grinblatt, M. (1999). Do industries explain momentum? *The Journal of finance*, 54(4):1249–1290.

[173] Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: Massive Text Embedding Benchmark. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2006–2029.

[174] Münnix, M. C., Shimada, T., Schäfer, R., Leyvraz, F., Seligman, T. H., Guhr, T., and Stanley, H. E. (2012). Identifying states of a financial market. *Scientific Reports*, 2:1–9.

[175] Nigam, A. and Caswell, N. S. (2003). Business artifacts: An approach to operational specification. *IBM Systems Journal*, 42(3):428–445.

[176] Noureldin, D., Shephard, N., and Sheppard, K. (2012). Multivariate high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics*, 27(6):907–933.

[177] Nyman, R., Kapadia, S., and Tuckett, D. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127:104119.

[178] O'Halloran, S., Maskey, S., McAllister, G., Park, D. K., and Chen, K. (2015). Big data and the regulation of financial markets. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, pages 1118–1124.

[179] Ong, H. and Shao, J. (2014). De-anonymising set-generalised transactions based on semantic relationships. In *International Conference on Future Data and Security Engineering*, pages 107–121. Springer.

[180] Onnela, J. P., Kaski, K., and Kertész, J. (2004). Clustering and information in correlation based financial networks. *European Physical Journal B*, 38(2):353–362.

[181] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji,

S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2023).

Gpt-4 technical report.

[182] Oyeniyi, T., Yang, Z., and Tortoriello, R. (2020). The Analyst Matrix: Profiting from Sell-Side Analysts' Coverage Networks. *S&P Global Market Intelligence*.

[183] Özgür, A., Cetin, B., and Bingol, H. (2008). Co-occurrence network of reuters news. *International Journal of Modern Physics C*, 19(5):689–702.

[184] Pacreau, G., Lezmi, E., and Xu, J. (2021). Graph neural networks for asset management. *Available at SSRN 3976168*.

[185] Page, S. and Panariello, R. A. (2018). When diversification fails. *Financial Analysts Journal*, 74(3):19–32.

[186] Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE.

[187] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

[188] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

[189] Park, J., Lee, M., Chang, H. J., Lee, K., and Choi, J. Y. (2019). Symmetric graph convolutional autoencoder for unsupervised graph representation learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:6518–6527.

[190] Parsons, C. A., Sabbatucci, R., and Titman, S. (2020). Geographic lead-lag effects. *The Review of Financial Studies*, 33(10):4721–4770.

[191] Peachey, J. and Webb, H. (2019). NLP-Driven Interpretation of Financial Regulations and Legal Texts. *arXiv preprint arXiv:1903.11474*.

[192] Pharasi, H. K., Sharma, K., Chatterjee, R., Chakraborti, A., Leyvraz, F., and Seligman, T. H. (2018). Identifying long-term precursors of financial market crashes using correlation patterns. *New Journal of Physics*, 20(10):103041.

[193] Piškorec, M., Antulov-Fantulin, N., Novak, P. K., Mozetič, I., Grčar, M., Vodenska, I., and Šmuc, T. (2014). Cohesiveness in financial news and its relation to market volatility. *Scientific reports*, 4(1):1–8.

[194] Poterba, J. M. and Summers, L. H. (1988). Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics*, 22(1):27–59.

[195] Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D., and Ben-Jacob, E. (2012). Quantifying the behavior of stock correlations under market stress. *Scientific Reports*, 2(1):1–5.

[196] Rapp, R. (2022). Using semantic role labeling to improve neural machine translation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3079–3083, Marseille, France. European Language Resources Association.

[197] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

[198] Roberts, S., Dong, X., Zohren, S., et al. (2023). Network momentum across asset classes. *ArXiv preprint arXiv:2308.11294*.

[199] Rosenfeld, N., Szanto, A., and Parkes, D. C. (2020). A kernel of truth: Determining rumor veracity on Twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*, pages 1018–1028.

[200] Sablayrolles, A., Jiang, A. Q., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B.

[201] Saha, S., Gao, J., and Gerlach, R. (2022). A survey of the application of graph-based approaches in stock market analysis and prediction. *International Journal of Data Science and Analytics*, 14(1):1–15.

[202] Salton, G. (1986). Recent trends in automatic information retrieval. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86, page 1–10, New York, NY, USA. Association for Computing Machinery.

[203] Samal, A., Pharasi, H. K., Ramaia, S. J., Kannan, H., Saucan, E., Jost, J., and Chakraborti, A. (2021). Network geometry and market instability. *Royal Society Open Science*, 8(2).

[204] Sandhu, R. S., Georgiou, T. T., and Tannenbaum, A. R. (2016). Ricci curvature: An economic indicator for market fragility and systemic risk. *Science Advances*, 2(5).

[205] Sawhney, R., Agarwal, S., Mittal, V., Rosso, P., Nanda, V., and Chava, S. (2022). Cryptocurrency Bubble Detection: A New Stock Market Dataset, Financial Task & Hyperbolic Models. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5531–5545.

[206] Schou, P. K., Bucher, E., Waldkirch, M., and Grünwald, E. (2022). We did start the fire: r/wallstreetbets,'flash movements' and the gamestop short-squeeze. In *Academy of Management Proceedings*, number 1, page 14028.

[207] Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2).

[208] Semenova, V. and Winkler, J. (2021). Social contagion and asset prices: Reddit's self-organised bull runs. *arXiv preprint arXiv:2111.13519*.

[209] Serur, J. A. and Avellaneda, M. (2021). Hierarchical PCA and Modeling Asset Correlations. *SSRN Electronic Journal*.

[210] Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., and Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, 3(2):279–317.

