University of Technology Sydney

Information and Sentiment in Financial Markets

Qi Zhang

Thesis submitted to Finance Discipline Group of the University of Technology Sydney for the degree of

Doctor of Philosophy

July 2024

Declaration

I certify that the thesis I have presented for examination for the PhD degree of UTS is solely my own work other than where I have clearly indicated that it is the work of others.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 52880 words.

Statement of conjoint work

I confirm that Chapter 2 is jointly co-authored with Jianxin Wang and Minxian Yang.

Signature: Production Note: Signature removed prior to publication.

Date: July 2024

Acknowledgements

I am deeply grateful to my supervisors Jianxin Wang, Minxian Yang, Wei Liu, Kathy Walsh, and Vinay Patel. Their feedback has been exceptionally insightful, encouraging, and helpful. Jianxin has been exceptionally supportive throughout my PhD study. Without him we would not even consider exploring a topic neither of us were familiar with, and this thesis would not be possible. We are especially grateful to Minxian, who provided critical support to the thesis despite his health conditions, and grieved by his passing. His enthusiasm to and solid knowledge of econometrics is a critical part of the first essay.

I received valuable feedback from UTS presentations and conference participants, including but not limited to Vitali Alekseev, Yushan Tang, Sheung Chi Chow, Jared DeLisle, Cynthia Cai, Kristoffer Glover, Ivan Medovikov, and an anonymous examiner. Their comments, questions, and suggestions are priceless, and are what shape this thesis.

I am grateful for scholarships and funds from UTS International Research Scholarship, UTS President’s Scholarship, UTS Vice Chancellor's Conference Fund, and UTS Business School funds.

Lastly, I am indebted to my parents for their love, support, and encouragement throughout my life.

Abstract

In this thesis, I revisit information and sentiment in financial markets using state-of-the arts technologies from econometrics and machine learning.

The first paper introduces a model that estimates the common and market-specific information flows when an asset is traded in multiple markets. The model does not require ultra-high frequency data and avoids the '*who moves first'* interpretation of price discovery. We show that at sampling intervals from 0.01 to 2 seconds, the common information across all markets accounts for 67% to 94% of the total information flow, and the listing exchange accounts for half of the remaining information. The common information between quotes and trade prices ranges from 58% to 85% of the total information flow, with the remaining information coming mostly from quotes. Trade prices have very little information.

The second paper measure news sentiment using BERT and explores return predictability based on a new database, Refinitiv Machine Readable News (MRN). The resulting portfolio achieves an annualized Sharpe ratio of 3.96, significantly higher than that of a passive investment (as proxied by S&P 500 index) and dictionary method, which achieves a Sharpe ratio of 0.32 and 2.94 respectively. We find that dictionary methods struggle to extract information from complicated texts compared with BERT. An interesting finding is that news of positive sentiment is tailored to fewer audiences, contain fewer topics, and are generally shorter. We further show that seasonalities and holiday effects do not appear to explain sentiment portfolio returns while article complexity does.

The third paper uses XGBoost, a powerful and state-of-the-art machine learning algorithm, to predict next-day volatility jumps and then form portfolio for the next day. We show that controlling for news sentiment and volatility measures and using over 1,400 news topics, next-day RSJ, a recent development of volatility jump measure, is reasonably predictable, and using the predicted RSJ to form one-day portfolio achieves outstanding portfolio performance where annualized Sharpe ratio is

2.06 with only 44 stocks available[1]. The performance is significantly higher than portfolio selection based on today's RSJ (0.619) or today's news sentiment (1.94). The improved Sharpe ratio is mainly from higher return. The portfolio earns a significant alpha relative to Fama-French 5-factor model benchmark. We show that investor attention and media coverage potentially explain the portfolio return.

[1] This effectively suggests a risky position because of the very small number of stocks available. See the paper for details. This, however, probably suggests that our model's performance, and hence the portfolio performance, would be even higher with larger dataset which makes effective hedging possible.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Sentiment and information are two critical aspects of finance. Traditional finance follows from information-based theory which assumes investors are rational, profit-maximizing individuals who have, and only have, their own well-being in mind. More recently, studies in psychology and sociology suggest that this is far from the case. Finance academics have been borrowing ideas from these advances in sociology, psychology, and more recently, computer science, to allow us better to understand, explain, and predict financial markets.

In the first paper, I revisit the classical price discovery metric as advanced by Hasbrouck (1995). This classical view of price discovery states that, when one asset is traded at different venues, whoever is the first to move must be the price leader. In other words, price discovery takes place at whichever market that is the first to react. However, different trading venues often react to the same set of information that is available to everyone, and attributing price discovery simply based on speed of reaction is a very coarse interpretation of the price discovery process. Instead, we ask: accounting for the public information, where does private information really come from? This residual information would help us shape our view of price discovery in the real world.

Interestingly, while sentiment analysis sounds like a recent development, first attempts to extract sentiment information from textual data can be dated back to at least the 1930s, when Cowels (1933) tried to subjectively give sentiment labels to Wall Street Journal front page articles and use the self-labelled sentiment to explain observed market conditions. Data and methodology limitations severely restricted his approach's performance, but his attempt was legendary considering how early it was. Today, we are equipped with large amounts of data, and advances in computer science allows us to extract sentiment information easily and accurately. In my second and third paper, we attempt to extract sentiment and information from news by using Transformers and XGBoost, two state-of-the-art machine learning algorithms, to predict sentiment and volatility jumps, respectively. We show that such practices are highly profitable where trading strategies based on this information would give

high risk-return tradeoff in terms of Sharpe ratio. We show that the portfolios' profits are unexplained by common risk factors in Fama-French Five Factor model.

While the first study utilizes traditional numeric data from conventional high-frequency data from Thomson Reuters, the second and third study utilize textual data from Refinitiv Machine Readable News (MRN) database. We hope this thesis would be beneficial to both financial academics and practitioners who seek to understand how modern technology could be used in conjunction with, or replace, traditional methods, and how they compare to the traditional methods.

We now give a more detailed but non-technical look at the three papers in this thesis.

1.1 Chapter 2: Common and market-specific information flows

Market segmentation is now a common feature of financial markets. increasingly, we see one asset trading at multiple venues across multiple trading exchanges within a country and cross-border. A natural question is then where information comes from. Traditional neoclassical economics states that it is the interaction between demand and supply that determines the equilibrium (or 'fair') price of the asset. In market microstructure literature, we attempt to understand how the market arrives at the market price. At a high level, investors interpret new information as they arrive at the market and incorporate that piece of information into asset prices. Depending on how efficient the market is (or what we believe the market is), market price may reflect all or partial information available today.

Under the classical view of price discovery after Hasbrouck (1991; 1995), we believe that information must come from somewhere, and only somewhere. In other words, there must be one market (or trading venue) that first incorporates the new information after asset prices, and we call this player the price leader. Whoever follows are simply following what the price leader is doing. Under this classical 'who moves first' view of price discovery, the critical job is simply to find who is the first to react when there is a price change. To do so, we reply on high-frequency data which allows us to identify the first mover.

However, some information is available to all. Such information is observed by everyone and is hence reacted to by everyone. By saying whoever is the first to react must be the price leader, and hence is more informative, is therefore not a perfect way of interpreting price discovery. We note that such *common information* does not necessarily equal to public information. Admittedly, a large part of common information is public information, such as news. However, some private information may also make a common knowledge across all *markets*. For example, an insider with private information may wish to exploit this piece of information. However, it is common practice that he or she tries to hide his knowledge of private information and not to submit large orders. An alternative way is to cut the orders into smaller pieces and send to all markets. In this case, the private information is effectively a common knowledge known to all market because the trader is cross-market.

In the first paper, we acknowledge the existence of common knowledge and hold that when controlling for the common knowledge that everyone reacts to, the residual part is the information carried by each market. We find that vast majority of information is indeed common information, and the difference between different trading venues is either economically smaller or statistically insignificant.

1.2 Chapter 3: Sentiment analysis with financial news

Having established that common information is the major source of information in financial markets, we now look at another popular topic in finance: sentiment analysis. As news is a major source of public information, understanding how news affects investor sentiment is important for us to understand the market. We do have a large literature on investor sentiment. In the early days, almost all sentiment analysis studies were based on numeric data. In these studies, researchers generally try to use numeric data (such as abnormal return, trading volume, etc.) to approximate investor sentiment. Large research institutions and government agencies also rely on expensive surveys to directly assess people's sentiment about market conditions. However, numeric-based methods are rough and indirect measures, and surveys are extremely expensive. In recent years, with advances in computer science

(specifically, opinion mining methods), which allows timely, efficient, and accurate extraction of sentiment information from textual data, we see a surge in financial sentiment analysis.

Finance academics tend to use what is available and prevalent method in our research toolbox and have been relying on so-called dictionary-based methods. This method essentially reduces textual sentiment analysis to word count of positive and negative words where the dictionaries are pre-specified. In this paper, we use Transformer-based model, which is (unarguably) the most important Natural Language Processing (NLP) architecture during the past decade. Essentially, this architecture asks each word in the document to look at words the document contains, noting the structure, interconnections of each word, words closer to and farther away from the word, et cetera. By doing so, it incorporates the structure, grammar, and contextual information in the text, and is a much more accurate representation of textual information.

Other than methodology, data availability has long been an issue in sentiment analysis. Because of unavailability of large volume financial news, we have been relying on alternative textual sources (such as twitter), some short versions of news (such as headlines only), or some rough estimate of news arrivals (such as the number of news). These methods are indirect estimates of news itself. In the second paper, we overcome this issue by using full news from Refinitiv Machine Readable News database, which contains historical news from Refinitiv (formerly known as Thomson Reuters). Having real-time news also allows us to directly observe news characteristics and is in itself valuable.

We show that the new method is especially good at extracting information from complex texts compared with the traditional dictionary-based methods for financial textual analysis. Also, there appears to be news heterogeneity: news of positive sentiment is tailored to fewer audiences, contain fewer topics, and are generally shorter. This is interesting as it suggests that merely the way the news is presented in appear to carry valuable information, at least for sentiment contents.

1.3 Chapter 4: News and volatility jumps

Among the different channels through which news affect the market, news trigger jumps in price and volatility dynamics. Traditional finance models have been relying on continuous models; when they use discrete models, it is usually a simplified version of the continuous model and in the limit, they approach the continuous version of the model. This is not surprising: these continuous models have been working well, and it usually takes a lot of effort, often through complicated stochastic processes, to merely identify the jumps.

In recent years, with advances in econometrics, there have been more and more studies on news (and generally, event)- triggered jumps in financial markets, and what implications they carry. Most are based on price jumps, one reason is that volatility is unobserved, hence identifying volatility jump is more tedious. Most if not all existing studies are based on macroeconomic news and firms' earnings announcement. This is not surprising: researchers typically have to manually collect these data, and the limited data size is manageable. However, any news announcement should cause jumps and manual data collection is unreasonable. With a recent development in econometrics, which simply uses realized semi-variances to approximate a relative volatility jump component, we can estimate volatility jump for each stock-day in a simple way.

Using this volatility jump measure, we show that news predicts next-day volatility jumps well. We rely on XGBoost, a machine learning algorithm, to capture the complicated inter-reactions and non-linear relationships among the abstracts of news. Using the predicted volatility jump to form portfolio is highly profitable.

We hope this thesis would shed light on future academic research and practitioners' practices. Nevertheless, we acknowledge that this study, and this strand of literature, is far from complete, and we wish we could enlighten future research in sentiment analysis, transfer learning of machine learning methods in finance, and asset pricing in general.

Chapter 2 Common and Market-Specific Information in Financial Markets

2.1 Introduction

In financial markets, we frequently see the same, or closely related assets, traded at different venues and the venues may or may not be from the same country: foreign currencies traded around the world, the same stock traded at NYSE and NASDAQ, or traded in different countries' exchanges, et cetera. Then a natural question is which market incorporates new information about the underlying asset faster and better, and how price discovery's efficacy depends on trading mechanisms, market liquidity, and the prevalence of asymmetric information (Yan and Zivot, 2010).

A popular measure of price discovery in multiple markets is Hasbrouck (1995) Information Share (IS), which defines a market's information share as the proportion of the efficient price innovation variance attributable to that market (Yan and Zivot, 2010). It propagates the '*who moves first*' view in incorporating new information and a market is assigned higher IS if it adjusts its price faster than another market. Hasbrouck (1995) further emphasizes that IS does not tell us the amount of information incorporated into prices and does not mean the market with highest information share necessarily has best price (lowest ask-bid spread). Putniņš (2013) notes that most price discovery studies either follow Hasbrouk's (1995) 'who moves first' view or do not explicitly define price discovery, and that being first to impound new information does not mean that market's price is more informative, as that market's price may be nosier hence less useful in assessing the asset's fundamental value. The advances in price discovery are generally limited over the years. In the meanwhile, we move to high-frequency, automated trading and calculating information share is becoming more and more resource-consuming because of the large number of parameters to estimate.

To estimate Hasbrouck IS, we need ultra-high frequency data where only one market moves at a time, such that we can uniquely assign the source of price discovery to one source.

Hasbrouck (2019) studies information share of IBM and NVidia in October 2016, and he uses dataset of 10^{-5} second dataset. The huge number of parameters (16 million coefficients using 10 seconds worth of lags) makes it practically impossible to estimate. Hasbrouck uses Heterogenous Autoregressive Model (HAR) to reduce the number of parameters, but the methodology still follows Hasbrouck (1995).

In this paper, we argue that different markets' prices have a common component due to common, public information, and a market-specific component due to market-specific information, and we investigate information content of public and private information. The rationale is that common information is made available to all market participants, and one market's seemingly high information content may be dominated by public information; it is each market's private information that is relevant to traders and policy makers. Public information includes everything everyone sees at the same time: macroeconomic and corporate news, social media, intermarket sweep orders, et cetera. Studying common component in price behavior is not new. For example, Hasbrouck and Seppi (2001) investigate common factors in price discovery process in equity markets. Starting with a model where stock returns are driven by signed order flows and public news, they find that stock returns and order flows contain substantial common factors. However, common factors have not been explicitly considered as a contributing factor in price discovery measures. The technique we use in this paper, decomposing variance into a common and a market-specific component, is also not new. See, for example, Altonji and Ham (1990) on Canadian employment; Stock and Watson (2005) on international business cycles, and Yang (2016) on U.S. macroeconomic time series. However, to our knowledge, it has not been used in market microstructure.

We argue that it is inappropriate to use ultra-high frequency data (1s in Hasbrouck, 1995; and $10^{-5}s$ in Hasbrouck, 2019) to investigate the informational contents of trades and quotes. At coarse resolutions, common component dominates; at finer resolutions, order-related information dominates; at ultra-fine resolutions, little or no information dominate, and race

for speed by market makers and other automated trades reduce informational contents at such reisolutions.

Historically, the literature tends to study information content of NYSE due to its large market share. For example, Deb, McInish, Shoesmith, and Wood (1995) find that NYSE does not respond much to other exchanges' prices; Hasbrouck (1995) finds that NYSE's IS was as high as 91.3%. In recent years, NYSE's contribution to price discovery drops as its market share drops. Ozturk, Van der Wel, and van Dijk (2017) find that considering listing exchange against non-listing exchanges, NYSE's (NASDAQ's) trades and volumes are only 30.7% and 27.5% (40.9% and 42.7%) respectively, while the information share of NYSE (NASDAQ) is 49.7% (39.7%) when allowing intraday variation and 61.9% (35.3%) when not allowing intraday variation. Therefore, although NYSE's IS has dropped, it is still large especially compared with its trades and volumes.

While NYSE specialist was measured in weeks (Hasbrouck and Sofianos, 1993), now it is in seconds, even milliseconds, and information in this high-frequency world are increasingly order-related (O'Hara, 2015). The effect of high frequency trading in financial market is far beyond higher speed and has changed the financial market. For example, other than the active side of trade, passive side limit order books also contain information. Today, order latencies (time taken for order to hit endpoint) are measured in milliseconds, and in some cases nanoseconds (O'Hara, 2015; Dungey and Mckenzie 2013). What is worse is that what is 'information' is unclear in the high-frequency world. Now a significant part of information is order based (Hasbrouck and Sofianos, 1993; O'Hara, 2015). That order flow information contains information has been studied in substantial literature. For a recent study, see Fleming, Mizrach, and Nguyen (2018). Specifically, the authors find that trades and limit orders both contribute to price discovery, and traders use limit orders to exploit their private information.

The rest of this paper is organized as follows. Section 2 conducts a literature review, Section 3 details the methodology, Section 4 presents empirical results, and Section 5 concludes with

suggestion for future research.

2.2 Literature Review

This section conducts a literature review of related fields. In section 2.1, we review what 'information' is defined and used in financial markets and price discovery frameworks used in market microstructure today; in section 2.2, we review how high frequency trading adds complications to the traditional price discovery framework; in section 2.3, we give a broad view of how asset pricing is relevant in asset pricing.

2.2.1 Information in financial markets

We have long known that it is the unexpected part in financial events that carries information and that is relevant to market participants. For example, Hasbrouck (1988) notes that it is the unexpected part in trades (trade innovations) that carries information and uses trade histories to assess the unexpected trade component. As price discovery is based on impounding new information into asset prices, a fundamental issue is than 'information'. Over the years, the research in informational economics and finance evolves as we see advances in technology and behavioral finance. Traditionally, numerous studies investigate public news and events' effects on financial markets. Such studies may include macroeconomic news announcements, earnings announcements, analyst forecasts, et cetera. See, for example, Fama, Fisher, Jensen, and Roll (1969) on announcement of stock splits on stock prices, Schwert (1981) on announcement of information on stock prices, Barclay and Litzenberger (1988) on equity issue announcement, and MacKinlay (1997) for a general review of event studies in economics and finance. More recently, researchers investigate social media and their implications for stock returns. See, for example, Ranco, Aleksovski, Caldarelli, Grčar, and Mozetič (2015) and Si, Mukherjee, Liu, Li, Li, and Deng (2013) on stock returns using sentiment analysis on twitter.

In price discovery literature, the definition of 'information' largely relies on Hasbrouck (1991a, b; 1995) framework, which we review below.

2.2.2 Hasbrouck (1991a, b) and Hasbrouck (1995) framework

Hasbrouck (1991a) uses a broad information set to assess the information contents of trade. Indeed, Hasbrouck (1991a, 1991b) set the scene for price discovery literature (especially measuring information content of a trade) that is used till today. The price series used is midquote. Assuming bid and ask prices are set symmetrically about expected value of the underlying security conditional on the information set prevailing at each time, at each trade, subsequent revisions in quotes convey private information prevailing at the time of each trade. Quotes may be revised with or without a trade. Trade and quote innovations hence reflect private and public information, respectively. The permanent impact of a trade innovation on quote reflects the information content. Efficient price,m_t, reflects expected security value conditional on public information. It is however unobservable. Using a vector autoregressive (VAR) model, one can account for serial correlation in trades. Trade innovation's permanent effects on prices can be estimated using impulse response functions (IRF). VAR models are flexible in that it accounts for subsequent revisions to initial over-reaction.

While we revise model set-up formally in methodology section, we note: First assume trades are either due to private information or liquidity needs; trades' impact on stock price may be transient or permeant (reflecting private information). The primary techniques used in Hasbrouck (1991b) are VAR model of Hasbrouck (1991) and random-walk decomposition of non-stationary series into a random-walk component and a stationary component. For the latter strand of literature, see, for example, Campbell and Mankiw (1987) and Stock and Watson (1988).

Any price series (trades or quotes) may be decomposed into an efficient term (expected asset value) and a noise term: $q_t = m_t + u_t$. By assumption, efficient price follows a random walk: $m_t = m_{t-1} + w_t$, where $E(w_t) = 0, var(w_t) = \sigma_w^2, E(w_t w_j) = 0 \; \forall t \neq j$. Further assuming $\{u_t, w_t\}$ are jointly stationary, we also have homoskedasticity. This tends to apply more to event-time than clock time, hence Hasbrouck (1991b) also estimates his models

using event-time sample; as a further remedy, Hasbrouck also reports his estimates based on different time of the day. Change in efficient price has a trade-correlated and a trade-uncorrelated component. The part of efficient price's variance correlated with trade is an absolute measure of information, and its proportion relative to total variance is a relative measure of informativeness. In other words, informativeness is defined as the proportion of variance in w_t that is explained by trade.

Later, market fragmentation became more and more common, and the same asset became more and more commonly listed on multiple markets away from their primary listings (Shapiro, 1993). Assuming the same asset's different prices in multiple markets share the same common efficient price (Garbade and Silber, 1983) and using essentially the same framework, Hasbrouck (1995) investigates multiple market's contribution to efficient price. In his words, '*price discovery in this framework refers to innovations in the efficient price*' (Hasbrouck, 1995). Hasbrouck also notes that by applying this method to multiple markets in one-second intervals, his price discovery metric, information share (IS), essentially measures '*who moves first*'. We however note that, albeit an intuitive interpretation, the econometric technique used is the decomposition of efficient price variance and attributing it to different markets. In this framework, one seeks to uniquely assign variance of efficient price to different venues; to do so, we need high frequency data (one-second in 1995, which is ultra-high frequency) where (hopefully) only one market moves at a time, hence correlations between different markets are close to 0, effectively measuring 'who moves first'.

2.2.3 Hasbrouck (2019)

In 2019, Hasbrouck (2019) discusses price discovery under, in his words, 'high resolution', as the US market is currently timestamped to nanoseconds, while trades are much fewer. Under such high resolution, we can truly identify which market moves first, a task we are unable to achieve under 'coarse' resolutions (such as 1-second). The large number of lags to include in VECM poses computational issues, making it too slow or practically impossible to estimate the model. Hasbrouck follows Corsi (2009) to use a heterogenous autoregressive

(HAR) approach to make the number of parameters manageable. Without HAR, the finest resolution Hasbrouck was studying would entail estimating 16 million coefficients.

The price discovery metric is still Hasbrouck (1995) IS. Hasbrouck (2019) estimates IS of IBM and Nvidia by comparing information share of their listing exchange against the best quote from non-listing exchanges. Using clock time of 1s, 0.1s, 0.01s, …, 0.00001s, Hasbrouck finds that IS of the listing exchange for IBM ranges from 0.136-0.932 at 1-second resolution to 0525-0.526 at 0.00001-second resolution and IS of the best quote from non-listing exchange ranges from 0.068-0.864 at 1-second resolution to 0.474-0.475 at 0.00001-second resolution. The range narrows down as the resolution gets finer.

In this paper, we argue that IS ignores public information and we seek to distinguish common information from market-specific, or idiosyncratic, information. While IS identifies each market's relative importance in incorporating information into prices, common information should have the same impact on all markets, and market-specific information that arises from each market is the relevant part in assessing a market's informativeness. As common information does not arise from each market, but from firms (such as annual reports), news and rumors, governments, and central banks (such as interest rate announcements), unexpected events that arise from the world (such as COVID-19), et cetera, we argue that public information should be considered as a standalone source of information separately from each market.

2.2.4 Information in high-frequency world

In high frequency world, investors still signal their information through trades: good (bad) news are reflected in buy (sell) orders, and the magnitude (where algorithms slice big orders into a series of small orders) signal the size of the news. However, different clienteles who pursue different strategies are attracted to different markets (O'Hara, 2015). Heterogeneity in investor clientele means different sets of private information carried by the different investors are clustered in different markets, giving rise to different informational contents of each market.

Trades and quotes may not be fundamental information related. For example, Zhang (2010) shows that HFT reduces markets' ability to incorporate information on corporate fundamentals into asset prices. To make the situation worse, it is not always the case that trades and quotes are even information-related in high-frequency world. O'Hara (2015) notes that market-making in high-frequency world is akin to statistical arbitrage across assets and markets, and statistical arbitrage by construction is based on statistical models which may or may not be information-based. To have a sense of the pace of automatic trading and order placement, O'Hara notes that as of 2013, '23% of all cancelled orders, and 38% of all cancelled quotes, occur within 50 milliseconds or less of placement'. Therefore, it is inappropriate to use high frequency data to assess markets' informational contents at clock time.

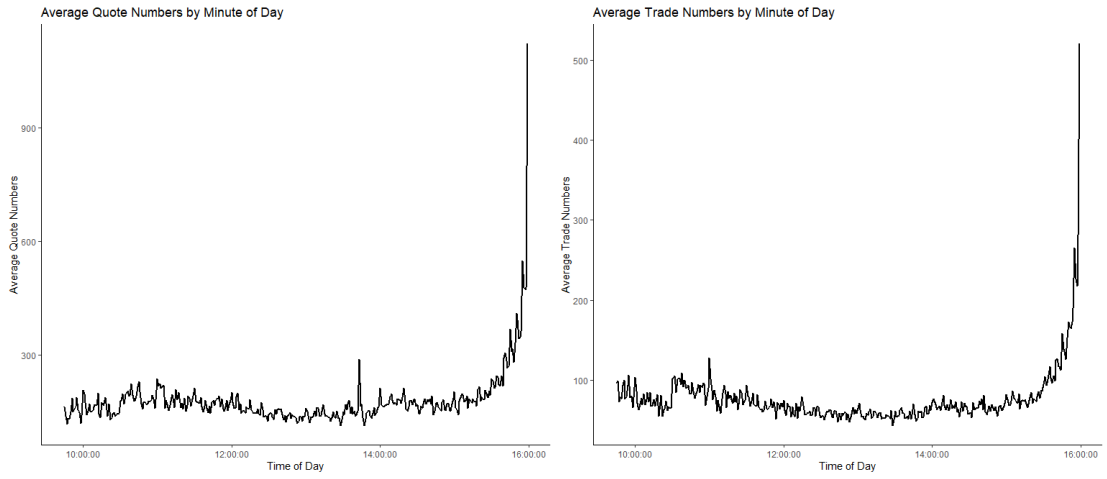


Figure 2-1 Average number of quotes (left) and trades (right) by time of the day

This graph shows the average number of quotes and trades for the 30 trading days from October 3rd, 2016, to November 11th, 2016, by time of the day.

The sparsity in trades and quotes poses another problem. On October 3rd, 2016, from 9:45 to 16:00, there were only 322,506 quotes for IBM, but if we consider clock time of 0.01s, there

would be 2,250,000 observations, which means at least 85.6% of quotes would be stale price; the case is even worse for the even finer resolution considered in Hasbrouck (2019). However, we know that it is new information that matters. Therefore, it is problematic when we estimate our model based on overwhelmingly old information that is not 'news'. This problem would be less severe in event time; hence we also report price discovery measures in event time other than clock time. Event time measures are beneficial especially with irregular trades, and IBM quotes tend to be clustered especially towards the end of the day and end of each hour, as illustrated in Figure 2-1, making event-time measures beneficial.

We note, however, that HFT does not always hinder price discovery. Brogaard, Hendershott, and Riordan (2014) investigate the role of HFT in price discovery. They find that HFT generally facilitates price discovery by trading in the direction of permanent price changes and against transitory pricing errors through marketable orders. Developing on Anderson and Bollerslev (2003) and Cao, Hansch, and Wang (2009), they further show that HFTs are correlated with macroeconomic news announcements and order book imbalances. They therefore provide evidence that in today's HFT world, orders are at least partially based on public information. If each market also trades on their own order imbalance, each market's order flow, which is observed by each market, is a source of market-specific (idiosyncratic) information.

We also note that competing on speed is expensive and economically wasteful. Hasbrouck (2019) follows Hasbrouck (1995) in calculating price discovery metrics and still follows the 'who moves first' view. However, it has been documented that speed should not be taken as information, and that the IS does not always reflect speed in impounding new information. For example, Anand and Subrahmanyam (2008) argue that besides reflecting adjustment for new information, the IS also reflects the magnitude of a market's adjustment, or response, for the new information. It is then questionable if the reaction is rational reflection of information or over-reaction. More recently, Menkveld (2016) provides an excellent review of high frequency trading, with an emphasis on economic interpretations. Among the competing

theories, we note the so-called high frequency trading arms race after Budish, Crampton, and Shim (2015). Under this school, high frequency traders all trade on the same public information, and this creates an expensive race for speed where traders equip themselves with expensive technologies that, hopefully, allow them to be milliseconds faster than their competitors. Budish et al. (2015) show that when a *public news* arrive at the market, high frequency traders enter the market; if the first high frequency trader to arrive is a market maker (HFM), he will update his quote; if he is, in Menkveld's (2016) word, a high frequency bandit (HFB), who trades aggressively to make money by trading against existing quotes, then he will trade against existing quotes and hence imposing adverse selection on HFMs. The HFMs will raise spreads on anticipating such costs. Collectively, high frequency traders involve in a wasteful race for speed, creating a dead weight cost. In multimarket set-ups we study here, such race for speed seems more relevant and more commonplace. Foucault, Kozhan, and Tham (2017) show that such a race for speed creates toxic arbitrage that widens bid-ask spread (Menkveld, 2016). We still observe race for speed among HFTs, and this is not surprising, as Biais, Foucault, and Moinas (2015) show that while socially costly, racing for speed allows one to find the best price in fragmented markets, hence is personally beneficial.

We direct interested readers to Menkveld (2016), but we note here that racing for speed does not imply high informational contents. It is the current market design that makes HFTs compete on speed to trade fast, while entering market early is expensive. Reacting on the same public information, it is the winner of speed who adjusts his price first, instead of winner of the information they carry. Therefore, we argue that one should not look at too fine resolutions such that technology, instead of information, dominates.

To sum up, while Hasbrouck (1991a, b; 1995) framework attempts to uniquely assign price discovery to each individual market, and to do so requires ultra-high frequency data where only one market moves at a time, we argue that it is inappropriate to use such ultra-high frequency data in the first place because trades and quotes at such high frequency may not be

information-based. Instead, we attempt to use ‘coarse’ frequency data where more than one market is allowed to move at a time, and we argue that at coarse frequencies, common component from public information dominates; as frequency gets finer, order-related information dominates; at ultra-high frequency which is used by Hasbrouck and generally price discovery framework in literature, competing on speed instead of information dominates and little if any ‘information’ is contained in trades and quotes at such fine resolution.

2.2.5 Price discovery and asset pricing

Our current discussion may be applied to more general asset pricing literature. While price discovery is important in asset pricing, it has been ignored in traditional asset pricing models. One reason is, in traditional asset pricing models, such as CAPM and the APT, market participants are assumed to carry identical information, or symmetric information among market participants. This view is apparently faulty because it at least ignores noise in the market, and noise should not be considered the same as information. In these classical asset pricing models, noise traders are not an important part, because classical asset pricing models argue that rational arbitragers would look for mistakes irrational traders make and, in the process, drive asset prices to their fundamental values (Friedman and Friedman, 1953; Fama, 1965). Moreover, the irrational traders who lose sufficiently enough money would leave the market. This view has been challenged by numerous academics, and one reason the opposite school presents is that arbitrage is costly (De Long et al., 1990; Pontiff, 1996; Shleifer and Vishney, 1997, among others).

Noise and noise traders are critical in financial markets. In his influential paper, Black (1986) emphasizes that it is uninformed traders’ irrational (noise) trading that makes financial markets function. The idea of noise traders is first studied by Kyle (1985) in market microstructure literature. They act on noise as if it was information. Noise and noise traders then attracted considerable attention in market microstructure academics.

Many studies document market anomalies and challenge the setup in traditional asset pricing

models that prices fully and rationally reflect public information. Although some argue that these anomalies are merely deviations from efficient prices just by chance and hence market is still efficient (Fama, 1998), others argue that some anomalies are too regular and strong to be temporary deviations that just happen by chance (Daniel, Hirshleifer, and Subrahmanyam, 1998). The latter view seems to receive stronger empirical support. For example, De Long, Sheleifer, Summers, and Waldmann (1990) argue that a number of market anomalies, such as stock market's excess volatility and mean reversion, are due to noise trader risk.

We have emphasized that noise is important in financial markets. In this paper, we focus on information in financial markets, as any price impact of noise should only be temporary and not reflected in prices in the long run. Price discovery in asset pricing has been documented in many studies. For example, O'Hara (2003) argues that price discovery is one of the two functions of markets and shows that by ignoring the impacts of market microstructure implications assuming symmetric information, traditional asset pricing models do not work. In reality, market participants carry different information and O'Hara (2003) further shows that allowing asymmetric information and incorporating market microstructure complications implies that investors would have different preferences for bonds and stocks. Specifically, investors with less private information would hold more bonds than risky equity.

2.2.6 Common information

Before we proceed, we take a note on what is public and private information. Macroeconomic and corporate news announcements, rumours, et cetera are made available to all market participants at the same time and as everyone sees the same information, they are common knowledge. However, in high frequency world, market and data providers sell public data to market participants seconds, even milliseconds, earlier than it is available to others (O'Hara, 2015; Easley, O'Hara, and Yang, 2014). When only some people see the information, it is private information, although the private information only lives for seconds (or milliseconds). It is unreasonable to expect such private information holders to exist in only one or some of the markets, we therefore note that the information contribution of each market contains a

short-lived private information component of public information. Common information may also contain the so-called sweep-to-fill orders by. In this case, the market participant places orders across all markets and consume liquidity across markets. They may place sweep orders because they have private information or because they simply demand liquidity.

2.3 Data and Methodology

We first give a detailed documentation of data obtaining, cleaning, and preparation process in this section.

Consistent with Hasbrouck (2019), we use IBM's one month's quotes data for October 3rd to November 11th in 2016. We use two sets of data in this paper. The first set of data is quotes data from WRDS Trade and Quote (TAQ) database, which is used for quotes analysis; and the second set of data is from Refinitiv TRTH, which is used for trades analysis. TRTH data only contains exchange IDs for trades. We report our results for the one day of October 3rd, 2016, and 30 trading days till November 11th, 2016. IBM's primary listing exchange is the NYSE but is listed on eleven (11) exchanges as of October 2016. We note that although Hasbrouck's dataset contains two timestamps, a participant time, and a securities information processor (SIP) time, the dataset we managed to obtain from WRDS TAQ database contains only one timestamp while TRTH data does contain two timestamps. However, we are unable to confirm if the two timestamps from Refinitiv TRTH are participant time and SIP time, because the three timestamps are not the same as each other[2].

Despite the mismatch in timestamp, we cross-check our trades data from Refinitiv TRTH and quotes data from WRDS with Hasbrouck's (2019) Table 5 (which we reproduce in Table 2-1), which reports the number of trades and quotes by exchanges. We note that both TRTH and WRDS data are consistent with Hasbrouck (2019) Table 5. We confine sampling window

[2] They typically differ only by a few milliseconds hence we think they relate to how the vendors process and record the data.

to 9:45 to 16:00 because the opening period is a single-price double-sided auction and is hence not continuous trading, consistent with Hasbrouck (2019) and empirical market microstructure literature (see, for example, Harris and Panchapagesan, 2005; Bessembinder, 2005).

Panel A. Number of IBM Quotes.		
Exchange	Number	Percent
NYSE	83,556	26.6
Bats BZX	38,674	12.3
Bats BYX	37,692	12.0
BX (NASDAQ)	35,116	11.2
IEX	32,125	10.2
NASDAQ	28,383	9.0
NYSE ARCA	20,688	6.6
NASDAQ PSX	13,523	4.3
EDGA	13,284	4.2
EDGX	11,283	3.6

Panel B. Number and Value of IBM trades.				
Exchange	Number of Trades	Percentage	Total Value	Percentage
FINRA ADF	4338	19.5	77,266,592.00	26.8
NYSE	4004	18.0	59,796,122.00	20.8
NASDAQ	4028	18.1	45,466,059.00	15.8
EDGX	1848	8.3	24,483,852.00	8.5
NYSE ARCA	1789	8.0	21,971,478.00	7.6
Bats BZX	2062	9.3	18,703,805.00	6.5
NASDAQ	2250	10.1	16,237,124.00	5.6
Bats BYX	1041	4.7	11,195,595.00	3.9
IEX	428	1.9	7,177,992.00	2.5
EDGA	412	1.8	4,956,413.00	1.7
NASDAQ PSX	82	0.4	853,433.00	0.3

Table 2-1 IBM quotes and trades updates on 20161003 by exchange (descending order)

This table is identical to Hasbrouck (2019) for IBM quotes on 20161003, confirming that we are using the same dataset. This table is in descending order of number (Panel A) and value (Panel B) while Hasbrouck (2019) is in alphabetic order.

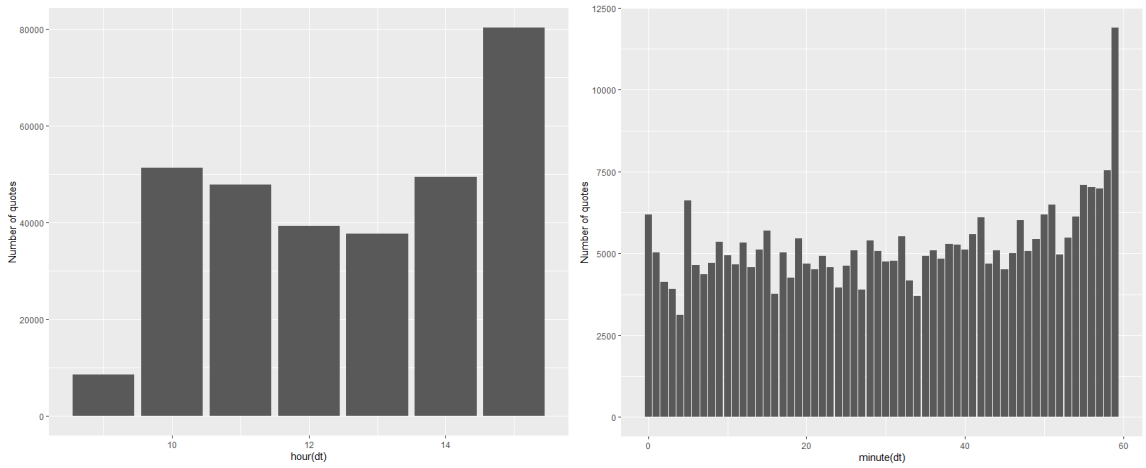


Figure 2-2 Distribution of IBM quotes

This graph shows the average number of total quotes for IBM in October 2016 across each hour (left) and minute (right). Note that the small number of quotes at 9-10am is because we only include 9:45am to 4:00pm data. This also brings about higher average number of quotes for minute 45 to minute 59. However, the large spike in minute 59 is still significant.

We take a first look at data. Figure 2-2 shows the average number of trades and quotes by minute of the day across the 30 trading days from October 3rd, 2016, to November 11th, 2016. We see that trades and quotes are very uneven across the day, with the striking but expected feature that at end of the trading day, investors, especially day traders, are busy closing their positions, leading to a huge spike in quotes and trades activities. While quotes show a small spike near end of lunch break, no such spike is observed for trades data. After the market opens in the morning, we see a slight increase in quotes and trades activities until 11am, but then both trades and quotes gradually decrease over the course of the day until the large spike at 4pm.

The listing exchange, NYSE, is labelled 'N'. If an exchange retracted ask (bid), we see an ask (bid) price of $0, and we let its bid (ask) equal $0 as well, consistent with Hasbrouck (2019). Then we get each exchange's ask and bid price series and use each series' most recent observation to 'fill forward'. Then at each point in time, we find the best bid and ask price from non-listing exchanges, and this becomes the basis of our analysis, other than listing exchange's ask and bid prices. Next, we convert the event-time observations to clock-time in

2s, 1s, 0.1s, and 0.01s.

In this paper, we consider clock time in 2s, 1s, 0.1s, and 0.01s data. While 1s data appears to be the most common approach in market microstructure literature (Hasbrouck, 1995; Shastri, Thirumalai, and Zutter, 2008; Anand, and Subrahmanyam, 2008; among others), trade time (event time) is another option one could consider, especially when trading is not at regular intervals. However, in trade time, non-synchronous trading is a problem, and we may end up with incorrect inferences when using data thinning algorithms to address this problem (Lehman, 2002). Considering price discovery in clock time also allows us to compare our results with existing literature.

The illustrate the issue of data sparsity, which arises when we raise trades and quotes data to much higher frequency (0.01s to 2s in our case), we show in Table 2-2 the percentage of instances where we see 0 market, 1 market, and 2 markets move across different resolutions from 0.01s to 2s for quotes (on the top) and trades (on the bottom) across the 30 trading days from October 3rd, 2016 to November 11th, 2016. At 0.01s resolution, we see no market move over 99% of the time for both trades and quotes. As we move to coarse resolution, this issue eases. At 2s, we see at least one market move for about half the time. This shows that there appears to be very sparse information arrival at high frequency if we consider price change as proxy for information arrival. See, for example, Du and Zhu (2017).

Sampling Interval	0.01 second	0.1 second	1 second	2 seconds
Mid-quotes (Listing vs Other)				
Random Walk Volatility	19.952	19.559	19.891	19.902
Average	0.465	0.617	0.757	0.816
St Dev	0.051	0.049	0.048	0.045
Min	0.360	0.508	0.649	0.705
Max	0.586	0.734	0.855	0.899
Quotes against Trades				
Random Walk Volatility	20.657	20.665	20.630	20.619

Average	0.246	0.341	0.438	0.501
St Dev	0.047	0.061	0.078	0.088
Min	0.123	0.176	0.206	0.222
Max	0.307	0.416	0.534	0.612

Table 2-2 Cross-market Return Correlations

This table reports summary of daily cross-market return correlations at different sampling intervals.

Now consider only instances where at least one market moves. For 0.01s, when at least one market moves, we see 27% of the time both markets move together. At 0.1s, this rises to 34%, and 49% at 1s; finally, at 2s, it's 57%. On average, we also expect the markets to move in the same direction because given sufficient time (and in this automated trading world, sufficient time can me seconds or even milliseconds), markets should react to the same set of news and make similar reactions. Therefore, we also report cross-market return correlation in Table 2-2. We see that even at 0.01s, return correlation is 0.465 for quotes and 0.246 for trades and against quotes. As we move to coarse resolutions, the correlation goes up steadily but much more so for quotes than trades against quotes. For quotes (listing vs the other exchanges), correlation is as high as 0.816 at 2s resolution. For quotes, the correlation is lower at 0.501. One possible explanation is that we have much move (about 14 times) quotes than trades data such that it is less likely to observe trades and quotes change at the same time. For quotes, about 33% of observations are from the listing exchange hence the sufficient data points bring up correlation. Our estimation is indeed closely related to correlation because it is based on the common component. However, they are two conceptually and statistically different method. This is evidenced in later sections where we see that the common component is much stronger than correlation.

In the first set of analysis, we assess informational contents of quotes from different venues. To do so, we compose two log-price series: log-midquotes for listing exchange (Lex) and log-midquotes from all other exchanges (Other), which is calculated from the best ask and bid from non-listing exchanges. This analysis is based on WRDS TAQ dataset with one timestamp. Consistent with Hasbrouck (2019) and literature, when one exchange poses a bid

or ask price of \$0, the exchange has retracted its quote, and that exchange is excluded from calculation of 'Other midquotes' until it poses a valid quote. In converting the event time quotes to clock time, we create a time grip of 2s, 1s, 0.1s, and 0.01s, and use the most recent quote to fill up each time grip. Then for each trading day from October 3rd, 2016, to November 11th, 2016, we estimate price discovery metrics including common and market information shares. We obtain a total of thirty estimates for each price discovery metric and report the trend over time.

2.3.1 Model set-ups

Without sacrificing generalization, we illustrate IS and CS in a two-market case for simplicity. The notion here mainly follows Hasbrouck (1991a, b; 1995) but we also note literature convention for generality purpose.

Let $\mathbf{p}_{i,t} = (p_{1,t}, p_{2,t})'$ be a vector of log prices of an asset traded at two markets. As $p_{1,t}$ and $p_{2,t}$ relate to the same underlying asset, they cannot deviate too much from each other because of arbitrage. It is further assumed that the price series contains a random walk component such that:

$$\mathbf{p_{i,t}} = \left(p_{1,t}, p_{2,t}\right)' \sim I(1)$$

$p_{1,t}$ and $p_{2,t}$ are prices of the same security traded at different venues, such as different exchanges; more generally, $\mathbf{p}_{i,t} = (p_{1,t}, p_{2,t})'$ can be two closely related assets, such as a stock and an arbitrary derivative security based on that stock. In some applications, $\mathbf{p}_{i,t} = (p_{1,t}, p_{2,t})'$ may be ask and bid prices if one believes they possess different information contents, and one finds it interesting to differentiate them. It follows that:

$$p_{1,t} = p_{2,t} + \mu_t$$

Where $\mu_t \sim I(0), \beta = (1, -1)$ is the cointegration vector such that $\beta \mathbf{p}_t = p_{1,t} - p_{2,t} \sim I(0)$. We also have $\Delta p_{i,t} = p_{i,t} - p_{i,t-1}$ is the log-return. It follows that $\Delta \mathbf{p_{i,t}} \sim I(0)$. The Vector Error Correction Model (VECM) representation using k lags is then:

$$\Delta \mathbf{p}_t = \alpha\beta' \mathbf{p}_{t-1} + \Gamma_1 \mathbf{p}_{t-1} + \Gamma_2 \mathbf{p}_{t-2} + \cdots + \Gamma_k \mathbf{p}_{t-k} + \epsilon_t$$

Where $\epsilon_t \sim N(0, \Omega)$. One can use ordinary least square to estimate the parameters $(\Gamma_1, \Gamma_2, \ldots \Gamma_k, \alpha, \Omega)$. Note that β is pre-specified by assumption; $\alpha \neq 0$ contains the error correction coefficients that measures each prices' expected speed in eliminating the price difference (Yan and Zivot, 2010).

The notation may also be written as:

$$\Delta \mathbf{p}_t = \alpha\beta' \mathbf{p}_{t-1} + \Sigma j$$

This makes it clear that the first term represents the long-run equilibrium of the price vectors, and the second term represents, among other things, short-run dynamics induced by market imperfections (Baillie et al, 2002).

Assuming log price changes are covariance stationary, we can use a Wold representation to write log price changes in terms of the innovations, or reduced-form shocks:

$$\Delta \mathbf{p}_t = \Psi(L)\, \epsilon_t$$

This is the vector moving average (VMA) representation found in some literature. We can use Beveridge-Nelson decomposition (Beveridge and Nelson, 1981) to decompose the common trends in prices:

$$\mathbf{p_t} = \mathbf{p_0} + \Psi(1)\Sigma_{s=0}^{t} \epsilon_s + s_t$$

Where $\Psi(1) = \Sigma_{k=0}^{\infty} \Psi_k$ is the is the sum of the moving average coefficients and represents the *cumulative effect* of shock ϵ_t on all future prices, hence is the long-run price impact of shock ϵ_t; $s_t = \Psi^*(L)\epsilon_t \sim I(0)$ is the noise term. By Hasbrouck (1995), as $p_{1,t}$ and $p_{2,t}$ are just the same asset's price at different trading venues, the long-run impact of the same shock ϵ_t should be the same on $p_{1,t}$ and $p_{2,t}$ in principle. Hasbrouck (1995) further shows that the rows of $\Psi(1)$ should hence be identical, that is, the long-run impact should be the same for all prices.

Let $\boldsymbol{\psi} = (\psi_1, \psi_2)'$ be the common row vector of $\boldsymbol{\Psi(1)}$ and let permanent innovation, or

efficient price return be:

$$\eta_t^P = \boldsymbol{\psi}' \boldsymbol{\epsilon_t} = \psi_1 \epsilon_{1,t} + \psi_2 \epsilon_{2,t}$$

Then $\eta_t^P = \boldsymbol{\psi}' \boldsymbol{\epsilon_t}$ is the permanent effect impounded into prices that is presumably brought about by the new information as captured in the reduced-form shocks $\epsilon_{1,t}$ and $\epsilon_{2,t}$. Under Hasbrouck framework, this is the common factor of the different price series. Other than the permanent price impact, we also have temporary price impacts. The transitory or temporary price impact may purely be from noise traders, or it may originate from normal operations in the market, such as bid-ask bounces and inventory adjustment. Then under Hasbrouck (1995; 2019), we can use a random walk representation to rewrite the price series using Stock and Watson (1988):

$$\mathbf{p}_t = \mathbf{p_0} + \mathbf{1_n} m_t + u_t$$

Where efficient price is a random walk: $m_t = m_{t-1} + \eta_t^P$, $u_t = \Psi^*(L)\epsilon_t$. The random walk representation implies that the cointegrated prices are comprised of three parts: 1) a constant (which is not necessarily explicitly considered in literature; for example, Hasbrouck, 1991b, which does not contain the constant term); 2) an unobserved common full-information value, m_t; and 3) a transitory pricing error u_t (Yan and Zivot, 2010).

Another representation of Stock and Watson (1998) is:

$$p_t = f_t + G_t$$

Which states that the observed price has two parts, a common factor f_t, and a transitory component G_t which only affects the observed asset price temporarily.

Hasbrouck (1995) IS and Gonzalo and Granger (1995) component share are both concerned with the common component. However, in Gonzalo and Granger (1995), they consider:

$$f_t = \Gamma p_t$$

Where Γ is the coefficient of the common factor. Gonzalo and Granger (1995) show that Γ is orthogonal to the error correction vector, α. Baillie et al (1995) note that Gonzalo and

Granger (1995) developed a statistical test to test if one of the factor components is the only contributor to common factor. Under their framework, we may consider f_t as forming a portfolio where the portfolio weights are defined by Γ. f_t is I(1) because the price series are I(1) and the error correction term is I(0).

Now it is clear that both IS and CS try to attribute the source of price impact to one of the markets. The difference is that Hasbrouck IS considers each market's contribution to innovations in efficient price's variance, while Gonzalo and Granger (1995) CS decomposes the common factor into a combination of the different price series. Both models rely heavily on $\alpha_{\perp}$.

2.3.2 Specifying IS

Having established the set-up, we can now specify IS and CS. As $\boldsymbol{\psi} = (\psi_1, \psi_2)'$ is the common row vector of $\boldsymbol{\Psi(1)}$, $\boldsymbol{\psi}\epsilon_t$ is then the incremental change in price permanently incorporated into asset price that, by assumption, is brought about by new information in the structural shock ϵ. The variance of $\boldsymbol{\psi}\epsilon$ is:

$$var(\boldsymbol{\psi}\epsilon_t) = \boldsymbol{\psi}\Omega\boldsymbol{\psi}'$$

If Ω is diagonal, $var(\boldsymbol{\psi}\epsilon_t) = \boldsymbol{\psi}\Omega\boldsymbol{\psi}'$ contains two non-zero terms representing the efficient price innovation from each market. Hasbrouck (1995) proposes that we can measure a market's contribution to price discovery using the share of the variance of $\boldsymbol{\psi}\epsilon_t$ due to market i:

$$IS_i = \frac{\psi_i^2 \sigma_i^2}{\psi_1^2 \sigma_1^2 + \psi_2^2 \sigma_2^2}, \quad i = 1,2$$

Where σ_i^2 is the i-th diagonal element of Ω. Note, $IS_1 + IS_2 = 1$ by construction. However, it is often the case that price innovations are not uncorrelated, hence Ω is not diagonal. When this happens, Hasbrouck (1995) suggests that we could use Cholesky decomposition on Ω. IS is then measured using the orthogonalized innovations. Let F be a lower triangular matrix such that $FF' = \Omega$. Then IS for the i-th market is:

$$IS_i = \frac{(|\psi' F|_i)^2}{\boldsymbol{\psi}\Omega\boldsymbol{\psi}}$$

Note that by construction, IS is sensitive to ordering. We obtain a higher and lower bound on market i's IS by re-calculating IS using reversed orders. In n-market case, we need to investigate all the possible permutations. Under this framework, stronger correlation between different markets gives a higher upper bound, and a lower correlation gives a smaller lower bound. The upper bound contains both the market's own contribution to efficient price innovation and its correlation with the other markets, while the lower bound contains only the market's own contribution (Baillie et al 2002). When using high frequency data, the higher and lower bounds are close (Hasbrouck 1995, 2019; Tse 2000). However, when the frequency is low, the upper and lower bounds could be very different because of the large correlation between markets (Huang 2000; Booth, Lin, Martikainen, and Tse 2002).

2.3.3 Competing price discovery metrics: Component Share (CS), Information Leadership (IL), and Information Leadership Share (ILS)

Another popular price discovery measure is Component Share (CS) based on Gonzalo and Granger's (1995) common factor component model. We first review its estimation method in this session.

Decomposing price vector into permanent and transitory component under Gonzalo and Granger (1995), we have:

$$\mathbf{p}_t = \mathbf{A_1} f_t + \mathbf{A_2 z_t}$$

Where $f = \gamma' \mathbf{p_t} \sim \boldsymbol{I}(\mathbf{1})$ is the permanent component, $z_t \sim I(0)$ is the transitory component, A_1 and A_2 are the loading matrices, γ' is the matrix of common factor weights. Granger and Gonzalo (1995) further defines $\gamma = (\alpha'_\perp \beta_\perp)^{-1} \alpha'_\perp$, where $\alpha'_\perp$ and $\beta_\perp$ are 2×1 vectors such that $\alpha'_\perp \alpha = 0 \; and \; \beta'_\perp \beta = 0$. One convenient choice of $\beta_\perp$ is $\beta_\perp = \mathbf{1} = (1,1)'$, and it follows that $\gamma = (\alpha'_\perp \mathbf{1})^{-1} \alpha'_\perp$. Booth et al. (1999), Chu et al. (1999), Harris et al. (2002), and Bailie et al (2002) suggest measuring market i's price discovery using:

$$CS_i = \frac{\alpha_{\perp i}}{\alpha_{\perp i} + \beta_{\perp i}}, i = 1,2$$

IS and CS attracted substantial confusion in the early days and considerable attention was paid to understand them. See, for example, special issue on price discovery on *Journal of Financial Markets* (Issue 3, 2002). They are both constructed under the 'who moves first' idea and are designed to tell us which market incorporates new information about the asset's fundamental value into its prices faster. For example, Hasbrouck (1995) explicitly states that IS is designed to measure 'who moves first in the process of price adjustment'. Baillie et al (2002), Hasbrouck (2002), and Lehmann (2002) all provide simple microstructure models and provide simple comparison of IS and CS (Yan and Zivot, 2002). The IS and CS are similar in some instances but may be quite different in other problems studied. While our study is an extension of Hasbrouck (1995) IS and we do not intend to answer which one is 'better', we review some debates here.

Baillie et al (2002) is one comprehensive comparison between component share and information share which gives empirical experiments on IS and CS. Deviating from information share, which defines each market's information contribution as each market's contribution to efficient price return's variance (which we follow in our model), component share's focus is the common factor and error correction process. Baillie et al (2002) show that if different markets' residuals are uncorrelated with each other, component share produce very similar results to information share. However, this is a very rare event, and it is hence unsurprising to see that information share and component share are usually very different in magnitude.

The idea behind component share is very straight-forward and follows directly from market features: if one asset is traded in different markets, cross-market arbitrage would keep their prices identical, or at least sufficiently identical when considering transaction costs; in other words, the prices are I(1) series. We say that the different price series have the same underlying stochastic common factor. When the common factor is unique, we usually consider the common factor the efficient price. Information share and component share are

both based on VECM model. Component share defines price discovery share as each market's contribution to the common factor, efficient price, as measured by each market's error correction parameter. Baillie et al (2002) gives an excellent example of how they differ: when market 1 and 2 are highly positively correlated and cointegrated, and assume market 1 responds to its deviation from market 2's price as measured by the error correction term but not the other way around, then component share suggests that market 1 contributes little to price discovery; information share would, however, assign information share to both market. IS and CS are therefore very different measures of the same issue.

By incorporating contemporaneous terms in VECM model, information share requires markets to be ordered and uses Chelosky decomposition. Because of the high correlation often observed in real data, information shares usually give a wide range. To make estimates more precise, Hasbrouck (1995, 2019) uses, wherever possible, high resolution data such that at each clock tick, often only one market moves. In 1995, this is 1-second; and in 2019, this is 10^{-5}-second. This, however, introduces more complications as it is questionable whether information is present at such high frequency trading data.

A few other studies try to answer the basic question: what IS and CS really measure, and which one is better. The question is not very clear-cut, because IS and CS are defined using reduced-form shocks, which are typically interpreted as forecasting errors and can be either informational or non-informational (Yan and Zivot, 2010). Yan and Zivot further use a structural cointegration model to show that neither measure on its own can distinguish the price discovery dynamics between markets; the CS does not reflect a market's response to new information at all, while IS cannot be interpreted unambiguously. Using simulations, Putniņš (2013) further shows that IS and CS are only consistent with the 'who moves first' view if the different price series contain the same level of noise; if not, they reflect both speed of adjustment for new information and relative avoidance of noise. Putniņš (2013) then proposes a new information leadership share as:

$$ILS_1 = \frac{IL_1}{IL_1 + IL_2}, ILS_2 = \frac{IL_2}{IL_1 + IL_2}$$

Where:

$$IL_1 = \left|\frac{IS_1}{IS_2}\frac{CS_2}{CS_1}\right|, IL_2 = \left|\frac{IS_2}{IS_1}\frac{CS_1}{CS_2}\right|$$

IS for each market is the simple average of all possible orderings under VECM.

2.3.4 Specifying common information share (CIS) and market information share (MIS)

We decompose reduced form shock into a common component, f_t, and a market-specific component, η_t, which are not correlated with each other:

$$\epsilon_t = \delta f_t + \Sigma^{\frac{1}{2}}\eta_t = \begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} f_t + \begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}\eta_{1t}\\ \eta_{2t}\end{bmatrix}$$

The common component should have the same effect on both markets as they relate to the same underlying asset. We may therefore impose a further restriction that $\delta_1 = \delta_2$. Then efficient price return is:

$$\begin{aligned} w_t = m_t - m_{t-1} = \psi\epsilon_t &= (\psi_1\ \psi_2)\left(\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} f_t + \begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}\eta_{1t}\\ \eta_{2t}\end{bmatrix}\right) \\ &= (\psi_1\delta_1 + \psi_2\delta_2)f_t + (\psi_1\sigma_1)\eta_{1,t} + (\psi_2\sigma_2)\eta_{2,t} \end{aligned}$$

Efficient price return's variance is of the form:

$$\sigma_w^2 = (\psi_1\delta_1 + \psi_2\delta_2)^2 + (\psi_1\sigma_1)^2 + (\psi_2\sigma_2)^2$$

MIS and CIS, the price discovery measure that considers common (public) information, for the common component (f_t), market 1, and market 2 are:

$$CIS = \frac{(\psi_1\delta_1 + \psi_2\delta_2)^2}{\sigma_w^2}$$

$$MIS_1 = \frac{(\psi_1\sigma_1)^2}{\sigma_w^2}$$

$$MIS_2 = \frac{(\psi_2\sigma_2)^2}{\sigma_w^2}$$

To estimate, we now impose the restriction $\delta_1 = \delta_2$. We have:

$$\Omega = \delta\delta' + \Sigma$$

Or:

$$\begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} = \begin{bmatrix} \delta_1^2 & \delta_1^2 \\ \delta_1^2 & \delta_1^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

We therefore have a system of three linear equations in three unknowns:

$$\omega_{12} = \delta_1^2$$

$$\omega_{11} = \delta_1^2 + \sigma_1^2$$

$$\omega_{22} = \delta_1^2 + \sigma_2^2$$

Which is just-identified. The solutions are hence:

$$\delta_1^2 = \omega_{12}$$

$$\sigma_1^2 = \omega_{11} - \omega_{12}$$

$$\sigma_2^2 = \omega_{22} - \omega_{12}$$

To estimate the parameters, we simply need to estimate the variance-covariance matrix, Ω, of the reduced-form shock, ϵ_t.