[211] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

[212] Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

[213] Sidorov, S. P., Faizliev, A. R., Balash, V. A., Gudkov, A. A., Chekmareva, A. Z., and Anikin, P. K. (2018). Company co-mention network analysis. *Springer Proceedings in Mathematics and Statistics*, 247:341–354.

[214] Sohn, K.-W., Kwon, S., and Choi, J. (2021). The global banking standards qa dataset (gbs-qa). *Proceedings of the Third Workshop on Economics and Natural Language Processing*.

[215] Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2020). Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

[216] Souza, T. T. and Aste, T. (2019). Predicting future stock market structure by combining social and financial network information. *Physica A: Statistical Mechanics and its Applications*, 535.

[217] Stammbach, D., Webersinke, N., Bingler, J. A., Kraus, M., and Leippold, M. (2022). Environmental Claim Detection. In *Annual Meeting of the Association for Computational Linguistics*.

[218] Strange, S. (1996). *The Retreat of the State: The Diffusion of Power in the World Economy*. Cambridge Studies in International Relations. Cambridge University Press.

[219] Sul, H. K., Dennis, A. R., and Yuan, L. (2017). Trading on Twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3):454–488.

[220] Tetlock, P. C. (2005). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

[221] Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American economic review*, 106(7):1577–1600.

[222] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine 2023 29:8*, 29(8):1930–1940.

[223] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Michael, E., Ranjan, S., Xiaoqing, S., Tan, E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models.

[224] Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

[225] Tumminello, M., Aste, T., di Matteo, T., and Mantegna, R. N. (2005a). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102 30:10421–6.

[226] Tumminello, M., Aste, T., Matteo, T. D., and Mantegna, R. N. (2005b). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426.

[227] Tumminello, M., Lillo, F., and Mantegna, R. N. (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior and Organization*, 75(1):40–58.

[228] Turner, E. (2021). Graph auto-encoders for financial clustering. *arXiv preprint arXiv:2111.13519*.

[229] Upper, C. (2011). Simulation methods to assess the danger of contagion in interbank markets. *Journal of Financial Stability*, 7(3):111–125.

[230] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[231] Vitali, S., Glattfelder, J. B., and Battiston, S. (2011). The network of global corporate control. *PLoS ONE*, 6.

[232] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science (New York, NY)*, 359(6380):1146–1151.

[233] Wan, X., Yang, J., Marinov, S., Calliess, J. P., Zohren, S., and Dong, X. (2021). Sentiment correlation in financial news networks and associated market movements. *Scientific Reports*, 11(1):1–12.

[234] Wang, C. and Luo, B. (2021). Predicting $ gme stock price movement using sentiment from reddit r/wallstreetbets. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 22–30.

[235] Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y., Yu, Q., Zhou, J., Yang, S., and Qi, Y. (2019). A semi-supervised graph attentive network for financial fraud detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 598–607. IEEE.

[236] Wang, G. J., Xie, C., Han, F., and Sun, B. (2012). Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence

from minimal spanning tree. *Physica A: Statistical Mechanics and its Applications*, 391(16):4136–4146.

[237] Wang, J., Zhang, S., Xiao, Y., and Song, R. (2022a). A Review on Graph Neural Network Methods in Financial Applications. *Journal of Data Science*, 20(2):111–134.

[238] Wang, K., Thakur, N., Reimers, N., and Gurevych, I. (2022b). GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

[239] Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M. (2021). ClimateBert: A Pretrained Language Model for Climate-Related Text. *ArXiv*, abs/2110.12010.

[240] Witts, D. W., Tortosa-Ausina, E., and Arribas, I. (2021). The irrational market: Considering the effect of the online community wall street bets on financial market variables.

[241] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). Bloomberggpt: A large language model for finance. *ArXiv*, abs/2303.17564.

[242] Xiang, W. and Wang, B. (2019). A Survey of Event Extraction From Text. *IEEE Access*, 7:173111–173137.

[243] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. (2024). C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

[244] Xue, J., Jiang, N., Liang, S., Pang, Q., Yabe, T., Ukkusuri, S. V., and Ma, J. (2022). Quantifying the spatial homogeneity of urban road networks via graph neural networks. *Nature Machine Intelligence*, 4(3):246–257.

[245] Yi, B. and Guo, S. (2022). Common analyst links and predictable returns: Evidence from china. *The North American Journal of Economics and Finance*, 63:101832.

[246] Yi, B. and Xiang, X. (2023). Pair analyst coverage and return comovement: Evidence from china. *Pacific-Basin Finance Journal*, 77:101908.

[247] Zhang, C., Bank, D., Song, C., Agarwal DBS Bank, S., Wu DBS Bank, H., Zhang DBS Bank, X., Jianan Lu DBS Bank, J., Agarwal, S., Wu, H., Zhang, X., and Jianan Lu, J. (2023). A Semantic Search Framework for Similar Audit Issue Recommendation in Financial Industry. *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 1208–1211.

[248] Zhang, S., Yin, H., Chen, T., Hung, Q. V. N., Huang, Z., and Cui, L. (2020a). Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 689–698.

[249] Zhang, W., Zhuang, X., Wang, J., and Lu, Y. (2020b). Connectedness and systemic risk spillovers analysis of Chinese sectors based on tail risk network. *North American Journal of Economics and Finance*, 54:101248.

[250] Zhao, L., Li, L., Zheng, X., and Zhang, J. (2021). A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. *Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021*, pages 1233–1238.

[251] Zhou, H., Li, M., Xiao, Y., Yang, H., and Zhang, R. (2024). Leap: Llm instruction-example adaptive prompting framework for biomedical relation extraction. *Journal of the American Medical Informatics Association*, 31(9):2010–2018.

[252] Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In Li, S., Sun, M., Liu, Y., Wu, H., Liu, K., Che, W., He, S., and Rao, G., editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.