To give a direct comparison of our technique and Hasbrouck IS, we note that the literature, pioneered by Hasbrouck IS framework, seeks to answer 'who moves first', and whoever is the first to incorporate new information into prices as shocks arrive the market is the price leader. To do so, one needs to use very high frequency data to (at least hopefully) remove inter-market correlation. If one uses 'coarse' resolution data which is not free from inter-market correlation, the higher and lower bound of Hasbrouck IS would be very high. By comparison, the MIS-CIS method in this paper explicitly acknowledges that correlation between different markets exist, and that this is a normal part of the market which we do not need to get rid of. We argue that the correlation comes from common information in the market which everyone in every trading venue sees, and this information is not from any individual market. It is hence inappropriate to assign the common information to any one market. At very 'coarse' resolutions (in this high frequency trading world, the word 'coarse'

is not used to its usual meaning, and we mean 2s for 2016, see empirical section), common information dominates. As we move to finer resolutions, we move from common to private and order-driven information. At very high frequency, trades and quotes become less information based. For example, competition for speed replaces information-based competition. However, ultra-high frequency data is exactly the dataset considered 'appropriate' under Hasbrouck framework.

2.4 Empirical results

This section presents empirical results of IS, ILS, and our IS-C first for October 3rd, 2016, and then the full month of October 2016. Our results for October 3rd , 2016 are similar to Hasbrouck (2019) and our IS-C further shows that it is the common component that dominates price discovery, instead of each market's unique contributions. The common component's importance (weight) drops as we go to finer resolutions and as we approach event time, the common component approaches 0. In Hasbrouck (2019), 10 seconds worth of lags are used throughout. For robustness, we include more lags than Hasbrouck and literature. So, for example, on 0.01s resolution, 1000 lags are used. In this paper, we used 100 lags on 2s resolution, 100 lags on 1s resolution, 1000 lags on 0.01s resolution, and 3500 lags on 0.01s resolution, significantly larger than Hasbrouck (2019) and literature. For event time analysis, we use 200 lags while Hasbrouck (2019) uses 10 lags.

	IS					ILS			MIS and CIS		
Frequency	Listing		LexMean	Other		OtherMean	Listing	Other	CIS	MIS(listing)	MIS(other)
2s	0.084	0.951	0.518	0.049	0.916	0.483	0.409	0.591	0.927	0.050	0.024
1s	0.138	0.942	0.540	0.058	0.862	0.460	0.384	0.616	0.886	0.087	0.027
0.1s	0.239	0.873	0.556	0.127	0.761	0.444	0.476	0.524	0.772	0.156	0.071
0.01s	0.296	0.774	0.535	0.226	0.704	0.465	0.524	0.476	0.645	0.208	0.147
event time	0.650	0.666	0.658	0.334	0.350	0.342	0.640	0.360	0.032	0.639	0.329

Table 2-3 IBM price discovery metrics for 20161003

This table shows price discovery metrics for IBM on different resolutions. Hasbrouck Information Share (IS) has an upper and lower bound and we also show their mean value for Listing Exchange (Lex) and Other Exchanges (Other). In this paper, we used 100 lags on 2s resolution, 100 lags on 1s resolution, 1000 lags on 0.01s resolution, and 3500 lags on 0.01s resolution

Sampling Interval	0.01 second			0.1 second			1 second			2 seconds		
Non-zero Returns	0	1	2	0	1	2	0	1	2	0	1	2
Mid-quotes												
Average	99.2%	0.58%	0.21%	94.3%	3.73%	1.94%	67.6%	16.6%	15.9%	52.5%	20.6%	26.9%
St Dev	0.2%	0.2%	0.1%	1.4%	1.0%	0.5%	5.8%	2.5%	3.7%	7.0%	2.3%	5.5%
Min	98.6%	0.4%	0.1%	90.5%	2.4%	1.0%	54.4%	11.3%	9.2%	37.7%	14.5%	16.8%
Max	99.5%	1.1%	0.3%	96.5%	6.7%	3.0%	77.7%	23.4%	22.2%	65.4%	25.1%	37.2%
Trade Prices												

Average	99.6%	0.38%	0.06%	96.2%	3.10%	0.65%	72.1%	21.2%	6.8%	54.4%	32.4%	13.3%
St Dev	0.1%	0.1%	0.0%	1.1%	0.9%	0.2%	5.2%	3.3%	2.1%	6.3%	3.1%	3.7%
Min	99.0%	0.3%	0.0%	91.9%	2.2%	0.4%	52.7%	16.9%	3.8%	32.3%	27.9%	7.9%
Max	99.7%	0.9%	0.1%	97.3%	6.6%	1.4%	78.4%	33.8%	13.5%	63.2%	43.0%	24.6%

Table 2-4 Frequency of Zero and Non-zero Returns

This table shows the percentage of non-zero returns for IBM's trade and quote prices across 0.01s to 2s intervals.

2.4.1 Comparing IS, ILS, and IS-C for 20161003

Hasbrouck (2019) conducts most of his analysis based on the one day of October 3rd, 2016. For comparison, we present in Table 2-3 IS, ILS, and IS-C for this day. For the listing exchange (Lex) and other exchanges (Other), we present their Hasbrouck information share (IS), and common component-market component (IS-C) shares. As information share is in the form of an upper and lower bound, we also present the mean value of the bounds for easier interpretation. We see that as the resolution gets finer, IS bounds become narrower and it has a limit in event time. By construction, only one market is allowed to move in event time, and it is essentially analogous to a sequential market instead of parallel market, as there is no 'parallel' component. IS suggests that, throughout all resolutions, listing exchange dominates price discovery, although the domination is larger in event time than clock time suggests. In clock time, listing exchange's domination is only marginal as the IS has a mean slightly larger than 0.5. As suggested by Yan and Zivot (2010) and then Putnins (2013), the IL and its share form, ILS, which is based on IS and CS, should be a more reliable measure of price discovery share. The ILS is more volatile than IS at different resolutions and suggests that the listing exchange does not dominate price discovery for 2s, 1s,and 0.1s data; at 0.01s and event time, the listing exchange dominates price discovery, and the domination is significantly larger at event time than clock time. They both suggest that the listing exchange's price discovery is significantly larger than suggested by its market share (around 20%). Hasbrouck (2019) suggests a possible explanation: designated market makers are present in the listing exchange, and they are critical in price discovery; for example, when due to the designated listing exchanges, liquidity would drastically fall when the listing exchange's trade is disrupted (Clark-Joseph, Ye, and Zi, 2017).

When moving to common component-market component share, we see that in clock time, the common component dominates price discovery across all resolutions, though its domination drops as we move to finer resolutions. This is as expected because in finer resolutions, the correlation among the two price series' returns drops, and we are more and more able to uniquely assign price discovery to one of the markets. In the limit, we have event time analysis, which by construction should have no 'parrel market' component and hence no 'common component'. Our analysis suggests that all markets respond to the same set of common information which have the same

impact on all markets, and each market's unique contribution is small.

We note that at event time, listing exchange and other exchange's price discovery shares are similar under IS and IS-C, confirming that our measure is indeed identical to IS when there is no common component.

2.4.2 Comparing IS, ILS, and IS-C for 201610

	IS		ILS		MIS and CIS		
	LexMean	OtherMean	Listing	Other	CIS	MIS(listing)	MIS(other)
2s	0.468	0.527	0.667	0.333	0.932	0.040	0.028
	-0.038	-0.055	-0.261	-0.261	-0.035	-0.039	-0.027
1s	0.457	0.543	0.668	0.332	0.902	0.051	0.047
	-0.058	-0.058	-0.180	0.180	-0.029	-0.036	-0.030
0.1s	0.438	0.562	0.622	0.378	0.797	0.094	0.109
	-0.074	-0.074	-0.086	-0.086	-0.038	-0.046	-0.048
0.01s	0.438	0.562	0.586	0.414	0.659	0.160	0.182
	(0.069)	(0.069)	-0.036	-0.036	-0.042	-0.038	(0.050)
event time	0.550	0.450	0.599	0.401	0.031	0.542	0.426
	-0.143	-0.143	-0.125	-0.125	-0.008	-0.140	-0.137

Table 2-5 IBM price discovery metrics for 201610

This table shows the price discovery metrics for October 2016, with standard deviations in parathesis.

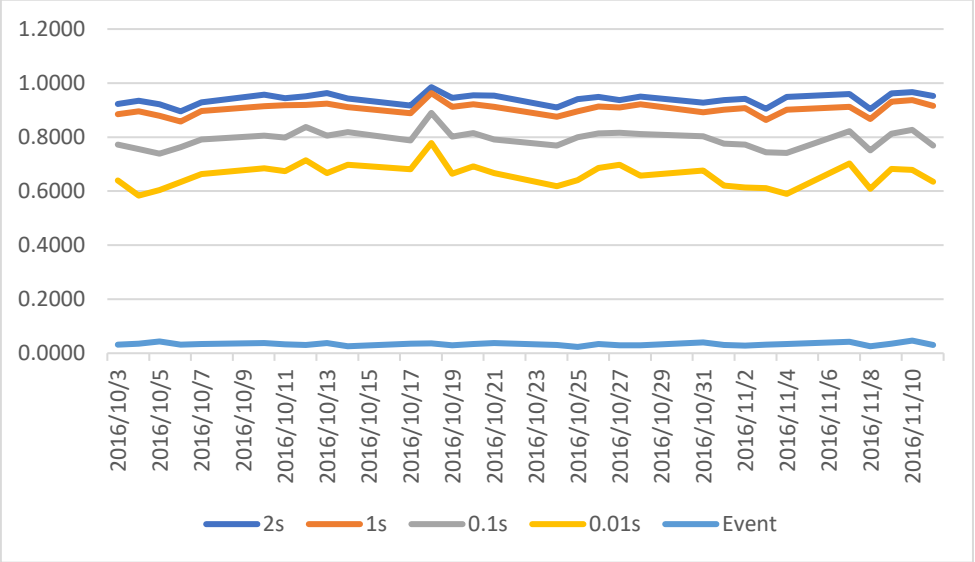


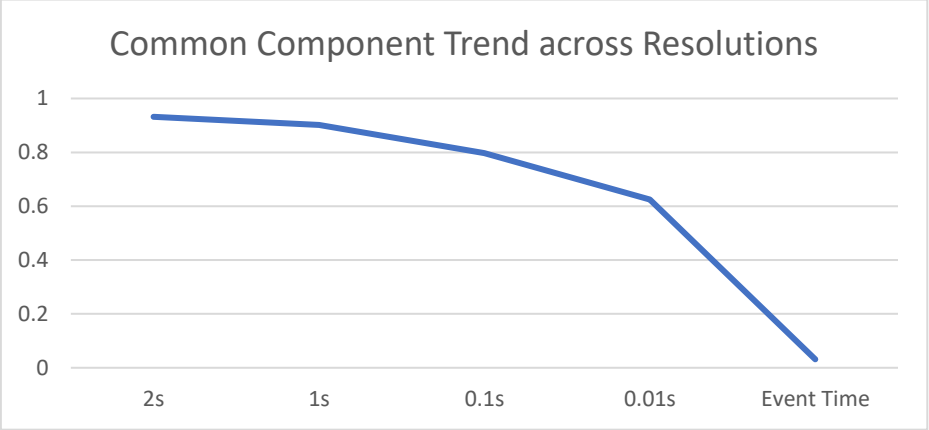
Figure 2-3 Trends in common component in 2016
This graph shows the common component's trend over time for October 2016 on 2s, 1s, 0.1s, 0.01s, and event time resolutions.

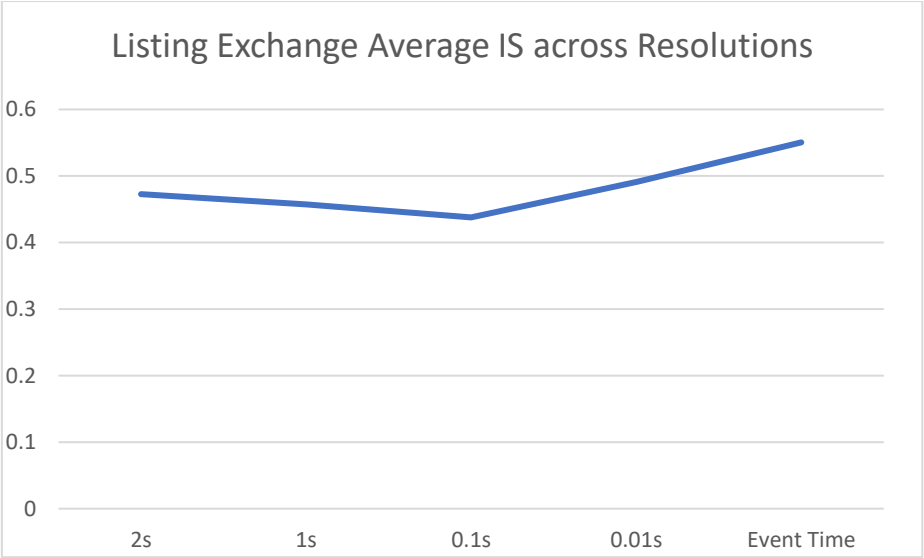
We proceed to provide the same set of analysis but now on 30 trading days from October 3rd, 2016, to November 11th, 2016, in Table 2-5, which presents the mean and standard deviation of the 30 trading days' estimates; Figure 2-3, which shows the trends in common component over time, and Figure 6-4, which shows average common component, listing exchange's IS, and listing exchange's ILS over time.

From Table 2-5, we note that listing exchange does not dominate price discovery in 2016 on average, but its price discovery share (around 0.45 at clock time, and 0.55 at event time) is still much larger than suggested by its market share. As we move to finer resolutions, IS gets more volatile. ILS, on the contrary, suggests that the listing exchange still dominates price discovery across all resolutions and its dominance drops as we move to finer resolutions. The standard deviation of ILS does not increase or decrease monotonically. The common component dominates price discovery across all clock time, but diminishes to near-zero at event time, as expected. Its standard deviation also does not show a clear pattern.

Figure 2-3 shows that at different clock time, we see moderate daily variations, but different clock time follow the same trend over time. As we move to finer resolutions, common component becomes less and less significant. In some days, each market's unique information is larger than

other days, maybe suggesting days with fewer news announcements hence lower common information component, or different clienteles using the different markets had access to information sources which are generating periodic data, et cetera.





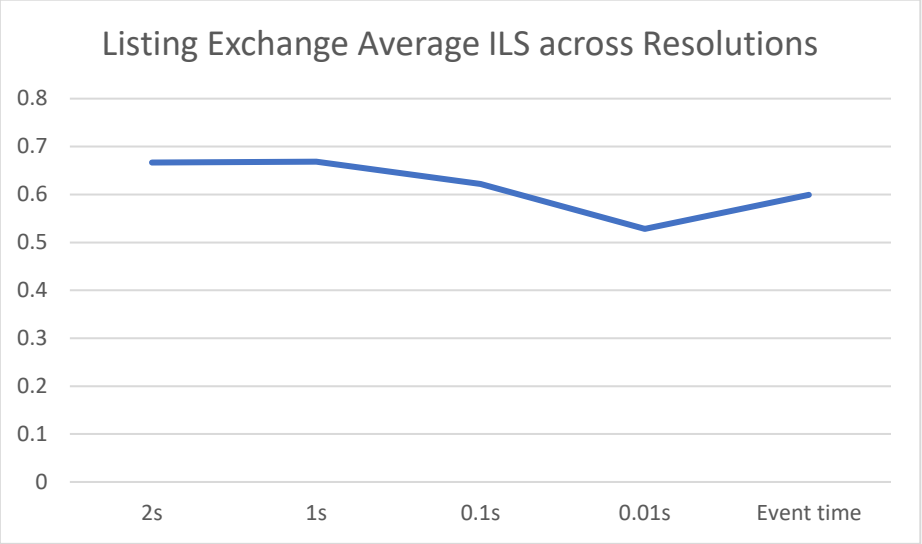


Figure 2-4 Common component, listing exchange's IS, and listing exchange's ILS over different resolutions

This pattern is made clearer in Figure 2-4: common component drops monotonically as we move to finer clock time resolution and then to event time. The listing exchange's IS drops slightly from 2s to 0.1s resolutions but increases as we move to 0.01s and event time resolutions. ILS generally shows a decreasing trend as we move to finer resolutions and then to event time, with a lowest point at 0.01s resolution.

Panel A. Common information share using all data points

Sampling Interval	0.01 second	0.1 second	1 second	2 seconds
Listing vs other exchanges				
Average	0.658	0.788	0.904	0.94
St Dev	0.042	0.039	0.022	0.02
Max	0.778	0.889	0.962	0.985
Min	0.583	0.674	0.857	0.895
Quotes and trades				
Average	0.594	0.748	0.858	0.899
St Dev	0.082	0.084	0.076	0.076
Max	0.702	0.863	0.963	0.993
Min	0.377	0.514	0.621	0.635

Panel B. Common information shares using data points within five standard deviations from mean of price and return

	0.01 second*	0.1 second	1 second	2 seconds
Listing vs other exchanges				
Average		0.708	0.817	0.854
St Dev		0.061	0.061	0.06
Max		0.844	0.926	0.96
Min		0.573	0.654	0.694
Quotes and trades				
Average		0.759	0.837	0.902
St Dev		0.075	0.057	0.069
Max		0.874	0.941	0.986
Min		0.505	0.731	0.639

Common information share in 2000

Sampling Interval	1s	10s	15s	30s
Average	0.115	0.577	0.666	0.698
St Dev	0.035	0.077	0.1	0.107
Max	0.171	0.717	0.956	0.846
Min	0.045	0.462	0.517	0.33
Quotes and trades				
Average	0.138	0.681	0.773	0.859
St Dev	0.088	0.154	0.131	0.162
Max	0.561	0.863	0.884	0.967
Min	0.039	0.187	0.26	0.307

Common information in 2008

Sampling Interval	1s	2s	5s
Average	0.671	0.809	0.928
St Dev	0.046	0.034	0.025
Max	0.765	0.865	0.981
Min	0.549	0.716	0.877
Quotes and trades			
Average	0.775	0.857	0.92
St Dev	0.053	0.048	0.051

Max	0.878	0.903	0.986
Min	0.63	0.723	0.799

Table 2-6 Common Information Share in 2000, 2008, and 2016
These tables show common information share of IBM in 2000, 2008, and 2016 for one month in October of each year. Panel A shows results using all data, and Panel B shows results using data points within five standard deviations of price and returns.

We next ask how outliers affect our estimates. Outliers in stock prices and returns may potentially disturb our results and make them less reliable. To do so, we filter away observations with stock prices and returns that are more than five standard deviations from their mean values. The estimated results are presented in Table 2-6. As a direct comparison, we show common information shares using all data points in the above panel and common information shares using all data within five standard deviations of mean price and return in Panel B. While common information share still dominates across all resolutions considered, we note that for both quotes (listing against other exchanges) and trades against quotes, common information generally decreases. While the estimates are not more precise, the extreme values in the estimates greatly reduce.

2.4.3 Results for 2000 and 2008

Next, we ask how the information share performs during earlier years. To do so, we perform our analysis for 2008 and 2000 at eight-year intervals.

The exact identity of exchanges is different in 2008 and 2000 are different from 2016 which we cannot confirm because of the lack of official dictionary for the exchange names. We are hence only able to identify the listing exchange, assuming that the listing exchange's symbol keeps constant over the years.

For 2000, we show common information share cross 30 trading days in October and November for listing against other exchanges in Table 2-6. The top panel shows information shares of listing against other exchanges while the bottom panel shows quotes against trades. 2000 sees much fewer trades and quotes activities and we see that at 1s interval, only 11.5% and 13.8% price discovery is explained by common information for listing against other exchanges, and trades against quotes, respectively. This is not surprising to us because in 2000, HFT had not kicked the market and 1s interval is very high frequency back then. As we move to lower frequency, more and more price discovery is explained by the common information. At 30s interval, around 70% and 86% information is explained by common information for listing against other exchanges and trades

against quotes, respectively.

We do the same and show results for 2008 in Table 2-6. Across 1s to 5s, common information dominates information share. At 5s resolution, over 92% price discovery is from common information.

As a direct comparison, we consider the information content at 1s for 2000, 2008, and 2016. The common information increases from 11.5% in 2000 to 67.1% in 2008 and to 90.4% in 2016 for listing against other exchanges; and from 13.8% in 2000 to 77.5% in 2008 and to 83.7% for trades against quotes. Over the years, as trading becomes more and more frequent because of advances in technology, common information's share increases sharply and steadily. This is because as speed raises, we are more able to trade on the same set of information.

2.5 Conclusion and suggestion for future research

In this paper, we re-examine price discovery by introducing a common component besides the market-specific component and we find that the common component dominates price discovery across 2s to 0.01s resolutions. The common component decreases monotonically as resolution gets finer and approaches 0 in event time. We note that the Hasbrouck (1991a, b; 1995) framework, which seeks to use ultra-high frequency data in clock time, is inappropriate to investigate the informational contents of trades and quotes, and it overlooks the common component in asset returns which is not from any individual market. There are fair variations across days in the common and market-specific components but at different resolutions, the patterns are the same. This is contrary to existing literature and framework which seeks to use ultra-high frequency data to identify which market moves first and hence leads price discovery. We also check for 2000, a pre-HFT period, and 2008. Common information, as expected, increases steadily and sharply over the years (controlling for sampling interval).

This framework can potentially be applied to other set-ups. For example, Hasbrouck (2003) investigates the informational contents of all index products (futures, ETFs, and E-minis) and finds that E-mini dominates price discovery among these index products and its domination is higher than its market share suggests. We may apply our metric to the derivatives market and investigate if it is E-mini that contributes the most to the derivatives' price discovery or it is common information that dominates. If E-mini still dominates price discovery, it is then interesting to investigate why E-mini contains more or superior information than the other markets.

Chapter 3 Portfolio Construction with News Sentiment using Large Language Model

3.1 Introduction

In recent years, more and more textual data are recorded digitally and made available for research. Unsurprisingly, we see a surge in textual analysis in finance (Gentzkow, Kelly, and Taddy, 2019) applied in multiple fields of finance and economics. Finance and economics literature, in trying to understand textual information, typically relies on traditional dictionary-based (or *bag-of-word* methods) and use limited data sources, such as front pages of *Wall Street Journal,* because of lack of datasets and machine learning techniques. Dictionary-based methods have many advantages, and hence their popularity. For example, they are simple to use and understand; they avoid researchers' subjectivity once the dictionary is selected; we can apply dictionary-based method to any lengths of texts; once dictionaries are made public, we can easily replicate other people's studies (Loughran and McDonald, 2016).

However, dictionary methods are overly simplistic. To let the machine understand text, dictionary method reduces the whole text to word counts of dictionary words, such as positive and negative sentiment dictionary in sentiment analysis. By reducing the whole text to word counts of dictionary words, dictionary method effectively throws away almost all information. This includes but is not limited to context, word orders, inter-connections of words, grammars, and structures. In 2017, Vaswani et al introduced the Transformer architecture based on self-attention mechanism, leading to BERT of Devlin et al (2018), a milestone of language model in machine learning. It significantly outperforms previous models in natural language processing tasks, including sentiment classification. Effectively, for each word, BERT asks it to pay attention to each word (including itself) in the text, hence considering the information missed by traditional dictionary methods, including grammar, context, inter-connections of words (such as 'not'), word orders, et cetera. By utilizing all information in the corpus, it represents (or *encodes*) text much more accurately. Subsequently, it outperforms all previous models in sentiment analysis. To our knowledge, the most comprehensive comparison of Transformer-based models against lexicon-based models in financial sentiment

analysis is Mishev et al (2020). In this paper, the authors compare performances of a large variety of textual analysis algorithms [3]. The authors find that transformer-based models significantly outperform other models in terms of accuracy and F1 score. Specifically, LM dictionary achieves an accuracy of around 65%, while transformer-based models' accuracies are around 30% higher.

In this paper, we overcome the limits of dataset by using Refinitiv Machine Readable News (MRN) database. This new database contains all American company news from 2001 to 2019 from Refinitiv (Thomson Reuters). MRN database, being more comprehensive, also allows us to investigate the informativeness (as implied by portfolio performance) of 1) long and more comprehensive news that are updated more slowly; and 2) short but more timely news.

For any generic language model (including BERT), we should further adapt it to domain-language. This is because the terms and jargons used in any field is different from generic language, and financial texts are no exception. As training is extremely expensive computationally[4], we choose FinBERT of Huang et al (2023), a pre-trained model using financial documents. The authors show that FinBERT, by 'speaking' finance language, achieves higher classification accuracy than the original BERT model, as it is fine-tuned using financial text hence is better able to understand information form financial texts. It also outperforms LM dictionary and other traditional machine learning algorithms.

MRN database contains two types of news: 1) news alerts, which are timely but short updates of an event as an event is known to the market; and 2) news articles, which are a more comprehensive version that are released to the market a while after the last news alerts update. While many previous studies rely on headlines only, we experiment portfolio performance when considering 1) news alerts only; 2) news alerts and articles' headlines; and 3) articles' body contents only. The comparison would be valuable to both finance academics and practitioners. If portfolio performance using only alerts achieves at least comparable results compared with news articles for finance academics, then researchers can continue using headlines only in their future research. This

[3] This includes dictionary, Word2Vec, ELMo, GloVe, BERT, and other transformer-based models.

[4] Specifically, the computing power required is beyond almost all universities' business schools' hardware. We hence also rely on this publicly available model. We attempted but was unable to do it on our own.

significantly reduces research cost, as databases containing news articles tend to be extremely expensive. For practitioners, news alerts are shorter and easier to process, allowing one to react faster to news. Practitioners may therefore continue using news headlines if portfolio performance does not see a significant improvement.

We do not have a definitive *a priori* expectation on which model should perform better, as we have reasons to believe both ways. On the one hand, headlines are designed to be concise but precise, capturing the most important and relevant aspect of an event and importantly, is less noisy. Therefore, models using headlines only may achieve better results. On the other hand, article bodies are much more informative and comprehensive, and may be especially important in complicated cases where we cannot summarize the event using a one-sentence headline. Consequently, we should believe that models using article bodies should achieve better results, and that the performance should be higher with longer articles. Without *a prior* expectations, we rely on model performance to check if alerts perform better than articles, vice versa.

The analysis in this paper is as follows. We first predict news sentiment using FinBERT model for each news article and news alert. We then form zero-cost long-short portfolios by going long the most positive-sentiment stocks and short the most negative sentiment stocks. We achieve annual Sharpe ratios of 2.79, 3.09, and 3.87 respectively under the 3 strategies. The model achieves significant alpha over the Fama-French five factor model. Our results suggest that reacting immediately but on incomplete information may not always be a good idea. If we believe that market prices reflect and incorporate any information and events that is already known to the market, then financial academics and practitioners should only react to news alerts, which are timely reflections of events that just occurred. This sheds new light to optimal trading for high frequency traders who seek to react to new information as quickly as possible as the new information hit the market, where doing so may unexpectedly reduce their performance. By considering long news with body contents only, we also have fewer pieces of news and hence number of stocks to work with, further reducing transaction cost. Previous studies which rely on headlines only are often forced to do so because of lack of data and lack of appropriate machine learning methods to extract sentiment information from long news bodies. Our study, by formally comparing portfolio performance based on headlines and articles bodies, fills this gap.

We also document an interesting pattern in news of positive, negative, and neutral sentiments: news

pieces of positive sentiment are tailored to fewer audiences, contain fewer topics, and are generally shorter compared with neutral and negative sentiment news. This suggests that news does not just differ in their sentiment and informational contents, and merely the way the news is presented may contain valuable information.

We note that there is a closely related paper investigating the same issue using similar datasets, but we arrive at drastically different conclusions. In the working paper of Jiang, Kelly, and Xiu (2023), the authors experiment with different language models using Refinitiv Machine Readable News for the US and global markets. The authors' sample is much larger than ours as they used both Refinitiv and third-party news while we only purchased news from Refinitiv. The methods are also different. The authors use large language models as text representation models, which allow machines to read and understand textual data, and train predictive models where they use market reaction of stock returns to automatically label news as positive and negative sentiments. This method's advantage over ours is their much larger training set, which almost surely improves model performance. Training cost prevented us from doing the same analysis in our paper, but we note that in human-labelled data, most news is labelled neutral sentiment. Specifically, over 65% of all human labelled news for training in FinBERT model is neutral sentiment. The training sample in FinBERT is probably the most comprehensive freely available source as the authors use a few sources. By labelling news just as positive and negative sentiment (because price reaction is either upwards or downwards), the authors' method is noisy (as acknowledged by the authors) and may not be an accurate reflection of human being's perception of 'sentiment'. For this reason, it probably should be termed 'text-derived information'. The authors' best-performing model is slightly better than ours, but this could be due to larger sample size (hence more stocks to choose from) or indeed due to higher model performance of the in-house trained predictive model. As a direct comparison, the authors also experimented with FinBERT model, where they use it also as a language model (converting text to high dimensional vectors for machines to understand), instead of classification model, and they achieve annualized Sharpe ratios of only 1.3. We also get contradicting results on alerts and articles: while we find that models using articles achieve better performance, Jiang et al (2023) finds that alerts produce higher Sharpe ratios. As we are using different methods, we are unable to positively reconcile the difference. However, the net portfolio performance after accounting for training cost and transaction cost because of much more stocks involved in alerts portfolios may further reduce the difference between our performances.

The rest of this paper is organized as follows: section 2 presents related literature, section 3 describes our data and data cleaning procedures, section 4 presents methodology on BERT. Section 5 presents results and robustness checks, and section 6 presents conclusions and suggestions for future research.

3.2 Related literature

3.2.1 Sentiment and opinion mining in finance and economics

In this section, we give a short lookback of investor sentiment research in finance and economics. We note that investor sentiment proxies in literature include but are not limited to:

1. Investor and consumer confidence surveys. For example, Charoenrook (2005) and Lemmon and Portniaguina (2006). They usually find negative relationships between investor sentiment and future stock market return, i.e., stock price reversals. Some studies do find no statistically significant results or positive relationships between investor sentiment and stock price, such as Solt and Statman (1988), Lee et al (2002), and Brown and Cliff (2004).
2. Proxies for investor sentiment using market-wide variables, such as Baker and Wurgler (2006, 2007).
3. Use news and social media to proxy for investor sentiment. This is a large and growing literature, and we also use financial news to extract investor sentiment. We note, however, it is less clear if we extract just investor sentiment or information embedded in texts. This is a common issue found in this strand of literature. See, for example, Chen et al (2013), Ke et al (2019), Ghiassi et al (2013), Bollen et al (2011), and Li (2021).
4. Online messages boards. In the early days, Yahoo! Finance and Raging Bull tends to be popular sources of small investor sentiment, now Twitter is the main source. See, for example, Das and Chen (2007) and Tumarkin and Whitelaw (2001).

We have long known that investor sentiment based not on (at least not entirely on) rational information possesses predictive power and moves the financial market (Baker and Wurgler, 2006). Interestingly, while textual analysis in finance appears to be a new idea, efforts in exploring how textual information could help with financial decision making may be dated back to almost 90 years ago. In 1933, Cowles (1933) tried to predict stock market by subjectively categorizing Wall Street Journal articles into bullish, bearish, and doubtful sentiments. At around the same period, academics

including Keynes (1936) have realized that investor sentiment affects stock market behavior and causes asset prices to deviate from their fundamental values.

After the pioneer studies, a developed version of bag-of-words based natural language processing model as applied to finance is experimented by a few researchers. An important study is Tetlock (2007). In this paper, the author uses General Inquirer (GI), a classical content analysis tool first developed in the 1960s, to analyze the contents of *Abreast of the Market* section of the *Wall Street Journal*. This technique counts words in 77 pre-determined GI categories from the Harvard psychosocial dictionary. Tetlock then uses Principal Component Analysis (PCA) to collapse the 77 categories into a single media factor, which is highly correlated with pessimistic words in the media, hence he calls it 'pessimism factor'. This measure is predictive of market price drops which is followed by reversion to fundamental value. Unusually high or low pessimism also predicts market turnover. As Tetlock himself points out, GI is only able to distinguish between positive and negative words, while in his study, he uses only negative words. GI itself is unable to capture contexts and more complex semantic meanings beyond each word themselves. In a follow-up paper, Tetlock et al (2008) shows that language used in company-specific words can predict firms' earnings and stock returns. The author's choice of dataset is popular in literature where researchers are (often forced) to use a small dataset, such as front page or certain column of *Wall Street Journal* or other news media, because: 1) alternative dataset is unavailable; 2) there are limited techniques for extracting information from full text; and 3) it is computationally expensive to use full texts. With better computation power, state-of-the-art machine learning technique, and Refinitiv MRN, which is a comprehensive dataset, we can overcome such limitations.

Ke et al (2019) is a recent study in textual sentiment analysis. In this paper, the authors use data from Dow Jones Newswire, which contains all historical news for the US companies from 1986 to April 2020. The authors start from the view that news simultaneously affect investor sentiment and market return and propose a three-step framework: 1) as positive sentiment drives up return and negative sentiment drives down return, we can use market reaction as a guide to automatically create dictionary of positive and negative sentiment; 2) use a two-topic model to estimate positive and negative sentiment scores; 3) predict sentiment scores of news articles. With the predicted sentiment score, the authors then go long the 50 most positive stocks and go short the 50 most negative sentiment stocks each day to forma a zero-cost portfolio. The authors achieve an annualized Sharpe

ratio of 4.3 overall. This method can be considered an 'advanced' dictionary-based technique while our method is large language model based. We note that the dataset used in this paper is more comprehensive than ours and contains third-party news. While we intend to follow a similar approach and use market reaction as guide to fine-tune a sentiment classification model, we were restricted by computation power and could not perform a similar analysis. Future researchers who have sufficient computation power may follow this approach and see if and how using market reaction in additional to large language model may further improve portfolio performance. We acknowledge that the dataset used in this paper is smaller than similar studies such as Ke et al (2019) and Jiang et al (2023). While the data limitation issue cannot be overcome, we use extensive manual inspection in data cleaning process to keep as much news as possible. See section 3.3 for details on data cleaning.

3.2.2 Sentiment and opinion mining algorithms

This section provides a lookback of sentiment and textual representation methods in computer science literature. While early studies in sentiment and opinion mining in computer science literature rely on simple lexicon-based models, they were quickly replaced by newer models, even if one tries to improve lexicon models by introducing rule-based models to simple lexicons. For example, while '*good'* carries positive sentiment, '*not good*' negates the positive meaning of the word '*good'*. However, natural language is too complicated to fit in any rule-based model, unless the rule is endless. A breakthrough in word representation is Word2Vec of Mikolov et al (2013a,b), which represents word by high-dimensional vectors and captures each word's semantic meanings. However, it has two main drawbacks: 1) it is incapable of handling words that are not seen in training sample; 2) words semantically similar would still be given two completely different encodings. For studies using Word2Vec on sentiment analysis, see, for example, Zhang et al (2015).

Global Vectors for Word Representation (GloVe) of Pennington, Socher, and Manning (2014) from Stanford University is much better at handling words not seen in the training set by considering the whole corpus. Generally, word embedding tries to represent each word in a high-dimensional space, and in the process, words semantically similar are close to each other. Unsurprisingly, higher dimension usually means better semantic meanings that we can capture, but it would be more computationally expensive. Word embeddings are also the basis of encoder-decoder frameworks including BERT.

Before BERT, arguably the best algorithm in word embeddings is Embeddings from Language Model (ELMo) of Peter et al (2018), which tries to incorporate context into word embeddings. In ELMo, the same word would be given different embeddings depending on the context. This greatly improves its performance because now we can capture the same word's different meanings.

As the literature on sentiment and opinion analysis in finance and computer science is huge, we have only presented the most important and relevant studies and algorithms in this section. Interested readers may refer to Appendix 3.7.2 for a more detailed literature review.

3.3 Data

We obtain American company's historical news from Refinitiv Machine Readable News database. The sole provider for our dataset is Refinitiv (formerly Thomson Reuters). Our sample contains 24 years' news from January 1st, 1996, to December 31st, 2019. Consistent with Ke et al (2019), we keep only news with one company tag, because sentiment content and information contents of individual companies are unclear when one piece of news relates to multiple companies. Refinitiv MRN contains two types of news: title-only alerts, and articles that have body contents. The dataset contains 59.2% alerts, and the rest are articles.

3.3.1 Data cleaning

We obtained each company's intraday, open, and close prices from the Refinitiv (Thomson Reuters) Tick History (TRTH) database. The risk-free rate was approximated by the T-bill rates, which were obtained from the Fed St Louis website. We aligned our portfolio analysis with the T-bill rates by adjusting the starting period of portfolio analysis to 2001. When a piece of news relates to specific companies, the news was tagged with the companies' Reuters Instrument Code (RIC), and we used the RICs to match and identify companies. The companies' RICs are in the format of "Ticker.Exchange," where the suffix identifies the exchange. For example, "AAPL.O" stands for Apple traded on NASDAQ. To identify and filter for US stocks (i.e., find the 'Exchange' part of the RICs), we used the latest US exchange suffix provided by Refinitiv. Unfortunately, since Refinitiv does not maintain a historical list of US exchange suffixes, we lost a small number of exchanges that existed sometime during the 26-year period and disappeared at some point in time. There are a total of 4,807,623 observations from 14,214 companies, where 4,389 firms still exist today, and 7,864 companies are historical and were delisted at some point in time. A total of 1,961 companies were not found in TRTH, but they only account for 17,518 pieces (or 0.36%) of news. Such news may

include failed IPOs, data errors, Refinitiv service alerts and maintenance, and so on. After excluding the companies not identified in TRTH, our total sample contains 4,790,104 pieces of news from 12,253 companies over a 24-year period. The original time stamp in the Refinitiv MRN is in UTC time, which we converted to the NYSE exchange time to align with the trading hours.

While on average, one company has 390 pieces of news, the distribution is extremely dispersed: at least 25% of the companies have at most four pieces of news, and at least 50% of companies have at most 27 pieces of news. Most of the news items, automatically issued at around 3:40 pm and 3:50 pm each trading day, concern the NYSE order imbalance information. A typical such 'news' reads: "*NYSE ORDER IMBALANCE <F.N> 98700 SHARES ON SELL SIDE.*" While they are posted as articles, their titles and bodies are identical. Such information was first available on October 12, 2015 and accounts for 647,975 (or 13.52%) of total observations. While they are potentially useful information, they merely contain numerical order imbalance information, and there is little room for sentiment analysis. We thus removed the news items concerning the NYSE order imbalance information and were left with 4,142,129 pieces of news from 11,968 companies, where 59.16% were headline-only and the rest were articles with body contents.

Minimum	25% Quantile	Median	Mean	75% Quantile	Max
1	4	27	390	229	31403

Table 3-1 Distribution of the total amount of news per company
This table shows the distribution of total amount of news of each company in our raw sample.

Minimum	25%	Median	Mean	75%	Max	SD
1	57	80	151.5	176	10793	204.55

Table 3-2 Distribution of article's lengths in number of words
This table shows the distribution of all articles' body lengths before cleaning.

Table 3-2 shows the distribution of article's lengths in number of words. The distribution is extremely dispersed with a very long right tail: while the mean number of words is 151.5, its standard deviation is 204.5, and this is stretched by a large number of very lengthy articles. The very

short articles are polluted by the way Refinitiv posts its news. We first filtered for news articles by applying the same filtering rules for alerts-only news pieces and gained a total of 2,339,138 news pieces. Some articles were exceptionally long, while others were exceptionally short. In Figure 3-1, which shows the distribution of articles' lengths, it is evident that the distribution of articles' number of words is extremely skewed. There are many articles with unreasonably short body content; specifically, at the 35th percentile, articles' length is only 11 words.

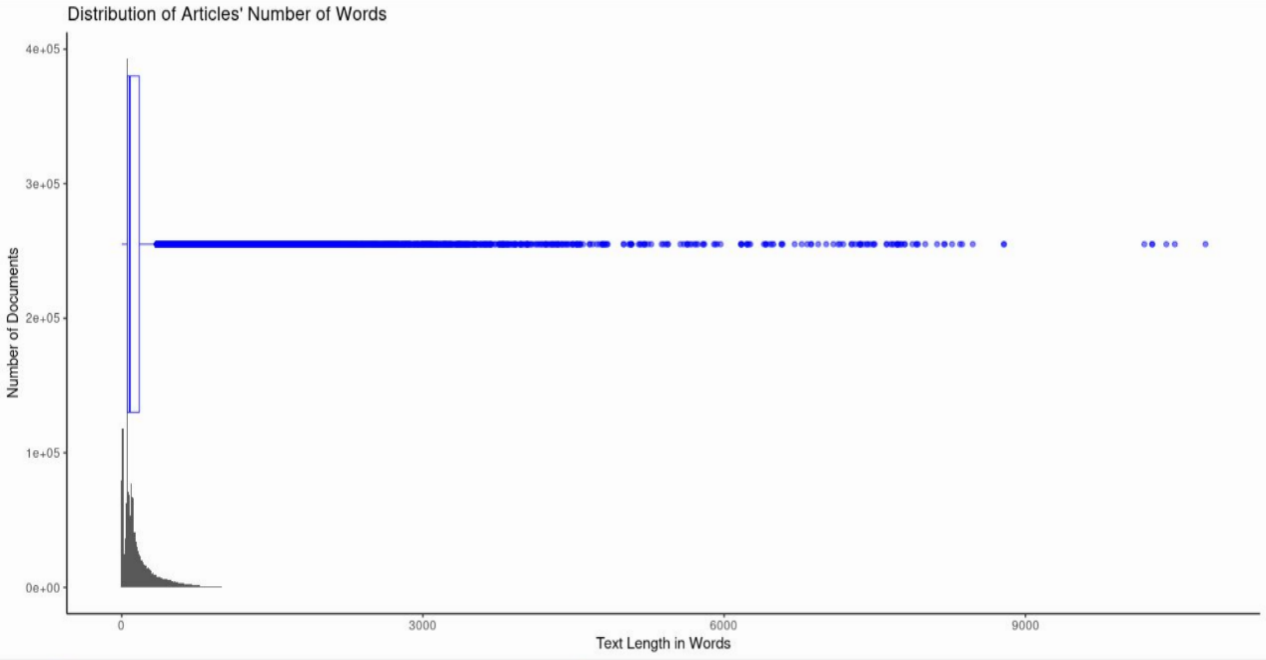


Figure 3-1 Distribution of article lengths in number of words
This graph shows the distribution of articles' lengths by plotting the boxplot and histograms of articles' number of words.

Since our dataset was considerably smaller than that of similar studies like Ke et al. (2019) and Jiang et al. (2022), we tried to keep as much news as possible using extensive visual and manual inspections. Detailed data cleaning process can be found in Appendix A.

After cleaning, we see a drastic decline in the number of articles with now only 631,521 articles after cleaning. We perform the same set of analyses as before with the FinBERT model. The processing time for articles was much longer than that of title-only news pieces.

3.3.2 Descriptive statistical analysis

Which companies attract reporter attention and have the greatest number of news articles? What are these articles about? To answer the first question, we plotted the top 30 firms' amount of news; to answer the second, we explored the topic codes. Each piece of news was tagged with a long list of topic codes covering a company's geographic location, industry, asset, events, and so on. Figure 3-2 plots the top 30 firms' amount of news. Five companies have over 20,000 pieces of news: General Motors, Boeing, Ford Motor, Citi Group, and General Electric. Interestingly, all the 30 companies are from NYSE, and 21 companies have more than 10,000 pieces of news. Not surprisingly, companies attracting the greatest number of news items are the largest ones.

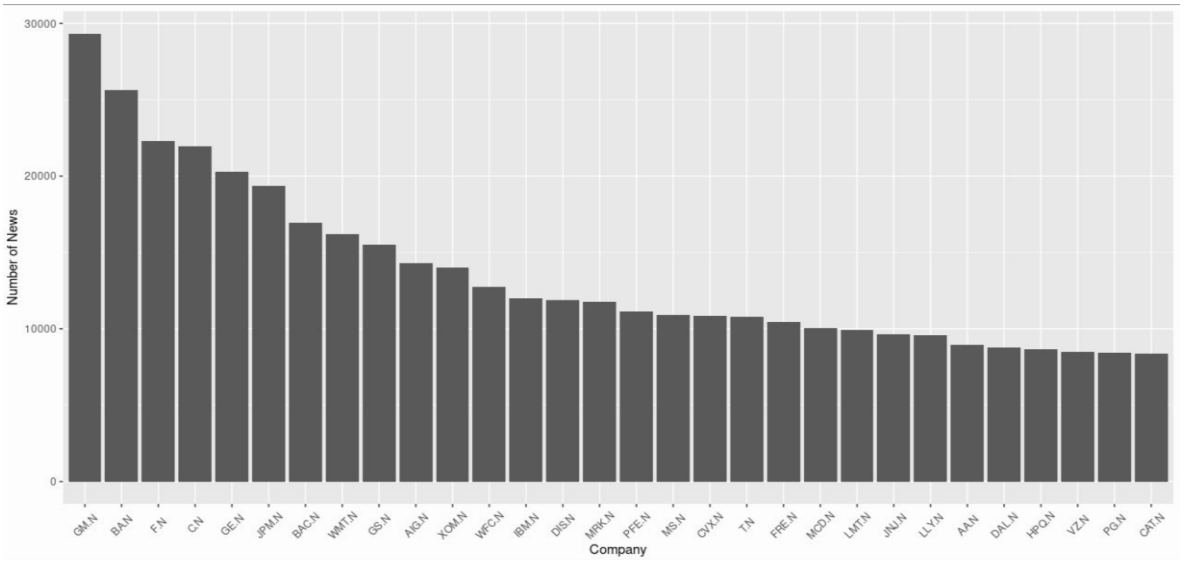


Figure 3-2 Top 30 companies' amount of news over full sample
This figure shows the number of news pieces from the top 30 companies by their RIC.

Figure 3-3 plots the top 30 topics. Since topic descriptions can be long, we kept only topic codes in this plot. The top four categories are US, America, North America, and company news. These four topics were not quite informative to us and are not plotted, as this was how we cleaned our data. The news concerned different topics, such as corporate events, mergers and acquisitions, financial events, consumer cyclicals, market events, and industrials.

We now consider how the number of news items evolve over the years, across months, and intraday. The patterns over time are largely consistent with intuition, although there are indeed some surprises.

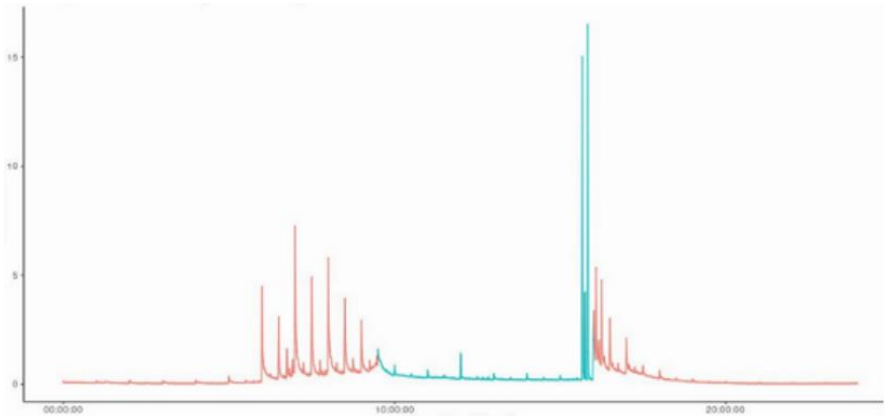


Figure 3-4 plots the average amount of news by the minute and hour of the day. The trading hours (9:30 am to 4:00 pm) are shown in blue, while the non-trading hours are shown in red. There is a very clear pattern intraday: news arrivals are more intense just before the market opens and closes, while they are calm during the day. There is a small flush of news during the middle of the day, after which the arrivals are low again. This pattern is consistent with the Dow Jones Text Feed & Archive database shown in Ke et al. (2019).

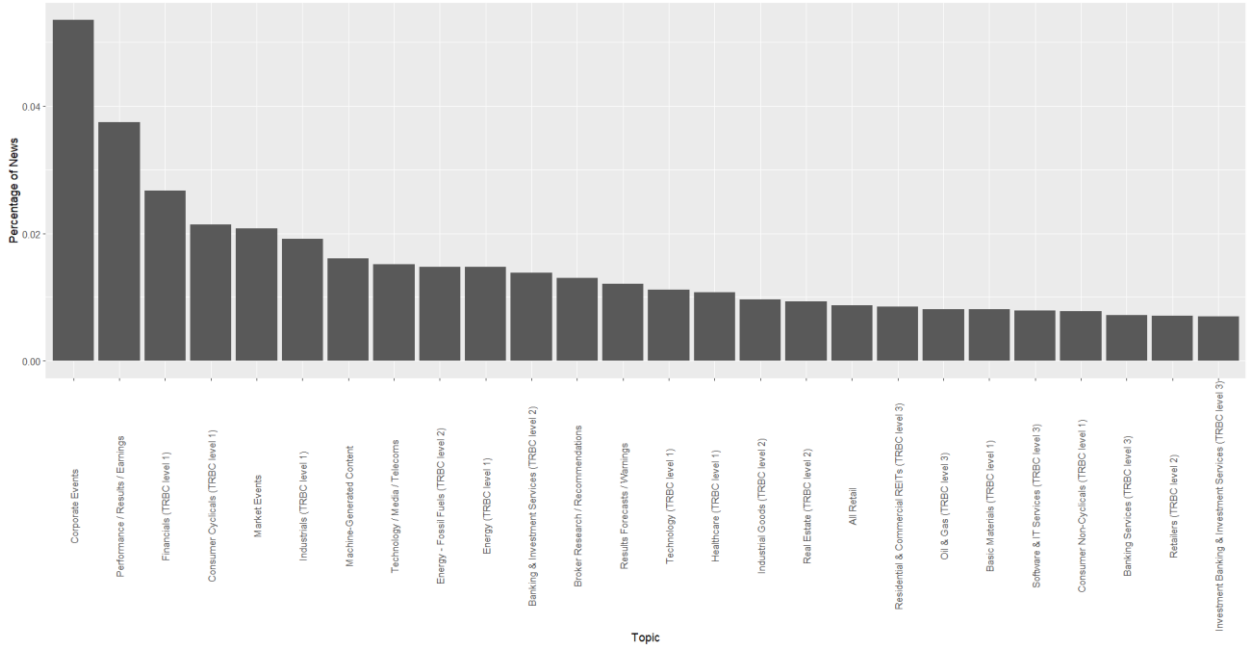


Figure 3-3 Top 30 topics in the full sample
This figure shows the top 30 topics that appeared in our sample after cleaning, including news alerts and news articles.

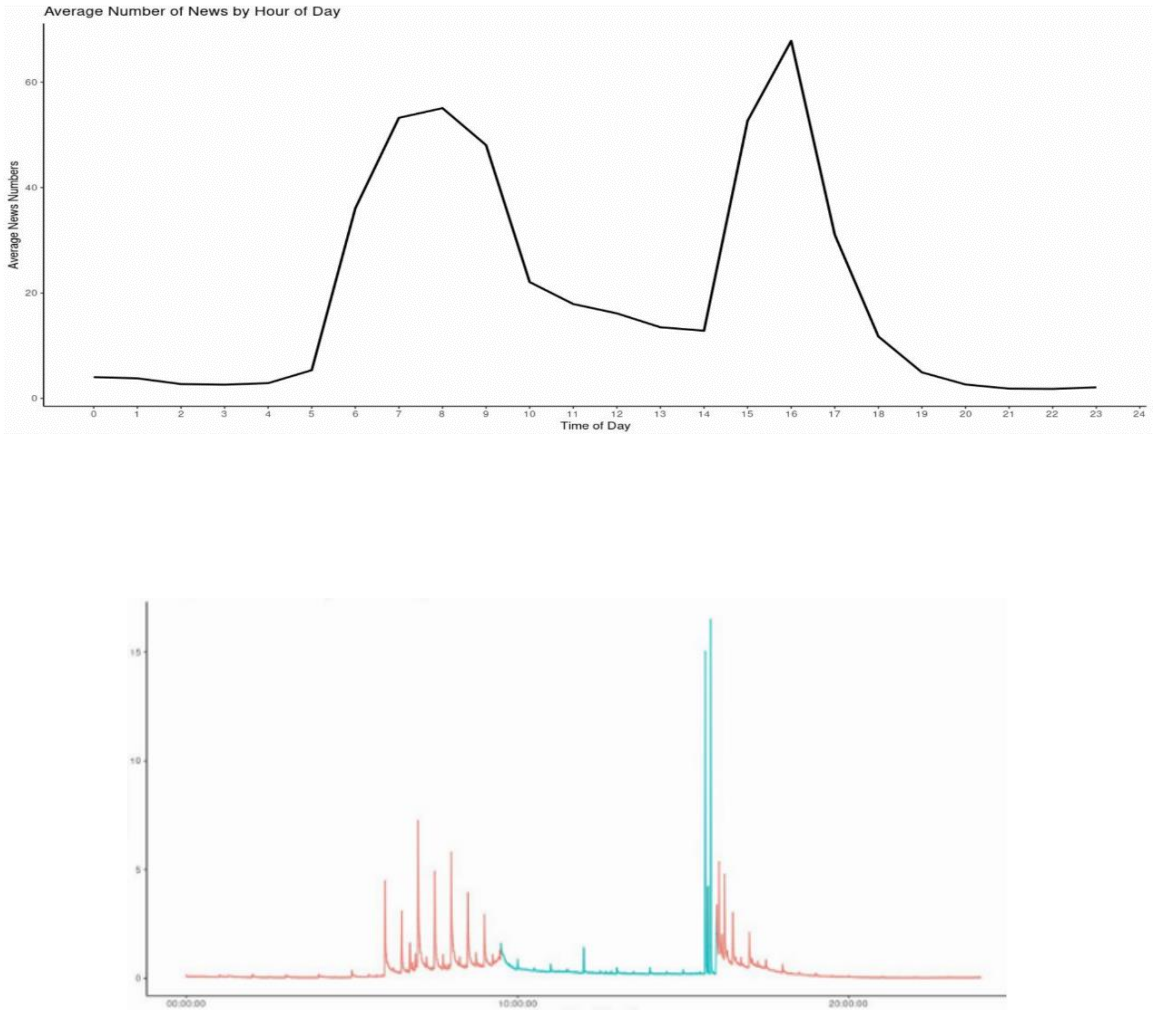
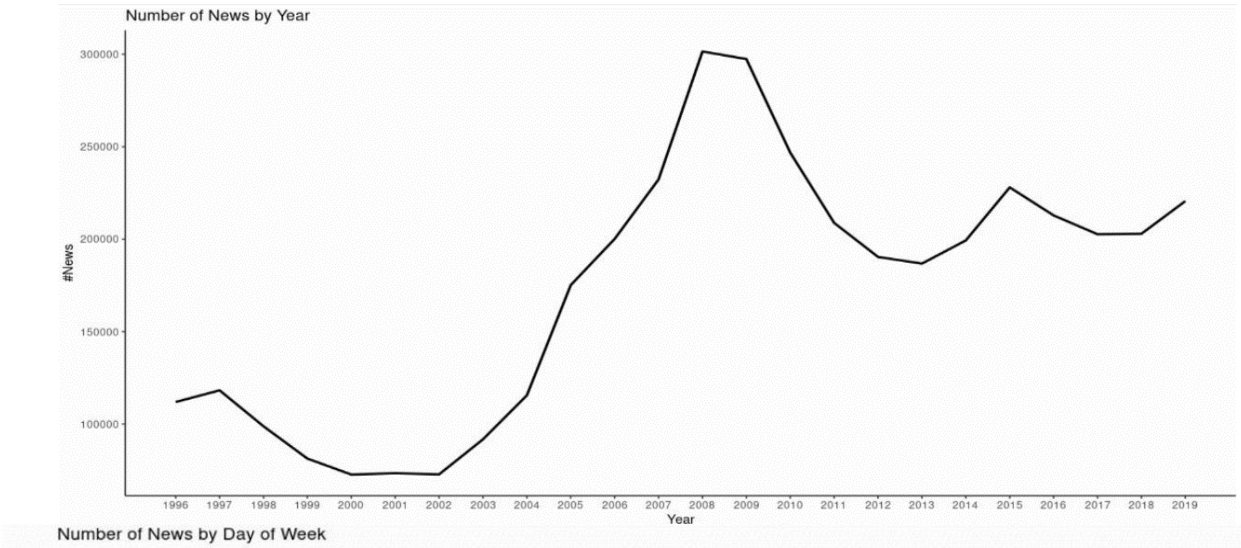
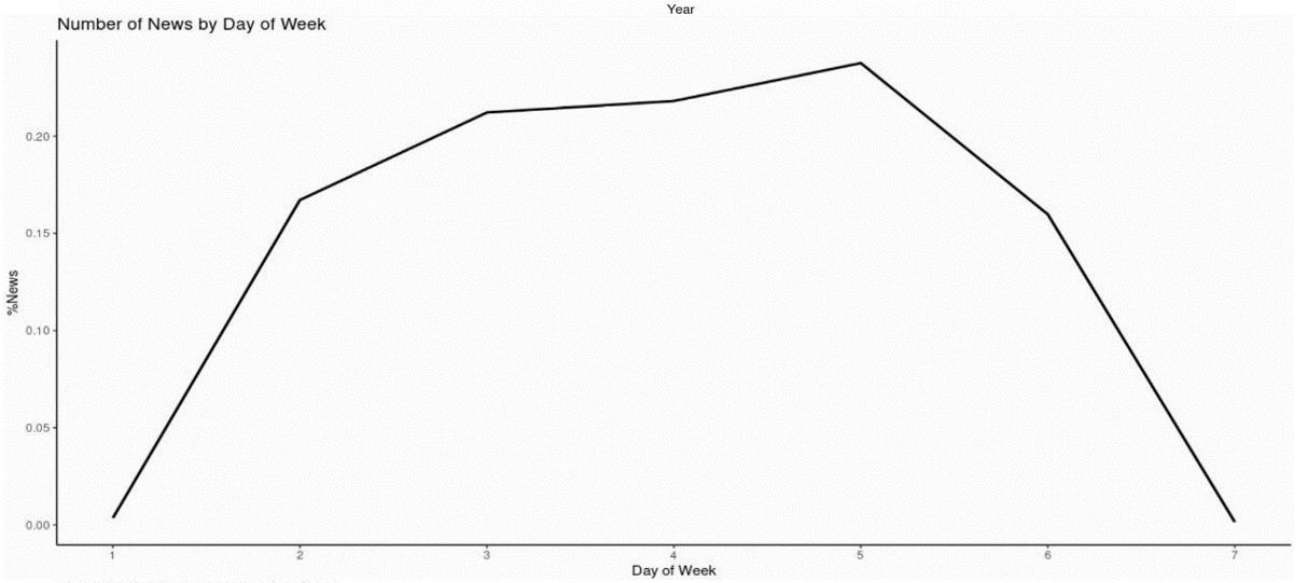


Figure 3-4 Average amount of news per hour (top) and per minute (bottom)





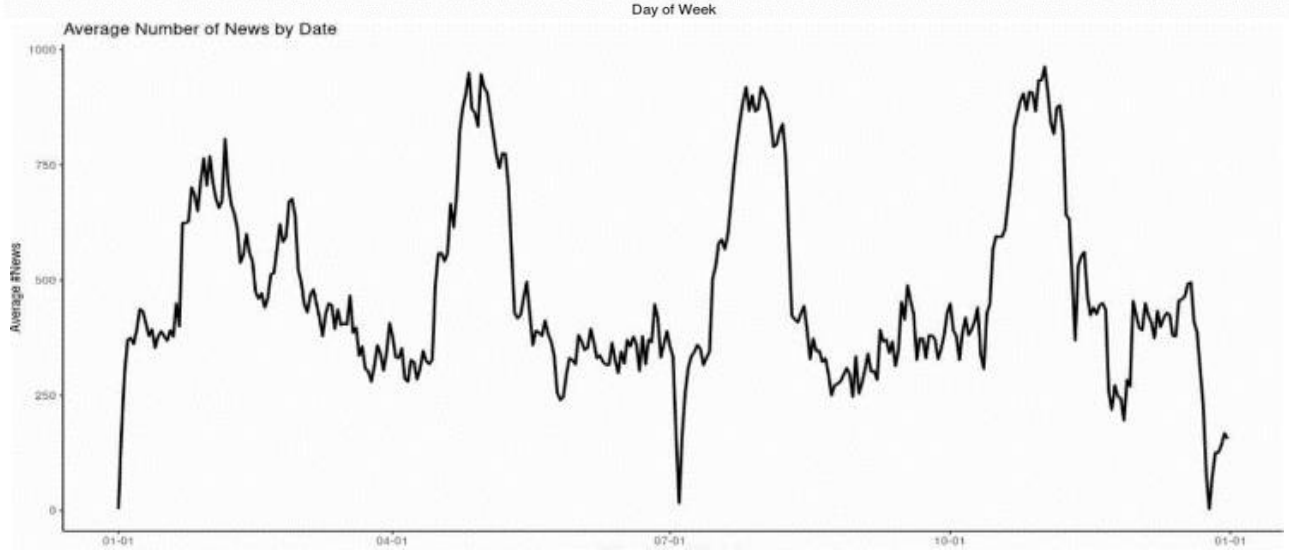


Figure 3-5 Distribution of news arrivals

This graph shows the amount of news by year (top), the percentage of news arrivals by the day of week (middle, where 1 indicates Sunday, 2 indicates Monday, and so on), and the number of news arrivals by the day of the year (bottom).

Figure 3-5 plots the number of news pieces by year, the percentage of news arrivals by the day of the week, and the number of news arrivals by the day of the year. While we can see a generally increasing trend in news arrivals, the market was especially turbulent and produced an exceptionally high volume of news during the 2008 Global Financial Crisis. During the beginning of the millennium, the market saw exceptionally low news arrivals. The months before the reporting period (March, July, September, and December) witnessed low levels of news arrivals, while the month of the reporting period witnessed a peak in news arrivals. This was expected, as we only had company-specific news in our sample and a considerable number of news reports were from or were about company events. Moreover, it is unsurprising that weekends saw minimal news arrivals. News arrivals, we found, tend to increase from Monday to Thursday and drop on Friday, when people appear to be in the weekend mode and produce less news. We also noted very strong seasonality and holiday impacts: around the reporting dates of each quarter, the market produced large amounts of news, and these effects would calm after the reporting days. New year, Christmas, and mid-year holiday seasons witnessed the lowest number of news arrivals.

3.4 Methodology

This section describes methodology used in this paper. Technical details on attention mechanism and Transformers architecture may be found in Appendix 3.7.2 for interested readers.

3.4.1 FinBERT

We know that we need to adapt any general model to domain-specific language to achieve higher performance. In other words, we need to make sure the model ‘speaks’ finance language, because the terms, jargons, et cetera used in financial documents are different from general language. In our baseline model, we use FinBERT of Huang et al (2023) which fine-tunes BERT using a huge dataset of financial documents, including: 1) 60,490 Form 10-Ks and 142,622 form 10-Qs of Russell 3000 firms during 1994 and 2019 from SEC website (2.5 billion tokens); 2) 136,578 earnings conference call transcripts of 7,740 public firms between 2004 and 2019 (1.3 billion tokens); 3) 488,494 analyst reports in the Investext database issued for S&P firms during the 1995-2008 (1.1 billion tokens).

The authors then train their model using labelled data from open sources for sentiment learning,

including Financial PhraseBank (4,845 sentences), AnalystTone (10,000 sentences) and FiQA (1,111 sentences). These sentences are human labelled into positive, negative, and neutral labels and represent sentiment contents as understood by human.

Before continuing, we note an important difference in literature using sentiment-implied returns. If we believe that after a news announcement, market participants react by doing two things:1) form sentiment; 2) react to the news, then market reaction is a natural guide to investor sentiment, and we can use market reaction as target variable to label news into positive, negative, and neutral sentiments. While this strand of study is interesting (such as Ke et al, 2019), this method intrinsically cannot identify investor reaction to information in the news and investor reaction to pure sentiment, and for this reason, it's probably safer to call this method 'soft information-based' study instead of sentiment based. Investor sentiment in computer science literature (including FinBERT of choice in this study) typically uses human-labeled texts, which is a more accurate indicator of investor sentiment. We tried to carry out analysis by fine-tuning BRET model using market reaction of news as guide but eventually could not do so due to hardware restrictions. Future studies may attempt to investigate how large language models can be used to directly explain and predict stock return directly, instead of through sentiment.

3.5 Results and robustness check

In this section, we present results of portfolio construction strategy together with robustness checks.

3.5.1 Baseline model results

We present the baseline results using the FinBERT of Huang et al. (2023) using alerts. After cleaning, we had 2,287,700 unique news alerts.

None of the training data in the FinBERT model contained Refinitiv Machine Readable News. In this sense, the predicted sentiment scores were all out-of-sample. To form a portfolio, we did the following procedure: for each trading day, we considered news arrivals from 9:00 am the previous trading day (Day t-1) to 9:00 am of the current trading day (Day t) and used this set of news to assess each company' sentiment. The trading day was aligned with the NYSE calendar. The FinBERT model has two outputs: a label (positive, neutral, or negative), and the probability of the predicted label. We converted the factor label of the FinBERT output into a numerical one as follows: if the predicted label was neutral, we assigned a score of 0; if the predicted sentiment was

positive, we assigned a score that was equal to the probability; if the predicted sentiment was negative, we assigned a score that was equal to the negative of the probability. For example, if the probability was 0.85 with a label of positive (negative), we assigned a score of 0.85 (-0.85). This was easy to use and fit our purpose: for each news item, we wished to ask how likely that it was positive. If a company attracted more than one news item in a given trading day, we then calculated a simple average of the scores.

Minimum News (N)	Days	%Sample
40	4534	98.50%
50	4517	98.13%
80	4409	95.79%
100	4297	93.35%
150	3992	86.73%

Table 3-3 Days with at least 40, 50, 80, 100, and 150 news alerts and as a percentage of total sample

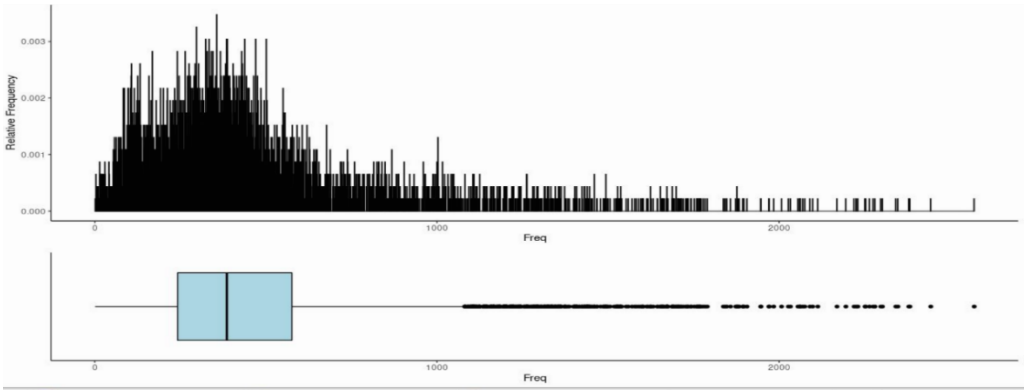


Figure 3-6 Distribution of news alerts by trading day

Table 3-3 shows the number and percentage of trading days with at least 50, 80, and 150 (X) news alerts. Figure 3-6 shows the distribution of news alerts on each trading day. Technically, we do not have a 'daily' portfolio; instead, for days with at least *m* number of news, we form a zero-cost portfolio by going long the top *n1* most positive sentiment stocks and going short the *n2* most negative sentiment stocks, where *n1* is the smaller of *N* or the number of positive or negative news if we have so few news on the day that we cannot find *N* companies for the day. We open the position at the beginning of each trading day using open price and close position the next trading day and use

open-to-open price to calculate return. For example, we form portfolios by considering only days with at least 50 news alerts. Then, for each day, we form portfolios by going long the 30 most positive-sentiment stocks by using funds from going short the 30 most negative-sentiment stocks. We close the position the next trading day and rebalance our position by re-selecting stocks into the portfolio. Gains and losses are reinvested. On days where we do not have 30 stocks that were positive (negative), we chose all the available positive (negative) stocks to form the long (short) legs of the portfolio. Such a portfolio is called a *50-30 portfolio*, and we experimented with different combinations of X-N and find the most optimal portfolio in terms of Sharpe ratio. We only consider days where we can form a long-short portfolio. In rare cases where the trading days contain only positive or negative news, we discard them from forming portfolios, such that we maintain a zero-cost portfolio.

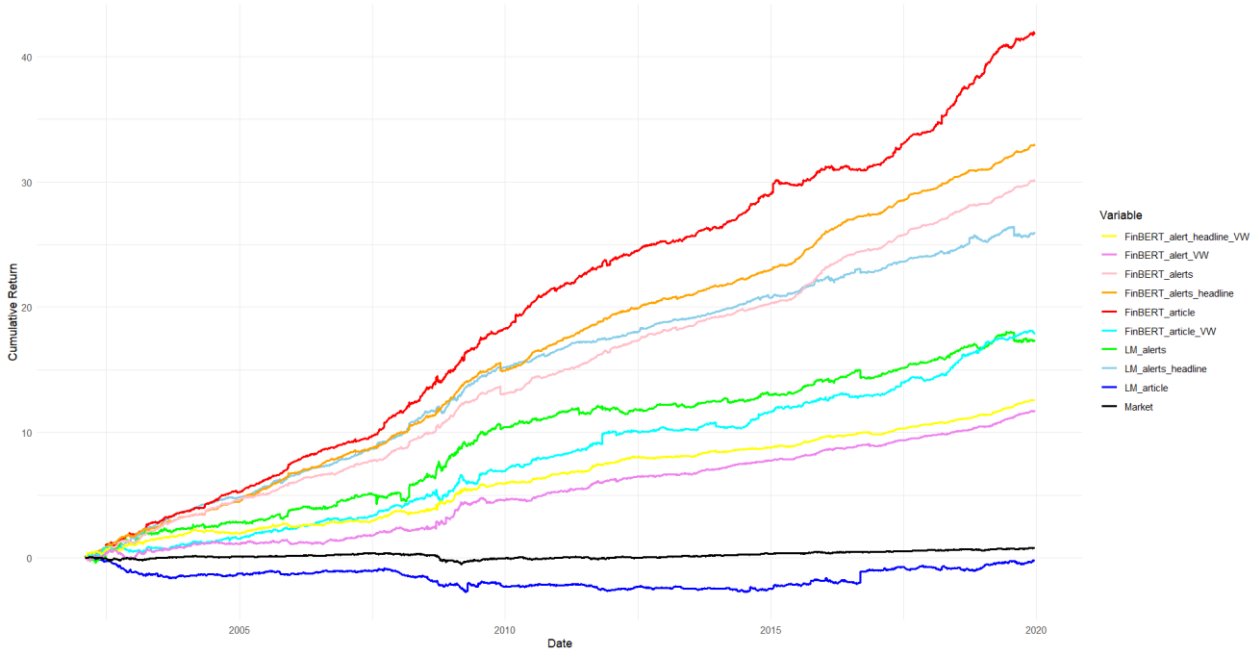
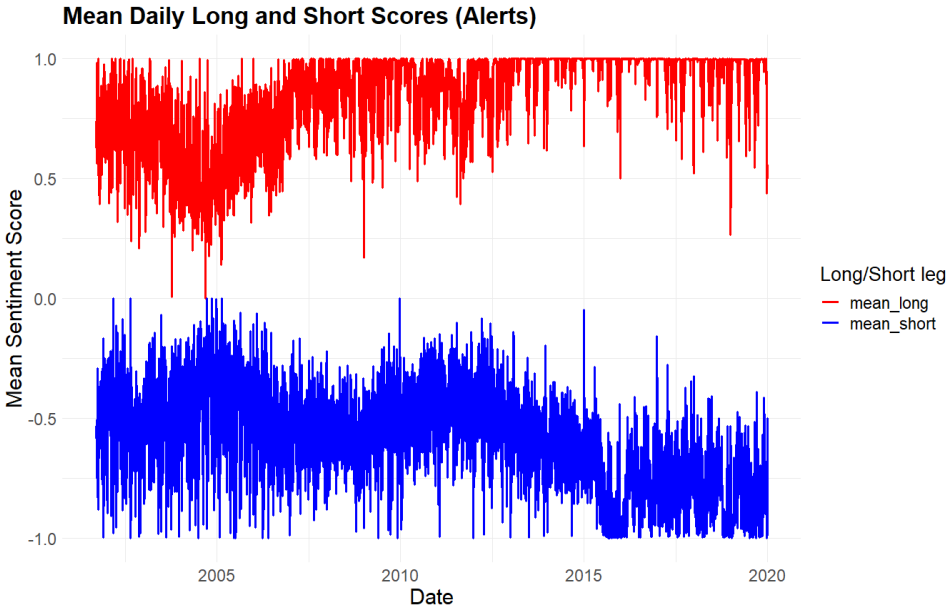


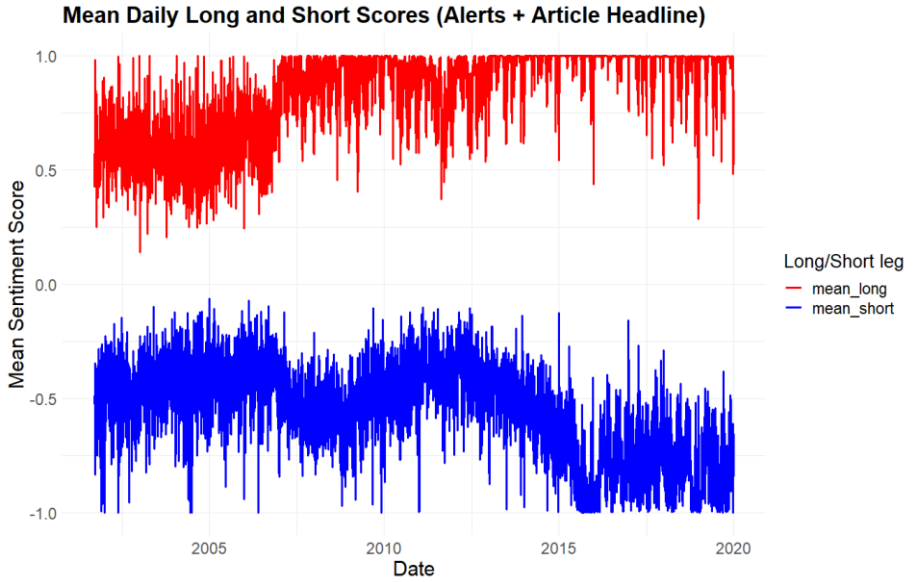
Figure 3-7 Cumulative daily log returns of S&P 500 index and sentiment portfolio

This graph shows the cumulative log returns using different models between September 2001 and December 2019. The portfolios were constructed based on the LM dictionary and the FinBERT model using alerts only, alerts and articles' headlines, and articles' body contents only. S&P 500 was included as a benchmark and proxy for a passive investment in the market. Gains and losses are reinvested in each trading day.

Figure 3-7 shows the cumulative log-returns of the best-performing (50-30) news alerts portfolio, which has an overall Sharpe ratio of 2.79. In the same graph, we also show the cumulative return of the S&P500 index and other sentiment portfolios over the same period. While equal weighting allows for higher diversification by allowing higher weights of small stocks, we also consider value-weighting. The S&P 500 index represents a passive investment that is still used actively in funds management, such as Vanguard. The baseline sentiment portfolio significantly outperforms a passive investment in the S&P 500 index. Moreover, the correlation between the daily S&P 500 index return and the sentiment portfolio return is very weak at -0.040, suggesting that portfolio returns do not seem to originate from market movement but from sentiment selection. We also found the S&P 500 index to have an overall annualized Sharpe ratio of 0.312 during this period, again suggesting the superior performance of the sentiment portfolio.

Throughout the sampling period, long legs tended to have more stocks than short legs. Specifically, in the long leg, there are 30 stocks (maximum) in 74.5% of the time, while in the short leg, the stocks are only in 26.6% of the time. The year-end periods around New Year's Eve tend to attract minimal number of stocks. This is because the days leading to New Year's Eve are not public holidays and were hence included in the sample; however, this period unavoidably attracts little investor attention and media coverage, and people are in a holiday mood. In the early part of the sampling period, both long leg and short leg witnessed a large number of days with much fewer than 30 stocks in each leg. This was also observed in Ke et al. (2019): in their early sampling periods, because of the small number of news pieces, not enough stocks were available for portfolio construction, and they were forced to use all available news for each day.


Mean Daily Long and Short Scores (Alerts)
Mean Sentiment Score
1.0
0.5
0.0
-0.5
-1.0
2005
2010
2015
2020
Date
Long/Short leg
mean_long
mean_short


Mean Daily Long and Short Scores (Alerts + Article Headline)
Mean Sentiment Score
1.0
0.5
0.0
-0.5
-1.0
2005
2010
2015
2020
Date
Long/Short leg
mean_long
mean_short

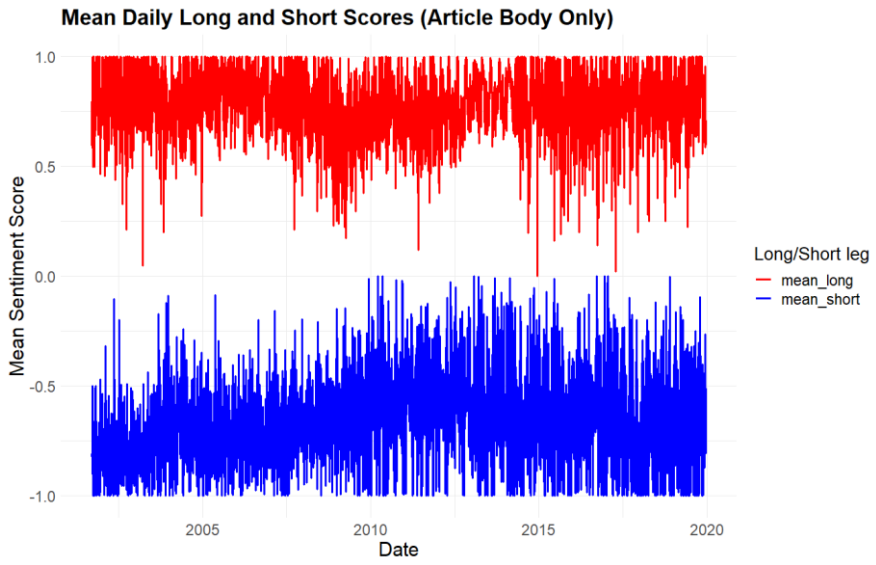


Figure 3-8 Mean score of long (red) and short (blue) legs by trading day
This graph shows the mean score of stocks in the long and short legs of the sentiment portfolio when considering alerts only (top), alerts and articles' headlines only (middle), and articles' body contents only (bottom).

Figure 3-8 shows the mean score of stocks in long (red) and short (blue) legs on each trading day. There is a generally increasing (decreasing) trend in the long (short) leg's sentiment score. There are two possible explanations to this observed phenomena: 1) because the FinBERT model was trained on relatively recent financial documents, they are naturally more applicable to recent years' news, as the jargon and vocabulary used in financial news are changing over the years, even if the language is in the same domain of finance; 2) over time, we see a larger number of news items released on each trading day. In the early days, since news items were very few, researchers were forced to choose stocks whose sentiment scores were not high (or low) enough. Both legs do exhibit a fair variation, indicating that tone of news is simply more extreme in some days than others. Moreover, positive and negative legs' mean sentiment score have a correlation of -0.45, which is moderately negative. This seems to suggest that the tone, or investors' sentiment after reading the news, is generally more extreme than moderate. In the long leg, there are significantly more days with sentiment scores closer to 1 than in the short leg with sentiment scores closer to -1.

The correlation between the daily portfolio performance and the total number of stocks in the portfolio, the number of stocks in the long leg of the portfolio, and the number of stocks in the short leg of the portfolio are very weak at -0.025, -0.056, and -0.028, respectively, suggesting that increasing the number of stocks does not seem to improve performance.

3.5.2 Articles

We now consider news articles in addition to headline-only news. We have noted that daily number of news is often a restricting factor especially during early days where we see significantly fewer news produced, this problem may be more severe for news articles because we have much fewer news articles. We first look at a comparison between news alerts and news articles' predicted labels. Table 3-4 shows the percentage of predicted positive, neutral, and negative labels for news articles and news alerts. We note that there is no economically significant difference between predicted labels in news articles and news alerts and that most news are predicted to be of neutral tone. This is consistent with previous studies where neutral sentiments dominate.

Label	Negative	Neutral	Positive
%Alerts	0.105	0.686	0.209
%Articles	0.108	0.669	0.222

Table 3-4 Percentage of Alerts and Articles Body's Predicted Labels.

Next, we ask: is informational contents of news articles reasonably captured in their titles? We perform two additional sets of analysis: one with all news alert headlines and news articles' body contents, and one with articles' body contents only.

We compare the same 50-30 portfolio based on news alerts only versus news alerts and news articles put together. We achieve overall Sharpe ratios of 2.79 (headline-only), 3.09 (headline-article), and 3.87 (article body only) respectively, a significant increase from pure alerts-based results. We note that as with other deep learning models, it is slow to run[5] and running speed grows almost exponentially with text length, and this may justify practitioners' decisions if they rely on headlines only. However, the performance improvements when using articles' body contents justify researchers' and practitioners' investments into computing power. The results are presented in

[5] Articles' body contents after cleaning took around one full week to run on High Performance Computing to get their predicted labels.

Figure 3-7 and Figure 3-8. While the mean scores of long and short legs of portfolio exhibit the same trend under the 3 portfolio strategies, there are significantly fewer stocks in each leg in portfolio relying on article body only because of much fewer news articles. However, we see much higher portfolio performance with fewer stocks, which translates to lower transaction costs, under article body-only model.

Our results have important implications for investors and algorithm traders: while it is believed that faster reaction to information is desirable and investors should try to improve their speed of reaction, it is probably not the case with real-time text feed. The significantly higher Sharpe ratio using articles' body contents and disregarding alerts suggest that with textual news data, it is advisable to wait until more information is released, and reacting too fast based on incomplete information may reduce one's profitability.

3.5.3 Breaking down portfolio return

	Alerts only	Alerts - headline	Article Body
Panel A			
Mean excess %return	0.75	0.81	1.08
Sd (%)	4.22	4.12	4.32
Sharpe Ratio (Annualized)	2.84	3.11	3.96
% Profitable days (excess return)	68.19	70.54	64.28
Panel B			
Mean excess % return (long)	0.2	0.2	0.2
Sd (% excess return) (long)	1.9	1.9	2.4
Mean excess %return (short)	0.6	0.6	0.9
Sd (%excess return) (short)	4.4	4.3	4.2
Total amount of news	**2,287,700**	**2,919,221**	**631,521**

Table 3-5 Portfolio Sharpe ratios (equal weighted)
Panel A shows the daily mean excess return, the standard deviation of the excess ratio, the Sharpe ratio of the portfolio, and the percentage of profitable days under the FinBERT model. Panel B shows the average and standard deviation of the daily portfolio returns by the long and short legs.

	Alerts only	Alerts - headline	Article Body
Panel A			

Mean excess %return	0.32	0.33	0.46
Sd (%)	3.63	3.63	3.62
Sharpe Ratio (Annualized)	1.38	1.45	2.01
% Profitable days (excess return)	58.79	58.91	29.66
Mean excess % return (long)	0.17	0.11	0.12
Sd (% excess return) (long)	1.70	1.75	2.38
Mean excess %return (short)	0.24	0.24	0.38
Sd (%excess return) (short)	2.52	2.34	3.73
Total amount of news	**2,287,700**	**2,919,221**	**631,521**

Table 3-6 Portfolio Sharpe ratios (value weighted)
Panel A shows the daily mean excess return, the standard deviation of the excess ratio, the Sharpe ratio of the portfolio, and the percentage of profitable days under the FinBERT model. Panel B shows the average and standard deviation of the daily portfolio returns by the long and short legs.

Table 3-5 and Table 3-6 show the portfolio's Sharpe ratio by breaking it down to daily excess returns and standard deviations in Panel A, and profitability by portfolio's long and short legs in Panel B. As we move from forming a portfolio using alerts to using articles' body contents, the portfolio's volatility remains relatively constant while the portfolio return improves, leading to improved Sharpe ratio. Across all portfolio strategies, equal weighting outperform value weighting. However, the proportion of profitable days does not improve: using alerts only yields 67.97% profitable days, while using articles' body contents only yields 63.73% profitable days. This is likely to be the direct result of the smaller sample size of articles. The alerts-article headline has the largest sample size, which includes both news alerts and news articles. However, relying on articles' body content allows higher overall profitability despite fewer days with positive excess returns. From Panel B, it is evident that, while the negative leg of the portfolio yields higher returns than the positive legs (3 to 4.5 times higher), they are also more volatile, and the standard deviation of the short legs' returns are about twice as large.

We also consider how the portfolio strategy compares when rebalancing the position weekly. We rebalance on Tuesday. The choice of rebalancing day is arbitrary, but is chosen to avoid Monday, which sees some public holidays, and Friday, which exhibits Friday effects. The portfolio performance is shown in Table 3-7. Unsurprisingly, portfolio performance is much lower than daily rebalancing. The informational content in news tends to be short-lived as fresh news become

available, and investors react to the latest sets of news. Advances in automated trading techniques facilitate the process.

	Alerts only	Alerts - headline	Article Body
Panel A			
Mean excess %return	-1.04	-0.95	-1.71
Sd (%)	9.05	8.73	10.82
Sharpe Ratio (Annualized)	-0.83	-0.78	-1.14
% Profitable days (excess return)	50.5	50	44.82
Panel B			
Mean excess % return (long)	-1.04	1.07	0.74
Sd (% excess return) (long)	4.83	4.45	6.21
Mean excess %return (short)	-2.1	-2.07	-2.65
Sd (%excess return) (short)	6.07	5.85	7.47
Total amount of news	**13,297**	**14,671**	**7,071**

Table 3-7. Portfolio performance of weekly rebalanced portfolio.

	Alerts	Alerts - Headline	Article Body
(Intercept)	0.743***	0.797***	1.036***
	(0.063)	(0.061)	(0.065)
Market	-0.061	0.018	-0.048
	(0.061)	(0.060)	(0.063)
SMB	-0.064	-0.020	-0.229
	(0.116)	(0.113)	(0.119)
HML	-0.336**	-0.366***	-0.476***
	(0.112)	(0.109)	(0.114)
RMW	0.034	0.192	0.098
	(0.167)	(0.162)	(0.170)
CMA	-0.140	0.010	0.064
	(0.210)	(0.204)	(0.215)
Information Ratio	0.177	0.195	0.245
R^2	0.004	0.004	0.008
Adj. R^2	0.003	0.003	0.007
Num. obs.	4496	4530	4287

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3-8 Fama-French 5-factor model for FinBERT model daily excess returns.
This table shows regression results by regressing daily portfolio returns (in percentage) on Fama-French 5-factors.

Table 3-8 shows Fama-French 5-factor model results of the 3 FinBERT models using daily returns where daily excess return is in percentage. We see that the only significant factor is HML, and the three models consistently earn significantly positive alpha, suggesting that the sentiment portfolio's returns are explained by the return differences in value and growth stocks, and the consistently negative coefficient of HML suggests that the portfolio is sensitive to growth stocks. Overall, sentiment portfolios achieve excellent alpha compared with Fama-French 5-factor model benchmark.

We have not considered transaction cost in stock returns. As a guide, when daily transaction cost is 0.7%, 0.8%, and 1.0%, the portfolio ceases to be profitable under alerts, alerts-headline, and article body strategies.

As a guide, when the daily transaction cost is 0.7%, 0.8%, and 1.0%, the portfolio ceases to be profitable under alerts, alerts-headline, and article body strategies.

3.5.4 Robustness check: dictionary methods

In this section, we compare FinBERT model's performance with dictionary methods. To be parsimonious, we perform sentiment analysis on headlines. Specifically, we use 1) the Harvard-IV4 dictionary, a general-purpose dictionary and 2) the Loughran-McDonald dictionary specifically designed for financial data. Sentiment scores based on dictionary method gives a continuous number from -1 (negative sentiment) to 1 (positive sentiment) based on polarity, where:

$$Polarity = \frac{Pos - Neg}{Pos + Neg}$$

Polarity is a simple but widely used metric in lexicon-based natural language processing models where Pos and Neg represent word counts of positive and negative words, respectively. Following standard pre-processing procedures, the headlines are stemmed, converted to lower cases, and removed of stop words before conducting dictionary-based models.

Correlation between different models

	LM	HIV4	FinBERT
LM	1	0.268	0.327
HIV4		1	0.012
FinBERT			1

Table 3-9 Correlations between LM model, HIV4, and baseline FinBERT model sentiment scores
This table shows the percentage of positive, negative, and neutral labels under FinBERT model for news alerts and news articles (body contents).

We start by noting that correlations between sentiment scores from the LM dictionary, Harvard-IV4 dictionary, and FinBERT models are very low as seen in Table 3-9. The general dictionary, HIV4, has very low correlation with FinBERT model and moderate correlation with the LM dictionary. While just moderately positive, FinBERT model and LM dictionary sees a higher correlation at 0.327, suggesting that the LM dictionary, which is tailored to financial text, captures financial sentiment better than the general dictionary of HIV4.

We then perform the same portfolio construction exercise as before. Using LM dictionary, we achieve Sharpe ratios of 1.59, 2.94, and 0.04 under news alerts, news alerts and article headlines, and article body contents, significantly lower than FinBERT model. Cumulative returns under LM dictionary are plotted in Figure 3-7. A striking feature is that when using only article body contents, LM dictionary and FinBERT gives the lowest and highest performance, respectively. While including news articles' headlines to news alerts significantly improves portfolio performance, the performance drastically drops when using article body contents only, right the opposite of what we observe in FinBERT model performance. This suggests that dictionary methods are not good at identifying sentiment and information contents from long and complicated texts. Using article body contents under LM dictionary would even give inferior performance than the market.

3.5.5 Robustness check: filtering news with low word-symbol ratio

So far, we have retained as much news as possible because of our smaller dataset compared with similar studies such as Ke et al (2019) and Jiang et al (2023). This contributed to the high volatilities of our portfolio. We have also used extensive manual checks and inspections for data cleaning to minimize the amount of news taken away. Due to the large amount of news, we acknowledge that our manual data cleaning is far from competent and exhaustive.

We note one critical point in terms of data cleaning: essential data cleaning may be minimal when using a pre-trained Transformer-based model. The rationale is simple: Transformer-based models

are first and foremost language representations. By inputting and interpreting textual news, it simply uses all information from its training set to interpret the text. Specifically, we expect FinBERT to correctly interpret (or *encode*) most news because FinBERT is trained on and is designed to interpret financial texts. Some news, however, may not contain textually relevant contents, such as tables (which contain mainly numbers) and certain alerts designed to warn readers of some upcoming events. Ex-ante, if the fine-tuning set of FinBERT contains such texts, then FinBERT can adequately interpret such news and hence adequately classify their sentiment contents. We do not have concrete evidence because it would be impossible to really tell if FinBERT's financial text inputs exhaustively cover the way language is used in Refinitiv MRN because of the large amount of data. However, as long as the text sources are different, some news in Refinitiv MRN would not be adequately accounted for (or *encoded* and *modelled*) by FinBERT, such as tabular data. As classification model essentially asks: how similar is the given text compared with the positive and negative sentiment texts learned in the my training phase? Such texts of unusual format or containing too little textual information would hence be classified as neutral, or giving low sentiment scores because the text is *not likely* to be of positive, neutral, or negative sentiment. As our portfolio construction looks for the most positive and negative sentiment news, such news would hence not be selected into our long and short legs.

We test this intuition by excluding news articles of low word-to-non-word ratios. We identify non-words by matching words starting with symbols, special characters, and numbers. Figure 3-9 shows the distribution of non-words in article body contents. While most articles seem to have a reasonable non-word ratio, some articles do exhibit too high non-word ratios. Without a guide on '*reasonable'* non-word ratios, we exclude articles with higher than 0.2 (20%) non-words. By doing so, we further filter away 48,781 articles.

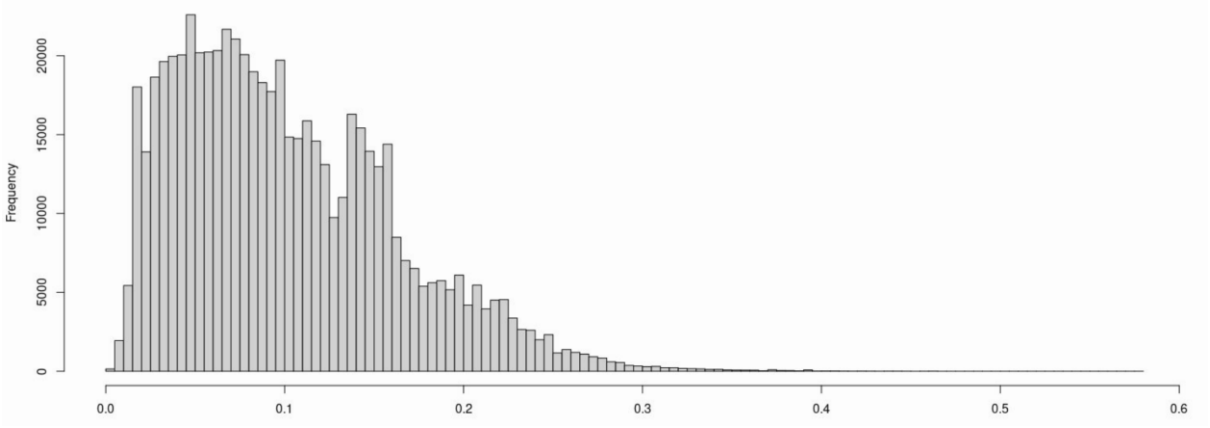


Figure 3-9. Distribution of non-word ratios in article body contents

Unsurprisingly, overwhelmingly much of the news (75.7%) is neutral sentiment, with 20.2% positive sentiment and only 4.1% negative sentiment among all news in the high non-word ratio news. While this provides partial evidence that news with high non-word ratios is potentially less informative in terms of their sentiment contents, it also confirms our ex-ante expectation: a huge proportion of such news would not have been selected into portfolio anyway. Excluding such news does not improve our portfolio performance. Portfolio Sharpe ratio with the smaller dataset drops slightly to 3.68. This is probably the direct result of the smaller dataset and hence fewer stocks to choose from, which at least reduces portfolio diversification. It is also possible that some news in the high non-word ratio articles still carry strong sentiment contents and would help us form portfolios.

3.5.6 Heterogeneity in news of different sentiments

In this section, we ask: is there heterogeneity in news? If news only differs in their informational and sentiment contents, then there should be no difference in anything but their contents. We test using a smaller sample in this section[6]. Specifically, we consider 2014-2019. We require, on average, a stock to have at least 30 company-days per year to include it into our analysis. This leaves us 44

[6] This section entails the use of high-frequency data, and we would easily exceed our faculty's download limit if considering all stocks.

companies in articles and 190 companies in headlines. The 44 companies are large companies in financial, manufacturing, automobile and airline, retail, food, consumer goods, high-tech, and financial industries and the list is shown in Table 3-10. We use the 44 companies that appear both in news articles and news alerts for analysis. 3 companies are now delisted: Raytheon Co (RTN.N), Twitter (TWTR.N), and United Technologies Corp (UTX.N).

RIC	n (articles)	n (alerts)	Company
BA.N	2834	5655	Boeing Company
GM.N	1977	3860	General Motors Company
TWTR.N	1874	2804	Twitter
JPM.N	1440	3284	JPMorgan Chase & Co
GS.N	1427	2549	Goldman Sachs Group
F.N	1395	3124	Ford Motor Co
WFC.N	1121	3255	Wells Fargo & Company
WMT.N	1056	2614	Walmart Inc
GE.N	1003	3585	General Electric Company
C.N	1000	3029	Citigroup Inc
XOM.N	991	2511	Exxon Mobil Corporation
BAC.N	822	2201	Bank of America Corporation
LMT.N	762	1988	Lockheed Martin Corporation
BLK.N	749	2405	BlackRock Inc
JNJ.N	688	2390	Johnson & Johnson
DIS.N	676	1636	Walt Disney Co
MRK.N	644	2431	Merck & Co Inc
MS.N	642	1873	Morgan Stanley
PFE.N	632	2550	Pfizer Inc
CVX.N	586	1998	Chevron Corp
ICE.N	566	1872	Intercontinental Exchange Inc
DAL.N	553	2363	Delta Air Lines Inc
T.N	544	2235	AT&T Inc
LLY.N	538	2486	El We Lilly and Company
NYT.N	531	702	New York Times Co
PCG.N	524	1787	PG&E Corp
MCD.N	513	1741	McDonald's Corporation
NKE.N	508	1543	Nike Inc
UAL.N	475	1669	United Airlines Holdings Inc
AIG.N	435	1653	American International Group
BX.N	435	1275	Blackstone Inc
CAT.N	427	2385	Caterpillar Inc

FCX.N	401	1798	Freeport-McMoRan Inc
VZ.N	400	1773	Verizon Communications Inc
TGT.N	383	1715	Target Corp
BMY.N	375	1955	Bristol-Ourers Squibb Company
LUV.N	361	1857	Southwest Airlines Co
KO.N	348	1499	Coca-Cola Co
UTX.N	337	1597	United Technologies Corp
UPS.N	336	1582	United Parcel Service Inc
IBM.N	334	1598	International Business Machines Corporation
MDT.N	317	1774	Medtronic PLC
BK.N	309	1227	Bank of New York Mellon Corp
RTN.N	302	1274	Restaurant Group PLC
Total	**32,571**	**97,102**	

Table 3-10 List of companies in news heterogeneity analysis.
This table shows the list of companies that we use in news heterogeneity analysis with high-frequency data.

		Length of Article Body	Number of Audiences	Number of Topics	Percentile of RV
alerts	Negative		3.7	23.3	73.2
	Neutral		3.6	23.9	64.5
	Positive		3.3	21.8	71.3
	F-Stat		174.2	203.1	697.1
article body	Negative	191	8.6	32.5	70.8
	Neutral	175	7.9	31.8	56.8
	Positive	187	7.5	28.1	76
	F-stat	43.35	90.6	84.3	965.2
article title	Negative		7.6	34.1	66.2
	Neutral		8.1	31.2	57.6
	Positive		7.6	29.3	70.4
	F-stat		32.8	48.3	433.2

Table 3-11 News heterogeneity
This table breaks down news length, news' number of audiences, number of topics, and each news-day's realized volatility percentile by the sentiment of the news.

We table the differences in length of article body, number of subjects, and number of audiences for alerts, news articles' headlines, and news articles' bodies together with their F-statistic in Table 3-11. All p-values are lower than 0.00 where we refrain from showing in the table to be parsimonious. We note that news does appear to exhibit heterogeneity beyond their contents, as there are economically and statistically significant differences in positive, neutral, and negative sentiment news' length,

number of subjects, and number of audiences the news is tailored to. Specifically, news of positive sentiment is tailored to fewer audiences, contain fewer topics, and are generally shorter.

We also check if they differ in volatility. To do so, we compute 5-minute realized volatility for each stock-day and compute the percentile of realized volatility from 20 days before and 20 days after each day. This hence can be considered a 'local percentile' of volatility. We see that there is also heterogeneity in volatility percentiles. It hence appears that positive, neutral, and negative news are intrinsically different and warrant investigation on their own. Previous studies which consider only sentiment-carrying news, i.e., positive and negative news only, potentially gives up valuable information. While the result on realized volatility is probably intuitive, it's probably surprising that there are statistically different length of article, number of news audience, and number of subjects in positive, neutral, and negative news.

3.5.7 What is the news talking about?

Figure 3-10 Word Clouds of news articles in long (left) and short (right) leg of portfolio

We present word clouds of news used in positive and negative legs of portfolio in Figure 3-10. Larger fonts indicate more frequent words. Note, however, this is different from sentiment words in traditional dictionary methods: words in dictionary-based methods are sentiment-charged, which are the only words the algorithm looks for when assessing documents' sentiment contents. The word cloud we present here concerns all words in a document and are not '*sentiment-charged*'. Instead, they give us a general impression of *what the news talks about*. We perform usual pre-processing

step by removing stop words, digits, turning words to lower-cases, and manually remove some common but meaningless words (such as ‘Reuters’).

We note that overwhelmingly many of the news pieces used in portfolio constructions talk about share prices and companies’ profitability, which is intuitive. They carry clean identification of news sentiment and is usually what moves stock market. Investor sentiments based on investor reactions to such news are carried forward to the next trading day, making it a profitable opportunity to trade on previous days’ sentiment. This, however, again suggests market inefficiencies: market doesn’t absorb all information into asset prices adequately and timely, leading to profitable opportunities using past news.

3.6 Conclusions and further research

Written in 2021, we are the first to investigate how LLMs help in portfolio construction and explain the portfolio’s performance with risk and behavioral factors. Deviating from previous textual analysis methodologies in finance, which are generally dictionary-based and reduces news (or generally, documents) to word lists of positive and negative words, transformer-based models can understand the semantics, grammar, context, and inter-connections among the words, hence greatly improving the algorithm’s ability to understand natural language. We use Refinitiv MRN database which consists of all historical North American company news from Refinitiv (Thomson Reuters) from 2001 to 2019. This contains much more information than previous studies, which typically use headlines of a small number of stocks’ news or certain columns of Wall Street Journal to extract sentiment information.

The comprehensive dataset allows us to explore features of news arrivals, which is also important for cleaning and structuring dataset. We observe strong seasonality of news arrivals where there are four ‘waves’ of news arrivals throughout the year, consistent with four reporting periods. News production is higher around seasonal reporting periods and significantly drop after that. During the trading day, news arrivals are significantly concentrated around market close, and news production during the day in other periods are (probably surprisingly) lower than market close period.

Our results show that sentiment portfolio significantly outperforms a passive investment in S&P 500 index and LM dictionary. The daily performance of our portfolio and market return exhibit very low correlations, suggesting that the sentiment portfolio’s return is not from general market movements. Using only article body contents significantly improves portfolio performance. While BERT models

(including FinBERT) are slow to run even on High-Performance Computers (HPCs), it's worthwhile to consider body contents of news pieces instead of using headlines only. Probably surprisingly, reacting too fast on incomplete textual information may be a bad idea as this reduces portfolio performance. This is good news to smaller investors: not having access to fast algorithms and forming portfolios based on a small number of complete news yields significantly better portfolio performance than reacting to every single piece of news that arrive to the market, hence significantly reducing their cost of trading while improving portfolio performance. Not surprisingly, BERT-based model significantly outperforms dictionary-based method.

An interesting finding is that there are economically and statistically different number of topics, audiences, and news length among news of different sentiment. We incorporated these variables in additional to common risk factors and seasonality proxies in explaining daily portfolio returns. We found seasonality, as a proxy of investor attention, holiday effects, and most common risk factors do not explain the daily sentiment portfolio's returns, while news complexity does, consistent with our ex-ante expectation. In future research, researchers may attempt to explore causes of this finding and attempt to investigate topic information in news.

We attempted to use market reaction as guide on news sentiment and train an in-house BERT model in addition to FinBERT model we employ in this study. However, hardware limitations prevented us from carrying out this practice. Future researchers who have sufficient computing power may attempt to train a regression model using market reaction of the stocks following the news as target variable, and this may potentially improve portfolio performance.

3.7 Appendix

3.7.1 Appendix A. Detailed Data Cleaning Process

To attain a sense of what the articles are about, so as to see if there were any ‘patterns’ or ‘series’ of news articles, we manually investigated short articles and noted the following features of the database: 1) most of the unreasonably short articles with one or two words concern information regarding the NYSE indication—such as the last, bid, and ask prices of a stock—with the body of the articles containing one or two numbers; 2) articles that are 11 words long are predominately about the NYSE order imbalance information that is automatically posted at the end of the trading day and about NYSE indications, which contain very similar information as the one- and two-word ‘articles,’ where their body just restates their headlines; 3) articles between 10 to 20 words are mainly reminders of upcoming events, such as corporate news announcements; 4) articles between 20 and 30 words are often brief notes to notify readers that there have been duplicated news items and that certain news should hence be disregarded; they contain links to the news press that the Reuters terminal users may access and include initial public offerings (IPOs) and seasoned equity offerings (SEOs); their body is unstructured and simply contains, for example, the issuer and the issued price. 5) articles between 30 and 59 words also contain predominantly bonds and equity issue information, notices of the press release available in Reuters terminals, and short news items from media. For the latter, the body contains links, contact numbers, and the identity of the media source. There is also some brief news in which the body merely restates the headline, sometimes with a rephrasing. The significantly longer length in the body compared with the headline is because of some fixed formats in the body, such as the date and source. We hence included media and briefs but only included them in our headline-only analysis. We excluded the words ‘brief’ and ‘media’ before running the classification. 6) Articles of around 57 words mainly contain Refinitiv’s (Reuter’s) legal declarations on news reposting. Such articles, while they often do contain valuable information on market and stock fundamentals, trades and quotes activities, and so on, are insufficient for textual analysis. We thus excluded all the articles with fewer than 60 words, which left us with 730,590 unique news articles. We acknowledge, however, that we did exclude some potentially useful articles. For example, many articles are briefs, which begin with ‘BRIEF’ in their body content, and we included all briefs regardless of the article’s length in preliminary screening. A closer look at briefs revealed that their body contents are usually unstructured, bullet points-like

sentences capturing the key points of some events, typically financial states. A comparison of a typical brief and an ordinary, non-brief news' body content is displayed in Figure 6 1. While readers can indeed form sentiments after reading such briefs, algorithms are typically unable to capture sentiment content, as sentiments are more likely derived from performance themselves (i.e., numerical information) instead of the use of words and phrases (textual information). In contrast, an ordinary non-brief piece of news is in the form of a typically well-structured article that has proper use of grammar and semantics. We thus used briefs only for title-only analysis.

Figure 6 1 The typical body content of briefs (left) and non-brief ordinary news (right)

Having dealt with short articles, we now turn to long ones. Since there is no guidance on the appropriate thresholds on article length, we used a subjective threshold and removed all the articles beyond the 95th percentile (434 words).

3.7.2 Appendix B. Transformers and BERT

This section provides a very high-level overview of Attention mechanism and BERT model. In *attention mechanism*, each token (i.e., word or pieces of word) is given an *attention score*, and each hidden state of each word directly considers each word (including itself) in each corpus. The *attention score* determines how relevant each word is to a given word for each hidden state. *Attention* provides a solution to information bottleneck problem of previous Sequence-to-Sequence models because now we establish direct connection from the decoder to each hidden state of the encoder to focus on a particular part of the source sequence, where the exact part of source sequence to focus on is given by the attention score. It also allows parallel computing, greatly improving performance.

To define attention, we consider encoder hidden states h_i (*values*) and decoder hidden state s_t (*query*). Then attention score is:

$$e^t = [s_t^T h_1, \dots, s_t^T h_N]$$

And attention distribution for each step is:

$$\alpha = softmax(e)$$

The attention output is:

$$a_t = \sum_{1}^{N} \alpha_i^t h_i$$

We note that there are several versions to compute attention score, e_t, such as:

1. Basic dot-product attention:$e_i = s^T h_i$.
2. Multiplicative attention: $e_i = s^T W h_i$, where W is a weighting matrix.
3. Additive attention: $e_i = v^T \tanh(W_1 h_i + W_2 s)$, where W_i is weighting matrix.

Araci (2017) provides an excellent summary of Transformer framework, and we reproduce here: '*The encoder consists of multiple identical Transformer layers. Each layer has a multi-headed self-attention layer and a fully connected feed-forward network. For one self-attention layer, three mappings from embeddings (key, query, and value) are learned. Using each token's key and all tokens' query vectors, a similarity score is calculated with dot product. These scores are used to weight the value vectors to arrive at the new representation of the token. With the multi-headed self-attention, these layers are concatenated together, so that the sequence can be evaluated from varying "perspectives".*'

The way *Transformers-based models* work is that we input documents (or sentences, or parts of sentences, et cetera, depending on the task) into the encoder part of *Transformer*, which converts the input to text embeddings (or features) and tries to understand the text through multiple transformer blocks, where each block contains *multi-head attention* for it to understand connections between words. For this reason, encoder-only models are typically good at tasks where it is crucial to understand the text, such as classification. The word embedding in transformer models typically use very large dimensional vectors to represent each word. For example, in BERT, each word is represented by a 768-dimensional vector. Large dimensions would increase precision but also increases computing burden. The *decoder* then takes as input the output of encoder. Therefore, the *decoder* would have all information from the input sentence and work sequentially (or piece-by-piece) when translating into the target sentence (as applied to translation problem as illustrated in the original paper of Vaswani et al, 2017).

Many models, including BERT, uses only the encoder part of transformer. Such auto-encoder

models typically tackle tasks where it is crucial to understand the language. They are 'bi-directional' because the attention layer at each state has access to information from all inputs, and it is perhaps more appropriate to call them 'all-directional', as opposed to older models where they go sequentially from left to right or from right to left, making it a much more accurate language representation. Specifically, we consider each word's relation to each other word to understand natural language. Other examples include ALBERT (Lan et al, 2019) and RoBERTa (Liu et al, 2019). Many such models contain 'BERT' in their names because they are improvements over the original BERT model. Previous models build on a uni-directional framework because: 1) we need a direction to generate well-formed probability distributions; 2) in uni-directional framework, we build each token's representation incrementally, while in a bi-directional framework, words can 'see themselves'. BERT's solution is to mask out 15% of input words and let the model predict the masked word that it never sees; to learn relationship between sentences, BERT's second task is next-sentence-prediction. BERT is trained on Wikipedia (2,500 million words) and BookCorpus (800 million words) for it to understand natural language and is effectively formed by layers of Transformers stacked together. Now both words before and after a central word are explicitly accounted for. BERT has two versions: BERT-base, with 12 encoder layers, hidden size of 768, 12 multi-head attention heads and 110 million parameters in total; and BERT-large, with 24 encoder layers, hidden size of 1024, 16 multi-head attention heads and 340 million parameters. This again illustrates the importance of sufficiently large dataset: deep learning models (and generally, all statistical models) are only as good as their training sets; because of the huge number of parameters, previous studies using human-labeled financial texts are far from sufficient for good results, and this may (at least partly) explain why in some previous studies, transformers failed to outperform lexicon-based models.

Because even the same word in financial documents may have different meanings, it is crucial that we consider the *context*. It is therefore a great improvement over previous lexicon-based models. For example, the word 'bank' may mean the organization where we borrow and lend money, but in industry-specific texts, even if it is from a financial news vendor. Consider the sentence '*Hundreds of thousands of Hindu worshippers flocked to the banks of the Ganges in India's West Bengal state*

Friday, braving a surge in Covid-19 infections to bathe in the waters of the holy river' from the news titled '*Thousands take holy dip in India's Ganges River amid Covid surge*[7]*'*. The meaning of *bank* in this context is obviously different from the more popular meaning of *bank* in financial texts. By ignoring contexts and such double meanings of single words, lexicon-based models are only capable of giving a coarse impression of a sentence's meaning.

Decoder models are another class of Transformer-based models. They only use the decoder part of the transformer architecture. Each hidden state at each word has access to only words generated so far, and they are typically good at generative tasks, such as next sentence prediction. Influential models include GPT of Radford et al (2018), GPT-2 of Radford et al (2019), and Transformer XL of Dai et al (2019).

Lastly, we have encoder-decoder models with use both encoder and decoder part of the transformer architecture. Both encoder and decoder use self-attention, but each word's hidden state in the encoder can 'see' all other words in the corpus, while in the decoder, each hidden state of each word only has access to all preceding words. Because of this, encoder-decoder models are best suited for generative tasks where input acts as guides, such as summarization. Important encoder-decoder style sequence-to-sequence models include BART of Lewis et al (2019) and Google's T5 of Raffel et al (2019).

3.7.3 Appendix C. A Comprehensive Literature Review on Sentiment Analysis Methods

The formal study of sentiment on financial market appears to begin after the Efficient Market Hypothesis of Fama (1970), one of the most important works in traditional finance research, where behavioral economists began to realize that the EMH could not explain observed market dynamics, and that market does not timely incorporate information into asset prices. For example, Shiller (1980) finds that financial market shows excessive volatility when new information arrives; Summers (1986) then finds that most empirical studies have little statistical power in testing the EMH, and that stock prices may not reflect rational, fundamental value. De Bondt and Thaler (1985) and Cutler

[7] Kwan, R., & Agarwal, A. (2022, January 15). Thousands take holy dip in India's Ganges River amid Covid Surge. NBCNews.com. https://www.nbcnews.com/news/world/thousands-take-holy-dip-indias-ganges-river-covid-surge-rcna11888

et al (1989) were among the first studies to establish how news affect market prices. Specifically, Cutler et al (1989) finds that macroeconomic news explains only less than 1/3 of the total variance in prices, and that market reaction to major political and world events are small. Barberis et al (2005) further establishes a model of investor sentiment and show that news can cause both over-reaction and under-reaction depending on how we measure sentiment and the timeframe considered. Ke et al (2019) argues that market is inefficient, and prices do not reflect all information, hence several days after news announcement, we still observe profitable opportunities.

As early as the early 2000's, academics started to learn that information posted on the internet affects stock prices, either because the postings contain new information or because they represent successful attempts to manipulate stock prices (Tumarkin and Whitelaw, 2001). However, extracting useful information from texts is difficult and early studies used coarse methodologies that unavoidably affected their results, as we are unable to tell if an insignificant result reveals the true relationship of interest or because the methodology fails.

One of the earliest studies of financial and accounting research using textual data is Ingram and Frazier (1980). In this study, the authors try to analyze the contents of firms' environmental disclosures by counting word frequencies. Another early study on financial sentiment analysis using textual analysis is Klibanoff, Lamont, and Wizman (1998). The authors simply ask if there is country-specific news appearing on front page of *The New York Times* and show that weeks with news reports do exhibit higher market reaction, and that investor sentiment is indeed affected by news arrivals. In an interesting study, Huberman and Regev (2001) studies the 'non-event' that a Sunday article on *The New York Times* on cancer drug caused EntreMed's stock price to rise from $12 at the Friday close, to open at $85 and close near $52 on Monday. This price effect appears to be permanent as it closed at above $30 in the following three weeks. It is called non-event because the drug had already been reported in *Nature* and other popular newspapers as far as five months before that time. Therefore, the price effect was driven merely by investor sentiment instead of solid material information. This study vividly shows how 'hard' information is not timely incorporated into asset prices and how investor sentiment moves the market. The fever even spread to other biotechnology firms and investors were chasing other firms in the same sector.

With the advent of internet, researchers began to explore information from the internet. One pioneer study is Antweiler and Frank (2004). The authors investigate messages on internet bulletin boards

on 45 DJIA stocks and find that these messages help predict stock volatility. The authors use a simple Baye's classifier and assumes words are independent of each other and classify messages into buy, sell, and hold signals. This is the coarsest form of so-called 'bag-of-words' based models in natural language processing, where we rely on dictionaries (more precisely, word lists of, for example, positive and negative tones) to assess overall sentiment of a sentence or longer document. By using this method, we essentially ask: how many positive and negative words does each sentence (or document) contains, after accounting for document lengths? While this method appears to be coarse, it is very simple to use and in early days, they achieve high precision and subsequently, high model performance for different tasks, and were favored for a long time.

The most popular generic dictionary designed for finance domain is introduced by Loughran and McDonald (2011). The authors note that 'In a large sample of 10-Ks during 1994 to 2008, almost three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts.' Using 10-K files, the authors examined all words that occur in at least 5% of all documents and designed their own dictionary consisting of 2,707 unique words in six categories (negative, positive, uncertainty, litigious, strong modal, and weak modal). The authors argue that some words unexpectedly apply more to certain sectors than the others, and they raise the overall precision of generic dictionaries. For example, words like 'cancer'and 'hospital' relate more to health and medical sector, and they proxy for industry effect rather than tone. The LM dictionary subsequently became the most popular dictionary in finance research that is still widely used today. As the authors argue, their dictionary is very large, extensive, and highly relevant, as they only consider words that are used by managers in 10K filings, hence were created with business communication in mind. We note that their dictionary is highly imbalanced: only 354 words are positive, and 2,329 words are negative. Many researchers further examine and modify the LM dictionary and devise their own dictionaries to apply to their specific research area. For example, Larcker and Zakolyukina (2012) create their own dictionary to detect managers' deceptive language during earnings conference calls. They created word lists measuring hesitations (hmmm, huh, and umm), extreme negative emotions (idiot, slimy, and disgraceful), and extreme positive emotion (tremendous, smashing, and swell).

We now give a more detailed review of sentiment and opinion mining literature in computer science and how finance and economics literature adopt such methods in financial sentiment analysis. We

also briefly extend the scope beyond sentiment analysis to show that natural language processing is a large and evolving sector in finance and economics literature.

Finding ways to extract information from texts and making use of them has become increasingly important as new techniques in NLP make it possible to extract information from financial texts. While we review some important applications in this section, it is far from exhaustive. A few recent papers give excellent reviews on NLP in finance. See, for example, Loughran and McDonald (2016, 2020), Gentzkow et al (2019).

Return prediction using textual data perhaps attracts the most attention from researchers. Prior to research using NLP methods, considerably many early studies investigate the relationship between information and asset prices. Many studies find that asset prices should reflect public and private information and demand shocks through rational and irrational trading. See, for example, Daniel et al (1998), De Lon et al (1990), Glosten et al (1985), among others.

In early days, measuring sentiment in finance literature tends to rely on numerical instead of textual data. One of the most widely cited literature in finance and economics for sentiment analysis is Baker and Wurgler (2007), which reviews developments in sentiment analysis in finance and provides a sentiment index. The authors note that early studies, which date back to the 1980s, tend to investigate sentiments' effects on aggregate stock market. However, research at this stage is very pre-mature and they do not try to explicitly state or investigate the role of sentiment. With advances in behavioral finance after De Long, Shleifer, Summers, and Waldmann (1990), researchers typically assume we have two types of investors: rational arbitrageurs who are free from sentiment, and irrational traders who are prone to sentiment trading. Prices deviating from fundamental value can then come from either limits to arbitrage or irrational, sentiment-driven trading. The sentiment measures they propose and survey are the basis of early studies, including: investor survey (such as Brown and Cliff, 2005), investor mood as measured by, for example, cold seasons (Kamstra, Kramer, and Levi, 2003), investor age (such as Greenwood and Nagel, 2009), mutual fund flows, which signal what stocks are favored by mutual funds and hence may proxy investor sentiment (such as Frazzini and Lamont, 2007), trading volume, which proxies which stocks are favored by investors especially with short selling constraints (such as Scheinkman and Xiong, 2003), et cetera. Constructing their sentiment index based on Baker and Wurgler (2006), which considers six indicators, including: trading volume, the dividend premium, the closed-end fund discount, the

number and first-day returns on IPOs, and the equity share in new issues, the authors find that 'stocks of low capitalization, younger, unprofitable, high-volatility, non–dividend paying, growth companies or stocks of firms in financial distress' are more sensitive to investor sentiment. To gauge a single index, the authors performed Principal Component Analysis (PCA) on the six components. One possible explanation for their finding is that such firms are more difficult to arbitrage and are more prone to valuation errors, or at least disagreements in valuation (Miller, 1977).

Researchers have long suspected that information on the internet and the availability of internet itself change investor behavior and have attempted to quantify and investigate how such information affect financial markets. One early study is Choi et al (2002). In this study, the authors try to answer the question: how internet affects investor behavior. The authors note that back then, the internet was considered a negative shock to financial markets and was often blamed to be the cause of excessive trading, excessive herding, higher volatility in the stock market, excessive risk-taking, the Internet "bubble" of the late 1990s, and the bursting of this bubble in 2000. However, much of the blame was pure suspect and there was little evidence backing up these claims. To contribute empirical evidence to this debate, the authors specifically tackle the issue: how allowing online trading affects trading volume and investor performance. The authors first investigate investor characteristics that make one more likely to participate in internet trading instead of traditional phone trading. They find that young, wealthy, male investors are the early adopters. Controlling for trends in stock trading, the authors then find that 18 months after the introduction of internet trading, internet trading nearly doubles traditional phone trading, but trading size is considerably smaller. The authors find no statistically significant evidence suggesting that internet and phone trading have any differences in performance. The authors note, however, while in the sample they consider, namely 401(k) plan, there's no direct transaction costs to investors, the more frequent trading does incur higher transaction costs to the funds level, and this is eventually born by all investors in the 401(k) plan.

Other than Choi et al (2002), some other early studies had also realized the link between small investor behavior and stock market activity (Das and Chen, 2007). For example, Wysocki (1998) simply uses message counts and finds that variation in daily message posting volume is related to news and earnings announcements. Tumarkin and Whitelaw (2001) investigates messages posted on Raging Bull with self-reported investor sentiment measure about how positive or negative they are

about a particular stock. The authors find that on days with abnormally high message activity, changes in investor opinion correlated with abnormal industry-adjusted returns. These event days also coincided with abnormally high trading volume, which persisted for a second day. However, we found that message board activity did not predict industry-adjusted returns or abnormal trading volume, consistent with market efficiency. Tetlock (2007) and Tetlock (2008) find that negative sentiments do predict downward movements in stock prices.

In Das and Chen (2007), the authors develop a method for extracting small investor sentiment from stock message boards. The authors note that internet message boards contain a variety of information including investor sentiment, investor insights, and investors' reactions to other sources of news. The messages posted online are not necessarily information but may also contain rumors and messages intended for market manipulation. As a result, internet message boards attract the attention of investors, corporate management, and regulators. The authors explicitly acknowledge that the way they define 'sentiment' is unavoidably noisy: they define 'sentiment' as net of positive and negative opinions, but it would they include sentiment, information, and measurement errors. They used techniques that were available back then in classifying messages into bearish, bullish, and neutral sentiments, such as support vector machines and Naive Bayes classifiers, which have accuracy of only 50%, close to a pure guess. They focused on tech stocks as they are actively discussed in message boards. They find that the aggregated sentiment tracks the index returns while such effect is quite weak for individual stocks.

Early studies including Tetlock (2007) use general dictionaries from psychology literature, but they are unable to capture semantic meaning in finance (or 'domain-specific language' in computer science literature). Consider a simple example: the word 'liability' tends to carry negative sentiment in generic language, but it is merely used to describe a company's financial position when discussing a company's assets and liabilities. Therefore, while early studies shed light on textual analysis in finance, their results and model performances tend to be poor. Also, generic dictionaries tend to correlate more with negative sentiment in finance, hence early studies tend to use negative words only (Loughran and McDonald, 2011; Tetlock, 2007).

Not all research on sentiment analysis in finance finds predictive power of sentiments in relation to stock market movements, though. For example, Kim and Kim (2014) investigates Yahoo! Finance messages boards' predictive power on stock return, volatility, and trading volume. The authors did

not find online messages board's predictive power on stock market. Instead, they find that stocks' past performance predicts message board messages' sentiment contents.

While traditional research in information economics also studies return prediction, sentiment and information theory differ drastically because price impact from information is permanent while sentiment effects are transitory. As noted by Tetlock (2007), 'The sentiment theory predicts short-horizon returns will be reversed in the long run, whereas the information theory predicts they will persist indefinitely'. Empirical research in sentiment analysis tends to find reversals in return following the initial reaction. For example, Hillert et al (2014) uses 2.2 million newspaper articles from 45 US newspapers and find that stocks with higher media coverage exhibit higher momentum and the momentum reverses in the long run. The momentum continues for up to 12 months after forming portfolio based on media coverage, where high coverage portfolio outperforms low coverage portfolio by about 40 basis points per month but drops thereafter. In 2 and 3 years after portfolio formation, high coverage portfolio underperforms low coverage portfolio by, on average, 22 basis points per month.

As researchers in economics and finance started to borrow from NLP methodologies, we see a growing literature using text and media as data sources to measure sentiment. An early study in this strand is Antweiler and Frank (2004). In this paper, the authors start with media's view that online forums move the market and systematically investigate if this is true. To do so, the authors collected 1.5 million messages from Yahoo! Finance and Raging Bull, which were the most popular online forums back then. As online messages are not labeled, the authors manually labeled 1,000 messages as 'Buy', 'Hold', or 'Sell' signals. The authors then implemented two simple machine learning algorithms: Naive Bayes Classifier, and Support Vector Machine to train their data. Their sample is highly unbalanced: of the 1,000 messages, 69.3% are hold, 25.2% are buy, and only 5.5% are sell signals. A feature of intraday trading in their sample is that trading around market open and close, especially for the first and last half hour, is significantly more than other periods. They argue that this is because small traders and investors think about trading strategy after work and there are news arrivals after market close, hence small investors place considerable number of orders for market open. Mutual funds and day traders close their positions near market close, driving up market close trading activities. This feature is still present today. The authors then used OLS, realized volatility models, and GARCH to study the relationship between online forum posting, sentiment, and stock

market behavior. The authors claim that they are the first to report a negative relationship between forum posting and next-day return, although this is economically negligible especially after accounting for transaction costs. Consistent with intuition, the authors find that disagreements in message sentiment causes more trading and higher volatility.

Research in this field typically uses short window of a few days after the relevant news release. Shiller (2000) argues that investors follow the printed word even though much of it is pure hype, suggesting that market sentiment is driven by news' content. Following Shiller (2000)'s argument and the formal evidence from Tetlock (2007) that negative words in Wall Street Journal predicts daily stock returns, Garcia (2013) revisits the issue by studying financial news from New York Times from 1905 to 2005. He shows that in hard times as proxied by recession, investors are more sensitive to news. Specifically, during recessions (expansions), a one standard deviation change in sentiment measure predicts 12 (3.5) basis points change in daily average of DJIA, where sentiment measure is the classic bag-of-word approach based on the number of positive and negative words in financial columns of New York Times. He also finds that both positive and negative words help predict returns while previously, Tetlock (2007) shows that only negative words have predictive power.

Ke et al (2019) is a recent study in textual sentiment analysis. In this paper, the authors use data from Dow Jones Newswire, which contains all historical news for the US companies from 1986 to April 2020. The authors start from the view that news simultaneously affect investor sentiment and market return and propose a three-step framework: 1) as positive sentiment drives up return and negative sentiment drives down return, we can use market reaction as a guide to automatically create dictionary of positive and negative sentiment; 2) use a two-topic model to estimate positive and negative sentiment scores; 3) predict sentiment scores of news articles. With the predicted sentiment score, the authors then go long the 50 most positive stocks and go short the 50 most negative sentiment stocks each day to forma a zero-cost portfolio. The authors achieve an annualised Sharpe ratio of 4.3 overall.

Hoberg and Phillips (2021) specifically deals with the issue of industry momentum using textual data. Drawing from previous literature, the authors suggest that industry momentum likely stems from investor inattention.

A handful of research use readily available sentiment data from commercial vendors. Groß-

Klußmann and Hautsch (2011) empirically examines high-frequency market reactions to an intraday stock-specific news flow. The authors wish to analyze to which extent high-frequency movements in returns, volatility and liquidity can be explained by the underlying mostly nonscheduled news arrivals during a day. To do so, they rely on Reuters NewsScope Sentiment Engine, which is a black-box engine that automatically analyses news when they arrive and produces a sentiment label (positive, neutral, and negative), novelty, and relevance indicator. The authors use 29,497 news headlines for 40 stocks from January 2007 to June 2008. The authors find that high frequency trading does react significantly to relevant intraday company-specific news arrivals, as expected; among other measures, volatility and cumulative trading volume are the most significant responders.

In another study, Uhl (2014) uses a much larger dataset of sentiment data from 3.6 million Reuters news articles from January 2003 to December 2010. The dataset again uses the black-box sentiment data from Reuters. The author acknowledges that there are two issues central to their study: 1) prior studies had no consensus on sentiment measure, hence one needs to carefully choose the sentiment measure; 2) timeframe also greatly affect stock prices post news arrivals. To tackle the first issue, the author chooses to use Reuters sentiment which gives sentiment score of positive, neutral, or negative for each news piece. To tackle the second issue, the author first note that prior studies either look at investor sentiment, such as Tetlock (2007, 2011) and Tetlock et al (2008), or investor sentiment, such as Brown and Cliff (2005). Studies on news sentiment typically consider sentiment effects at short intervals up to a few days while studies on investor sentiment typically consider longer timeframe of monthly sentiment effects. Other asset classes may see 'sentiment effects' lasting a few years. For example, Menkhoff and Rebitzky (2008) finds that sentiment effects in the foreign exchange market may last up to two years. The author then chooses to form monthly sentiment index using all the news sentiment available. The author then uses VAR model to assess the dynamics of sentiment and finds that positive and negative Reuters sentiments do affect stock returns, although negative sentiment's effects are larger; fundamental factors, such as the Conference Board Leading Economic Indicator, do not have a measurable effect on stock returns and the author proposes that this is because market participants can quickly incorporate fundamental information into asset prices hence they are not significant in analysis spanning months.

Sentiments in financial market are also interesting to policy makers and market participants rely on a range of hard (such as unemployment rate, price index, et cetera) and soft variables (such as survey-

based methods) in forecasting future economic conditions. For the latter, consumer sentiment by the University of Michigan and the Conference Board appears to be the most popular used by practitioners and policy makers (Shapiro, Sudhof, and Wilson, 2020). A recent sentiment analysis study in finance and economics literature using 'soft' variables is Shapiro et al (2020). In this paper, the authors experimented with different sentiment analysis tools and propose their own sentiment score measure to extract sentiment information from news. The authors purchased 238,685 economic and financial news from LexisNexis from 16 major newspapers (Atlanta Journal-Constitution, Boston Globe, Chicago Tribune, Detroit Free Press, Houston Chronicle, Los Angeles Times, Memphis Commercial Appeal, Miami Herald, Minneapolis Star Tribune, New Orleans Times-Picayune, New York Times, Philadelphia Inquirer, San Francisco Chronicle, Seattle Times, St. Louis Post-Dispatch, and The Washington Post) from January 1980 to April 2015. The authors purchased only news with sufficient contents (news not labeled as 'brief', 'summary', or 'digest') and are long enough (longer than 200 words) to exclude articles that appear elsewhere and reduce noise, because 'very short articles are likely to be more noisy'. As an additional step, they only include articles that include 'said, says, told, stated, wrote, reported', because they consider such articles reporting opinion of someone or some group of people, hence carrying strong sentiment signal. As news articles are by construction unlabeled, the authors asked 15 research assistants at the Federal Reserve Bank of San Francisco to hand-label 800 news articles into: Very Negative (1), Negative (2), Neutral (3), Positive (4), and Very Positive (5), and the 800 labeled data forms the basis of their training sample. We note that their labeled sample is very small compared with their sample size, accounting for only 0.335% of their total sample. The authors very explicitly distinguish sentiment and information contents. They state that:

"By sentiment, we mean the tone/feeling/emotion expressed of the article rather than the economic substance of the article. For example, If the writer is talking about a report of very high GDP growth but is expressing concern that this reflects overheating of the economy and monetary policy being behind the curve, then this could be the writer expressing negative sentiment even though he/she was talking about high growth."

The authors compare the performances of various lexicons with advanced ML models. To our surprise, the authors find that BERT models perform almost as good as lexicon-based models which combines LM and HL lexicon in predicting news sentiment, and that LM+HL lexicon performs

almost as good as VADAR. However, their findings are not too surprising. As is with other deep learning models (Marcus 2018), BERT requires a large training sample due to the large (usually several millions) number of parameters to learn. The authors' very small training set is hence unable to allow BERT to work to its full capacity. As is with other lexicon-based sentiment analysis, the overall sentiment for a text is simply the difference between its proportions of positive and negative words. Because of the theoretical advantage of transformers and the fact that BERT's superiority in sentiment classification has been confirmed in many other studies, we believe the lower performance of transformers is largely because the BERT model the authors are using are not designed specifically for finance domain, and their very small training sample is especially problematic for deep learning models including BERT.

There are many other applications of NLP in finance but they are not the focus of this study: Manela and Moreira (2017) constructs a news-implied volatility measure from Wall Street Journal front page and their findings are consistent with recent theoretical advances suggesting disaster risk is an important source of volatility; Jeon et al (2021) studies stock price jumps and they find that news frequency, tone, and uncertainty and they find that news flows can explain jump intensity and jump-size distributions and explain an important fraction of variations in the jumps across individual companies; Huang et al (2023) finds that institution trading on stocks tend to concentrate on the first release of a series of news and such trading predicts returns over weeks and suggests that institutional investors facilitates price efficiency by quickly interpreting public information and incorporating public information into asset prices; Engle et al (2020) studies climate news and shows how to use a synthetic portfolio to hedge climate risk.

Apart from empirical research, some studies also attempt to give theoretical grounds to news trading. For example, Foucault et al (2016) is a theoretical paper on news trading. The authors construct a model where the speculator's private signals can be used to forecast both short-run price reactions to news arrival and long-term price changes.

A handful of research also investigates volatility and information. In an early study, French and Roll (1986) examines information arrival during market open and market close. Based on the notion that volatility may come from private information, which is revealed through trading during trading hours, or public information, which may arrive either during trading hour or non-trading hour, or

irrational trading, the authors find that return volatility is mainly from rational trading driven by private information. Their conclusion appears to be supported by later research. See, for example, Ito et al, 1998; Chordia et al (2011). However, their findings are subject to different interpretations. For example, Hong and Stein (2003) notes that 'Roll (1984, 1988) and French and Roll (1986) demonstrate in various ways that it is hard to explain asset price movements with tangible public information'.

In a recent study, Boudoukh et al (2019) revisits this issue using textual data. The authors wish to investigate the saliency of news instead extract sentiment contents from news and they wish to match companies with events such as such as new product launches, lawsuits, analyst coverage, news on financial results, and mergers. Because of computation limitations, the authors only consider S&P 500 companies that have at least 20 trading days in the sampling period. To do so, they use two methods: 1) visual information extraction platform (VIP), which uses a mixture of a rule-based information extraction platform and a trained support vector machine classifier, to identify event instances for companies and measure sentiment from text contained in financial news. The authors apply VIP to Dow Jones Newswire from January 1, 2000, to December 31, 2015. 2) A commercial product, RavenPack. When news articles are released, RavenPack would automatically process the news, with the algorithm being a black box, and produce 16 fields including timestamp, company identifier, relevance score, et cetera. RavenPack recommends a relevance score of at least 90 and such news are typically highly relevant of company events. The authors are then able to assign variance in stock prices that are due to arrival of firm-specific events, or 'fundamental information in news'. The authors find that 49.6% (12.4%) overnight (trading hour) idiosyncratic volatility is due to fundamental information related to company events. The authors also document a large negative correlation over time (i.e., −0.50) between average idiosyncratic volatility during overnight hours and the contribution of identified news to overnight return volatility, arguing that the benefit to private information production has decreased due to the increase in publicly available information.

Market participants and policy makers rely on very different indexes of investor and consumer sentiment. While labor-intensive and very time-consuming and expensive to construct, survey-based methods, notably, Michigan Consumer Sentiment index and the Conference Board's Consumer Confidence index, are widely used in economics. The most popular method in finance and economic

literature to extract sentiment from texts is lexicon-based method. This method relies on pre-defined 'dictionaries', which are 'sentiment-charged' lists of words, and by giving each word a score (such as 1 for positive, 0 for neutral, and -1 for negative), it effectively asks: in each corpus, how many words belong to each dictionary? In early studies, generic dictionaries from psychology and sociology literature are used, such as Bollen et al (2011). However, researchers quickly realized that context, or domain-specific dictionaries are necessary for accurate sentiment classification, and tailoring dictionaries to specific needs of each research greatly improves performance, because words have very different meanings in different contexts. For example, the word 'liability' does not always carry negative sentiment in financial news, but in general texts, they are very negative words (Loughran and McDonald, 2011). Other words that appear frequently in finance documents but do not necessarily carry negative meaning include tax, cost, capital, board, and depreciation. Dictionaries may contain only single words ('unigrams') or phrases containing several words (n-grams), although unigrams are much more prevalent than n-grams. This method belongs to 'bag-of-word' representation of texts, which ignores the context, inter-relatedness of word, sentence structure, et cetera, and considers each word in isolation. One popular generic dictionary designed for finance research is Loughran and McDonald (2011). While this dictionary is 10 years old, it is still widely used today. See, for example, Shapiro and Wilson (2021), which uses meeting transcripts of the Federal Open Market Committee to study the committee's loss function. Common dictionaries include:

1. Harvard General Inquirer (GI) Dictionary. This dictionary is mostly seen in early studies where dictionaries catered specifically for finance and economic research was not available, although some recent studies do use this dictionary (such as Heston and Sinha, 2017). This is also used as the basis of Loughran and McDonald (LM) dictionary. It consists of 3,626 words labeled 'positive' and 'negative'.

2. Loughran and McDonald (LM) (2011) dictionary, which was subsequently updated in 2014 and 2020. It consists of 2,707 words in 2014 version. The words are labeled 'positive' and 'negative' and are sourced from 10-K files of publicly traded companies.

3. Hu and Liu (HL) (2014) dictionary. This dictionary is very popular (with over 8,500 citations) in general sentiment analysis literature but has limited use in finance and economics literature, because while it is very large (with 6,786 words), it is sourced from online movie reviews

where reviewer themselves label their review as ‘positive’ and ‘negative’.

The dictionaries can be very different from each other. Shapiro et al (2020) directly compares GI, LM, and HL dictionaries. They find that 58% of LM dictionary words are not found in the other dictionaries, and only 31% of GI words are not found in the other two dictionaries. For the common words that appear in two dictionaries, the different dictionaries tend to agree on their sentiment contents: the disagreement rate of HL-LM dictionaries is only 0.9%, with HL-GI and LM-GI being 1.4% and 2.7% respectively. While LM dictionary is tailored to finance and economic literature, its small size means very often, news may use words that are not found in it (LM, GI, and HL dictionaries only classify 2.8, 6.4, and 4.4% words in their sample corpus). The authors hence suggest combining the different lexicons.

We are interested in studying sentiment in financial markets because they ultimately move the market. For example, Baker and Wurgler (2007) sentiment index is successful in explaining cross-section of stock returns and is one of the most popular sentiment indexes by far. A few studies investigate stock behavior in high and low sentiment periods where the classification is given by Baker and Wurgler (2007) index. For example, Stambaugh, Yu, and Yuan (2012) investigates how investor sentiment explains stock market anomalies. The authors find that high sentiment follows high anomalies, and that the short leg is the profitable one. Yu and Yuan (2011) show that the positive relationship between market return and volatility is gone with high sentiment period. More recently, Jiang, Wu, and Zhou (2018) show that many anomalies are sensitive to investor sentiment. Pástor, Stambaugh, and Taylor (2017) show that funds trade more when sentiment is high, and that there is a positive relation between fund turnover and return.

It is also interesting to consider the types of texts inputs used in financial studies. The exact types of financial documents to use in financial sentiment analysis differ in different studies and novel dataset in recent years greatly facilitates sentiment extraction. Broadly, sentiment may be extracted from corporate reports (see, for example, Li 2006), social media such as twitter (such as Bollen, Mao, and Zeng, 2011), partial texts in financial media (such as Tetlock 2007), and in more recent years, full texts of historical financial news (such as Ke, Kelly, and Xiu, 2019). Because of data availability, social media such as Twitter and StockTwits, which could be considered Twitter for investors, were the main source of data in early studies, as such data tends to be freely available and covers sufficient large number of stocks. Bollen et al (2011) is the most widely cited study to use

Twitter to predict stock market movements. The authors note that traditional asset pricing theories and studies under the efficient market hypothesis (Fama, 1996; Fama, Fisher, Jensen, and Roll, 1969) imply that stock market prices should follow random walks and should be unpredictable because new information is unpredictable, but multiple strands of literature, including socioeconomic theory of finance and behavioral finance, have been challenging this view of stock market. See, for example, Prechter and Parker (2007), Smith (2003), and Nofsinger (2005). Based on research from psychology (such as Dolan, 2002) and behavioral finance (such as prospect theory of Kahneman and Tversky, 1979) that emotions do affect behavior and decision-making, they authors try to extract public sentiment from a collection of tweets from February 2008 to December 2008. To do so, they used two sets of tools: OpinionFinder, which classifies daily moods into positive and negative sentiment, and GPOMS, which classifies daily moods into six categories: Calm, Alert, Sure, Vital, Kind, and Happy. Both classifiers are based on lexicons and potentially asks how many words in the tweets belong to each category's dictionary. The dictionaries are pre-defined from psychology research. The authors find that sentiments do predict Dow Jones Industrial Average in coming days.

Recent studies also attempt to use advances in machine learning beyond textual information. For example, Obaid and Pukthuanthong (2021) constructs sentiment index from news photos and they find that their Photo Pessimism index predicts market return reversals and volume.

Apart from finance and economics literature, a large and growing number of machine learning literature attempt to measure and classify financial sentiment analysis. Machine learning methods are especially suitable for this task because financial sentiment analysis at is core is a classification problem. In early studies of this strand of literature, researchers typically use simple supervised classifiers to assess sentiment contents of texts and are typically reliant more on bag-of-words approach with significant effort in feature engineering. See, for example, Turney and Pantel (2010). Ghiassi, Skinner, and Zimbra (2013) introduces a Twitter-specific lexicon and uses Dynamic Architecture for Artificial Neural Networks (DAN2) which outperforms a simple Support Vector Machine (SVM). Wang et al (2015) experiments with common classifiers in classifying financial sentiment using StockTwits sample into 'bearish' and 'bullish' and finds that SVM outperforms Naive Bayes and Decision Trees with an accuracy of 76.2%. Later, deep learning methods became popular because of their superior classification performance.

While early studies in sentiment and opinion mining in computer science literature also relies on

simple lexicon-based models, they were quickly replaced by newer models, even if one tries to improve model performance by introducing rule-based models in addition to simple word counts. For example, while 'good' carries positive sentiment, 'not good' negates the positive meaning of the word 'good'.

However, natural language is too complicated to fit in a rule-based model, unless the rule is endless. More recently, researchers using lexicon-based methods attempt to account for contexts in addition to words, such as VADER of Hutto and Gilbert (2014). VADER is a simple rule-based sentence-level classifier comprising of 1) a large dictionary, which assigns each word a score from -4 to 4 for most negative to most positive; and 2) a set of heuristic rules to determine each word's context in the sentence. The heuristic rules are very simple and accounts for common sentence structures, such as 'but' reverses sentence's meaning, words like 'very', 'a bit' modifies intenseness of a sentence's sentiment, et cetera.

A breakthrough *word representation* method is Word2Vec of Mikolov et al (2013a,b). This model has two versions: Continuous Bag-of-Words, which predicts central word based on surrounding contextual words, and Skip-Gram, where central words are used to predict surrounding, contextual words. The authors experimented the performance of Word2Vec on a range of NLP tasks including sentiment analysis. However, it has two main drawbacks: 1) it is incapable of handling words that are not seen in training sample; 2) words semantically similar would still be given two completely different encodings. For studies using Word2Vec on sentiment analysis, see, for example, Zhang et al (2015).

One important development after *bag-of-words* is Global Vectors for Word Representation (GloVe) of Pennington, Socher, and Manning (2014) from Stanford University, which is one way of word embeddings. Generally, word embedding tries to represent each word in a high-dimensional space, and in the process, words that semantically similar are close to each other. Unsurprisingly, generally the higher dimension, the better semantic meaning we can capture, but the more expensive training we face. Word embeddings are also the basis of encoder-decoder frameworks including BERT. This strand of method is also called distributional semantic model, (DSM), vector space model of meaning, and semantic space model of meaning. In GloVe, each word is still represented by a dense vector that incorporates its semantic characteristic, hence words of similar meaning are close to each other in the high-dimensional vector space, such as 'very' and 'extremely'. Lexicon-based models

typically account for document length and words that appear only in certain documents through 'Term Frequency - Inverse Document Frequency' (TF-IDF). For research using GloVe on sentiment analysis, see, for example, Rezaeinia, Rahmani, Ghodsi, and Veisi (2019).

One breakthrough in *word embeddings* is a pre-trained model called Embeddings from Language Model (ELMo) of Peter et al (2018). Previously, word embeddings would produce the same static regardless of different texts. In ELMo, the same word would be given different embeddings depending on the context. This greatly improves its performance because now we can capture the same word's different meanings.

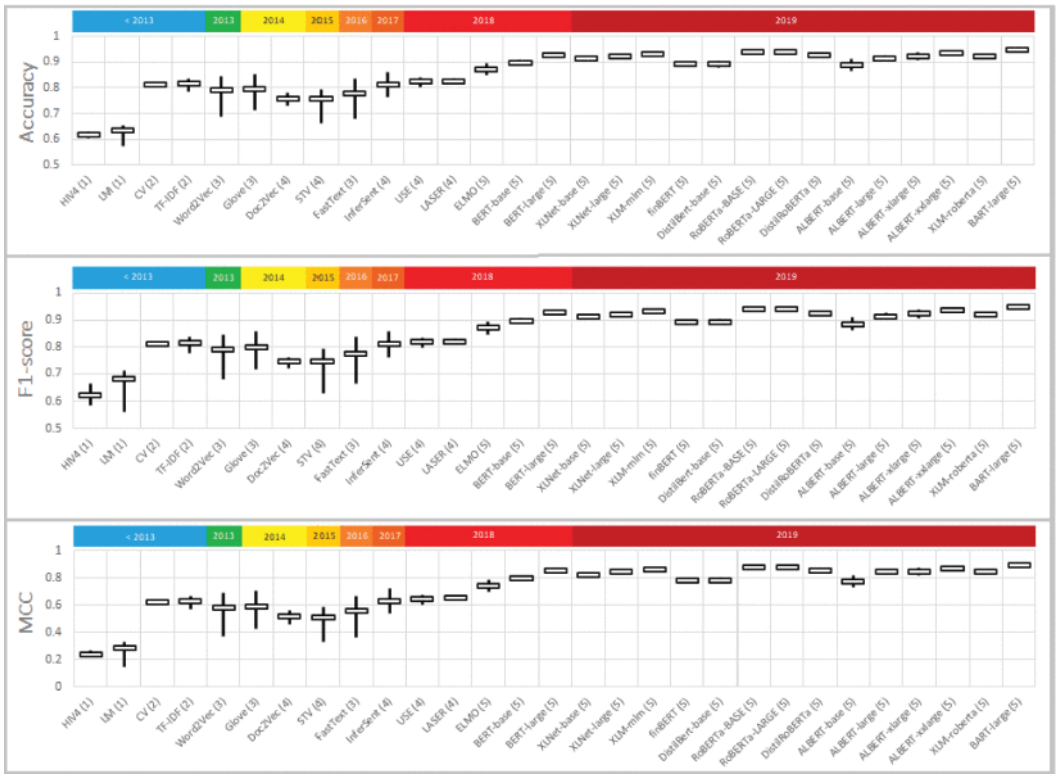
Recent developments in machine learning, especially deep learning methods, attempt to let the algorithm learn the context, and they typically achieve much better performance than bag-of-words and rule-based models. This is not surprising: each word has a meaning, but the meaning is deeply affected by all words in front of and after the word. By ignoring the context, it is not surprising that traditional language models perform poorly in understanding a corpus, and hence poor performance of models relying on the language model. In plain vanilla deep learning models, each node, or hidden state, takes input as the output of its previous state and performs a non-linear transformation. The exact non-linear function may vary, such as ReLU, sigmoid, and Adam.

Textual analysis in finance is not limited to English. In recent years, as China attracts attention from almost all fields in economics and finance, a growing number of literature studies textual analysis in the Chinese financial market. A recent work in progress is Fan, Xue, and Zhou (2021). Deviating from previous studies, which typically use topic models or dictionaries to reduce dimensionality and ignores semantic sequences and structures, the authors propose a method to account for the whole corpus. The authors propose a 'Factor-Augmented Regularized Model for Prediction (FarmPredict) on stock returns by extracting the hidden topics (factors) from all words with consideration of structure and interactions of phrases or words'. The authors use a three-step framework: 1) use PCA to convert articles into vectors of hidden features consisting of multiple factors and idiosyncratic components; 2) screen the idiosyncratic variables by their correlations with beta-adjusted returns; 3) use a simple LASSO model to predict return using hidden factors and the screened idiosyncratic components. The authors use data from Sina Finance and predict returns from 2015 to 2019. The authors then long the 50 most positive sentiment (return) companies and short the 50 most negative companies with daily rebalance. They find that for the Chinese market, positive leg's returns are

much more important than negative leg, likely due to China's short-sale constraints; and that 7 days before news publish, returns already started to react, likely due to information leakage. We believe part of the pre-news reaction is brought about by the feature of their dataset: Sino Finance does not always provide timely information on the latest news, many news articles are in-depth analysis of events already occurred some time ago, hence market reaction was already in place. It is, however, interesting to note that market reaction can last up to 7 days before Sino News, a major public, online, and freely available news provider in China, publishes news articles. Still, return prediction is overwhelming: FarmPredict achieves a Sharpe ratio of a stunning 9.37, higher than all previous models.

More recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, and Toutanova, 2018), based on transformer architecture of self-attention mechanism of Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, and Polosukhin (2017), outperforms all previous models in almost all NLP tasks including sentiment analysis. Transformer is an encoder-decoder architecture. Models based on transformer may use encoder only, decoder only, or both encoder and decoder.

To our knowledge, the latest and most comprehensive comparison of transformer-based models against lexicon-based models in financial sentiment analysis is Mishev, Gjorgjevikj, Vodenska, Chitkushev, and Trajanov (2020). In this paper, the author acknowledges that the problem of unavailability of large, labeled dataset and the lack of domain-specific model for financial sentiment analysis, and uses publicly available labeled dataset to assess the performance of lexicon and transformers-based models. Their results are opposite to Sharpiro et al (2020) and show that transformer-based models significantly outperform lexicon models. We reproduce their Figure 9-9 here:



Clearly, Transformer-based models significantly outperform previous models including all lexicon-based models across accuracy, F1 score, and MCC: the authors experimented with different versions of lexicons and introduced machine learning techniques in addition to plain vanilla lexicon model, such as TF-IDF, and cross validation. Lexicon-based models are around 60%-80% accurate with F1 score also around 60%-80%, while different Transformer-based models achieve accuracy and F1 scores both over 90%. The best-performing model is BART-large, which achieves accuracy and F1 of 94.7%. The model of choice in this paper is ALBERT-xlarge, the second-best performing model with an accuracy of 93.6% and F1 score of 93.5%, While BART performs slightly better, it is much more resource-consuming with 406 million parameters. ALBERT-xlarge is designed to be light on resources with only 58 million parameters in its xlarge-V2 version. As our dataset contains all historical news of US companies from 1996, model performance would be an issue and the next-best model which is cheap to train is critical for our study.

Before BERT, the best-performing sentiment classification model relies heavily on convolutional neural networks (CNN) and LSTM, which introduces a forget gate at each hidden state to teach the model how much previous information to keep and to drop. See, for example, Wang, Huang, Zhu, and Zhao (2016). GloVe provides pretrained word embeddings for each word and largely ignores

context. As context does affect each word's meaning, especially for domain-specific tasks, BERT-based models, which directly incorporates context, is expected to outperform almost all previous models. However, as with other deep learning models, we need a large training set.

Traditionally, language models are trained to predict the next word in each text and more recently, researchers in NLP apply language models to general downstream tasks including sentiment classification. Normally, such models are pre-trained on a large sample and then fine-tuned using domain-specific texts (Kant et al, 2018). BERT is one such model that could be fine-tuned using finance texts. As is with other NLP tasks (see, for example, McCann et al 2017), pre-training models specifically on financial text greatly improves performance. A version of BERT specifically trained for finance domain is FinBERT of Araci (2019). The author uses an open-source dataset from Reuters TRC2 comprising of 1,800,370 news stories covering the period from 2008-01-01 to 2009-02-28 or 2,871,075,221 bytes. The author only uses a subset of TRC which consists of '46,143 documents with more than 29M words and nearly 400K sentences' after filtering for 'some financial keywords' to limit sample size for training and make their sample more relevant. This dataset is unlabeled. They also use a labeled dataset of Financial PhraseBank from Malo et al (2014), where the authors ask 16 annotators with adequate business education to hand-label around 4,845 phrases and sentences sampled from financial news texts sourced from LexisNexis into 'positive', 'neutral', and 'negative'. The annotators were asked to label based on how they think the sentence would affect the company's stock price. The authors use Financial PhraseBank to run their main set of analysis and set 60% as training set, 20% as validation, and 20% as test set. As is with routine procedures, re-training model on domain-specific texts greatly improves the model's performance (see, for example, Howard and Ruder, 2018). The author then trains his model by adding a dense layer on top of the usual BERT model, as is the suggested method for arbitrary downstream NLP tasks by the original author of BERT (Devlin et al 2018). The author achieves an accuracy of 86%, cross entropy loss of 37%, and F1 score of 84%. While their performance is not too eye-catching, this is probably because there are indeed different ways to interpret the same piece of news. For the subset of Financial PhraseBank with agreement rate of 100% (about 70% of total sample), BERT retrained on financial texts achieves an accuracy of 96%. This represents another advantage of using market reaction to label news sentiment: market reaction represents the average opinion of all market participants; while we cannot observe how many investors think a news is bullish and how many think it is bearish, we can infer the average opinion from market prices, and this is what is

important for our purpose: formulating a long-short portfolio consisting of stocks that are most likely and least likely to rise, instead of precisely working out the public's opinion. In terms of fine-tuning, their best results are achieved with slanted triangular learning rate, gradual unfreezing, and discriminative fine-tuning.

Chapter 4 Boosted Returns with News: News, Volatility, and Portfolio Implications

4.1 Introduction

While in everyday life and in many financial economics applications, we tend to use the word '*volatility', 'risk',* and '*uncertainty'* interchangeably, they essentially mean drastically different things. '*Uncertain'* implies that we do not know the outcome of an event, and the outcome may be good or bad, while '*risky'* almost unambiguously mean the bad side of a possible outcome[8]. Semivariance, realized semivariance, and after that, realized signed jump variation (SJV), are built upon this very simple idea. Barndorff-Nielsen et al (2008) proposes realized semivariance by decomposing realized variance to two parts relating to positive and negative price movements using high-frequency stock prices, hence '*good'* and '*bad'* volatility, and derived its statistical properties. Notably, the difference between positive and negative semivariance removes the continuous part of realized volatility and leaves us the jump component. Patton and Sheppard (2015) uses the difference between the positive and negative parts of realized variance, SJV, to show that negative part of realized variance has strong predictive power of future volatility. Bollerslev et al (2020) uses relative version of the SJV, which they term Relative Signed Jump Variation (RSJ), to show that there's a strong (negative) predictive power of RSJ of future stock return. Bollerslev et al (2020) proposes the question of the sources of the predictive power of RSJ.

In this paper, we use Refinitiv Machine Readable News (MRN) database to shed some light on the predictability of RSJ and its predictive power of stock return. In traditional asset pricing models, return and volatility are considered continuous processes. When asset pricing models do consider discrete processes, they are often simplifications of the continuous model and in the limit, they would approach their continuous counterparts. In recent years, literature starts to consider price and volatility jumps. Notably, such jumps tend to be event-driven. Most of these studies tend to focus on

[8] Researchers in utility theory may define 'risk' as a situation where we know the possible outcomes of an event, and 'uncertainty' as a situation where we do not know the possible outcomes of an event. See, for example, Park and Shapira (2017).

macroeconomic news (Rangel, 2011; Lahaye et al. 2011; and Huang, 2018). Admittedly, macroeconomic news is much more manageable and when one tries to focus on market returns, macroeconomic news serves their purpose right. The large number of company-specific events provide a rich source of data for practitioners who wish to investigate profitability from company events, for academics and practitioners who wish to learn new insights from company-specific data, among others. Company news, which captures the latest company-specific events, is a wonderful proxy for triggers of volatility jumps.

The causality channel we propose in this paper is very intuitive and simple. Firm-specific events trigger volatility jumps, which can be used to form portfolios. We approximate firm-specific events by firm news from Refinitiv MRN database. We construct abstractions of news, including news sentiment scores, audiences, number of audiences, topics, number of topics, firm and time fixed effects, et cetera, and use them as input for our regression. We choose XGBoost, a popular and powerful machine learning algorithm, to predict next-day RSJ using today's news.

Specifically, Refinitiv MRN database contains all historical news from Refinitiv (formerly known as Thomson Reuters). We use 44 stocks with the largest amount of news from 2014 to 2019 (inclusive) and ask: controlling for news sentiment and today's stock volatility measures (including realized variance, positive and negative semivariance, and today's RSJ), can we predict tomorrow's RSJ? To assess this predictive power, when forming portfolios based on the predicted RSJ, is the portfolio profitable? Our results show that next-day RSJ, which is bounded between -1 and 1, is reasonably predictable with predicted root mean squared error (RMSE) of 0.325. Forming portfolios by going long (short) the top (bottom) 10 percentile stocks gives an annualized Sharpe ratio of 2.06, significantly higher than portfolio return bases on today's RSJ (0.619) and best-performing sentiment portfolio (1.94). We note that we do not technically have a 'portfolio': because of sparse news arrivals and small number of stocks to work with, we have, on average, only 5.6 stocks in the 'portfolio' each trading day, effectively suggesting a risky position. In a sense, makes the portfolio return more attractive. With more stocks and news to work with, we expect to have better portfolio performance with effective diversification. Our results suggest that compared with today's price and volatility dynamics, what's happening in the market, and hence what's discusses in the news, are a vital source of volatility jumps. Using this predicted volatility jump to form portfolio achieves significantly higher portfolio return than using today's actual RSJ. We also note that Bollerslev et al

(2020) found that when selecting stocks using RSJ, small companies are more profitable, while the stocks we use in this paper are large ones, again suggesting that we could potentially improve portfolio performance further when using smaller stocks.

There have been some attempts to investigate how textual news can be used to predict market-wide volatility. For example, Manela and Moreira (2017) uses historical *Wall Street Journal* front pages' titles and abstracts dating back to 1890 to construct a *news implied volatility (NVIX).* Deviating from ARCH family models, the authors use monthly Support Vector Regression (SVR) based on occurrences of *n-grams* (n words that tend to occur together) to directly predict volatility. This measure of volatility peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises. Periods of high volatility are followed by high returns and are high before economic disasters, consistent with intuition and existing literature.

While studies on the source of market-wide volatility is important, when forming portfolios, if one is not on a passive investment strategy, stock-specific volatility dynamics is a more important and relevant measure. In this paper, we use RSJ to investigate portfolio construction decisions based on what's going on in the market as implied by news topic codes. We show in methodology section that SJV (and its relative form, RSJ), is an estimator of volatility jump.

One possible explanation to the observed portfolio return is that investors overreact to news of certain topics, and this induces reversals in volatility jump the next trading day. Such topics should be hard to pick up by human traders but easier by algorithms. Information on, for example, less than expected annual profit would be quickly incorporated into asset prices because such numerical data is easily fed into algorithms and investors easily understand their implications, inducing less reversals. The very large number of topics makes it almost impossible to hand-pick the topics that we think would be useful, but algorithms are good at finding the most relevant and best performing variables and hence in a sense, automates topic code selection for us. In this paper, we rely on XGBoost to model the predictability of RSJ. We note that many previous studies choose to use headlines only in their research, many of which use *Wall Street Journal* as their data source. The choice of headline-only as data source appears to be from:1) unavailability of full text; 2) unavailability or computationally expensive techniques to extract full text information. In this paper, we overcome the first issue by using all news from Refinitiv MRN database, which contains two types of news: 1) news alerts, which are headline-only real-time updates to some events that just

happened in the market; 2) news articles, which are a more detailed and comprehensive discussion of news.

A convenient feature of Refinitiv MRN database is that it contains detailed topics of the topics discussed in the news. Of the total topic codes of over 2,400, around 1,400 appear in our sample. This greatly helps us investigate the informational contents of topics: previous studies' textual data usually (if not always) contain no topic information, and researchers are forced to use techniques such as Latent Dirichlet Allocation (LDA) to group texts into different topics. See, for example, Grape (2011). This method classifies documents into mutually exclusive and exhaustive topics. This is, however, far from reality. One document can well accommodate multiple topics. In our sample, on average, each piece of news relates to around 30 news topics covering the geographic location, business sector, event type, et cetera. This allows a much more detailed and comprehensive analysis of topic information, which presumably improves the portfolio performance built after the exploration of topic information.

As RSJ entails high-frequency data where we would easily exceed database download limits, we include a stock into our sample only if it has more than, on average, 30 stock-days with news from 2014 to 2019. As the dataset would be highly sparse and selecting topics to include/exclude from analysis would be practically impossible, we use XGBoost to automatically select topic codes and predict next-day RSJ by regressing next-day RSJ on today's stock sentiment (where sentiment score is obtained by applying FinBERT model to news texts), realized volatility, positive and negative parts of the realized volatility, and the 1,400 topic codes. We note that XGBoost only selects around 150 topic codes out of the 1,400 available. An additional benefit of XGBoost is that it works exceptionally well on sparse dataset and interested readers may refer to Chen and Guestrin (2016).

To explain our portfolio returns, we attempt to use proxies for investor attention and investor recognition. Specifically, we include the mean number of audiences and volume shocks of stocks used in daily rebalanced portfolio. We find that they indeed significantly explain the daily return of portfolio. Specifically, a greater number of audiences would reduce portfolio return while higher volume shocks to the portfolio would increase portfolio return.

We now give a short lookback of related literature and how they relate to this paper. The study of semivariance and RSJ follows from volatility literature. Of the observed characteristics in asset risk-return studies, the leverage effect is a long-observed and long-studied characteristic of asset returns.

First noted by Black (1976), leverage effect refers to the negative correlation between asset return and changes in volatility. Namely, asset prices tend to rise when volatility drops, and they tend to gain values when their volatilities are lower, vice versa. A large literature has attempted to explain leverage effect and explore its implications. See, for example, Bollerslev, Engle, and Nelson (1994) and Andersen et al (2006). In Barndorff-Nielsen et al (2008), the authors propose a new estimator, realized semivariance, that's very simple and clever. The authors propose this estimator as an extension to downside risk literature and is a natural extension to semivariance, and is subsequently adopted in other strands of literature, such as Patton and Sheppard (2005), which uses realized semivariance to investigate predictability of equity price volatility, and Bollerslev, Li, and Zhao (2020), which investigates the cross-section of stock returns. These studies use the same framework of Barndorff-Nielsen et al (2008) and we review its framework in methodology section.

The relationship between news, sentiment, and volatility has long been investigated in literature. We note that, in early studies, the word '*news'* is used in a broad and usually numerical manner. For example, an early and influential paper, '*No news is good news: An asymmetric model of changing volatility in stock returns'*, by Campbell and Hentschel (1993), investigates the role of past 'news' on stock volatility. The word 'news' in their paper refers to past stock prices (and hence past returns and volatilities), instead of textual news in more recent contexts. This paper is one of the first to systematically investigate volatility feedback effect and allows volatility to respond differently to past volatility information. The authors find that their model works better than previous models of their days and better captures the phenomenon that higher (lower) volatilities tend to be followed by higher (lower volatilities). A large literature, especially in the early days, follow similar structures and propose various ARCH-family models when dealing with past 'news' information to investigate the relationship between asset returns and volatility. Such models typically focus on market level instead of individual stocks. See, for example, Engle and Ng (1993), Pagan and Schwert (1990), and Chou (1988), among others.

On the other hand, more recent studies utilize textual information to investigate how *textual news* can be used to explain and predict asset volatilities and returns. This advance is made possible by recent developments in machine learning which allows us to extract information accurately, automatically, and efficiently from textual data, such as Manela and Moreira (2017).

Baker et al (2019) constructs a newspaper-based volatility tracker based on frequency of words from

three sets of topics the authors manually construct: economic, market, and volatility. The authors use newspaper articles from 4 newspapers where they can download data from: *Miami Herald, Dallas Morning News, San Francisco Chronicle,* and *Houston Chronicle*. The authors then construct word lists of 20 policies and classify each news into one or more of the policy news if the news contains any words in the word list. The authors find that 72% of EMV articles discuss the Macroeconomic Outlook, and 44% discuss Commodity Markets. Policy news is a major source of volatility. Each policy's contribution to volatility varies greatly over time.

The rest of this paper is constructed as follows. Section 2 gives details of our dataset and gives descriptive statistics of Our database. Section 3 gives methodology and results of XGBoost and construction of RSJ. Section 4 gives an economic explanation of portfolio returns by drawing references from literature on investor recognition, volatility, and volume. Section 5 concludes.

4.2 Data and Descriptive Statistics

In this section, we describe the sources of the dataset we employ in this study, data cleaning methods, and the descriptive statistics of features of our sample to get a first impression of our study.

4.2.1 Refinitiv Machine Readable News (MRN)

We use the same news dataset as Chapter 3 and use Secured Overnight Financing Rate (SOFR) as a proxy for one-day risk-fee rate for the US. SOFR is obtained from Federal Reserve Bank of St. Louis. High frequency stock price data is obtained from Thomson Reuters Tick History (TRTH) database. Fama-French factor data is obtained from the official website of Kenneth French[9]. We obtain two sets of sentiment scores, one set based on articles and alerts' headlines, the other based on articles' body contents, both of which are based on FinBERT model as applied to Refinitiv Machine Readable News based on all US companies from September 2000 to December 2019. Details can be found in the previous chapter. FinBERT yields two outputs: a 'most likely' label of positive, neutral, or negative sentiment; and the probability of the label. We convert the probability to sentiment score by 1) if sentiment label is neutral, the sore is 0; 2) if sentiment is negative, then

[9] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

negative of the probability is sentiment score; and 3) if sentiment is positive, then the probability is sentiment score.

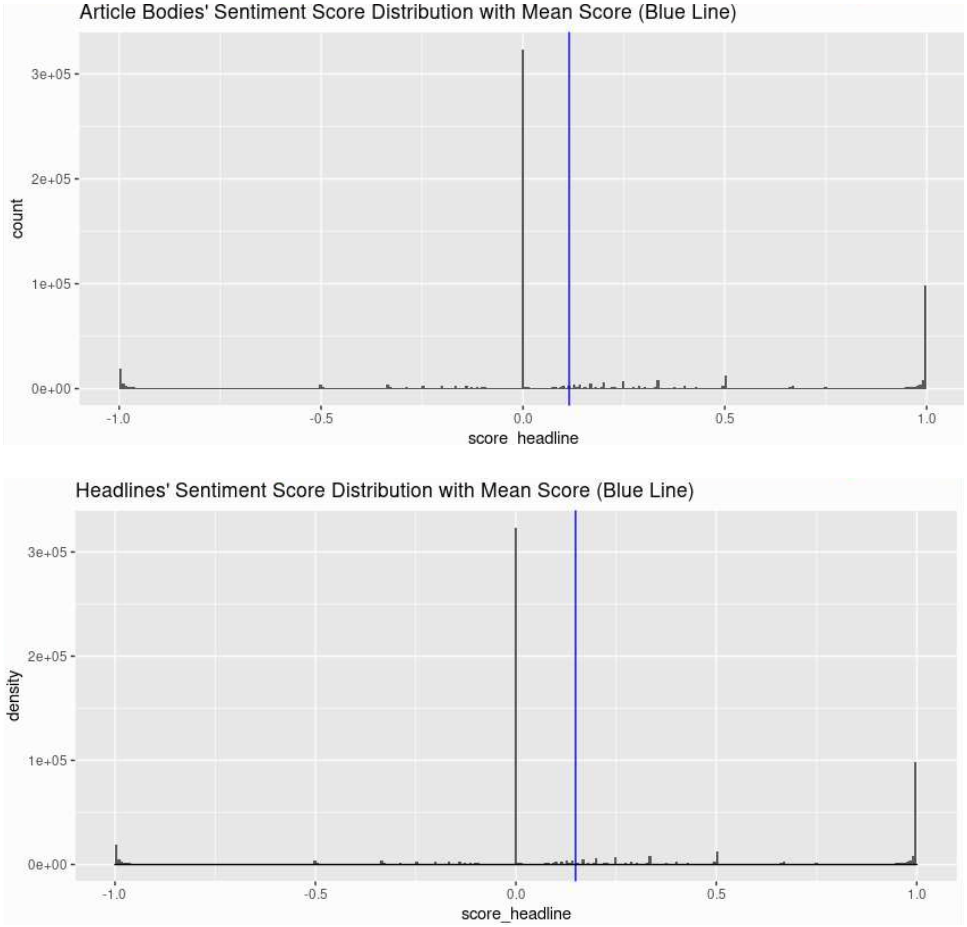


Figure 4-1 Distribution of the 44 companies' sentiment scores.

In total, we have 731,238 company-day sentiment scores based on headlines and 385,841 company-day sentiment scores based on article bodies. We take a first look at the sentiment scores in Figure 4-1.

As neutral labels dominate in both headlines and articles, we see a sharp peak of score 0; positive labels occur more frequently in both sets of sentiment scores and hence mean scores are both positive.

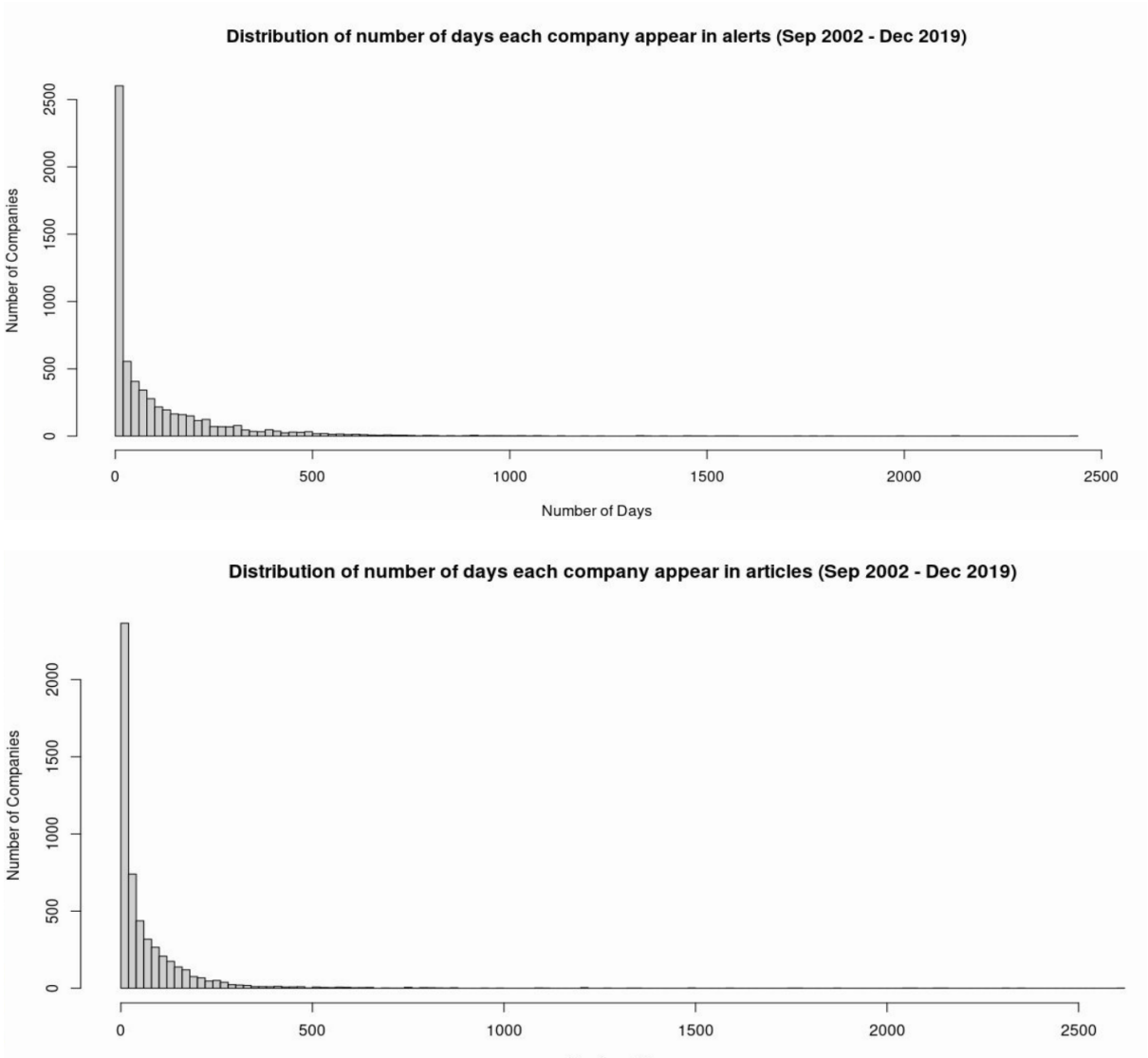


Figure 4-2 Distribution of number of days where each company has at least one piece of news per day

Another feature of our database is that news is concentrated among a small number of companies that attract public and media attention. This is clearly demonstrated in the histogram of number of days companies feature in news pieces. This is made even clearer in the quantile table showing number of company-days we have for news headlines and news bodies at 0%, 1%, 25%, 50%, 75%, 99%, and 100% percentiles from 2001 to 2019. See Figure 4-2 and Table 4-1.

Distribution of Number of Days Each Company Appear in News Alerts and Articles (Percentile)

	Mean	0%	25%	50%	75%	100%
N(Alerts)	107	1	5	37	140	2438
N(Articles)	73	1	4	27	88	2608

Table 4-1 Number of Days Each Company Appear in News Alerts and Articles

Without a better guide on minimum number of observations, we include a company into our analysis if it has at least 30 trading days for each company each year on average. This leaves us 44 companies in article analysis and 207 companies in headline analysis.

4.2.2 Companies and Sentiment

We consider 2014-2019 only[10]. We still require, on average, at least 30 company-days per year. This leaves us 44 companies in articles and 190 companies in headlines. We use the 44 companies that appear both in news articles and news alerts throughout this study. 3 companies are now delisted: Raytheon Co (RTN.N), Twitter (TWTR.N), and United Technologies Corp (UTX.N). The 44 companies are large companies in financial, manufacturing, automobile, airline, retail, food, consumer goods, high-tech, and financial industries and the full list is presented in Table 4-2.

RIC	n (articles)	n (alerts)	Company
BA.N	2834	5655	Boeing Company
GM.N	1977	3860	General Motors Company
TWTR.N	1874	2804	Twitter
JPM.N	1440	3284	JPMorgan Chase & Co
GS.N	1427	2549	Goldman Sachs Group
F.N	1395	3124	Ford Motor Co
WFC.N	1121	3255	Wells Fargo & Company
WMT.N	1056	2614	Walmart Inc
GE.N	1003	3585	General Electric Company
C.N	1000	3029	Citigroup Inc
XOM.N	991	2511	Exxon Mobil Corporation
BAC.N	822	2201	Bank of America Corporation

[10] We limit Our analysis to 6 years because high-frequency data is large and hence, we would easily exceed TRTH download restrictions with too many stocks. This is the same dataset we use in Chapter 3 section 3.5.5.

LMT.N	762	1988	Lockheed Martin Corporation
BLK.N	749	2405	BlackRock Inc
JNJ.N	688	2390	Johnson & Johnson
DIS.N	676	1636	Walt Disney Co
MRK.N	644	2431	Merck & Co Inc
MS.N	642	1873	Morgan Stanley
PFE.N	632	2550	Pfizer Inc
CVX.N	586	1998	Chevron Corp
ICE.N	566	1872	Intercontinental Exchange Inc
DAL.N	553	2363	Delta Air Lines Inc
T.N	544	2235	AT&T Inc
LLY.N	538	2486	El We Lilly and Company
NYT.N	531	702	New York Times Co
PCG.N	524	1787	PG&E Corp
MCD.N	513	1741	McDonald's Corporation
NKE.N	508	1543	Nike Inc
UAL.N	475	1669	United Airlines Holdings Inc
AIG.N	435	1653	American International Group
BX.N	435	1275	Blackstone Inc
CAT.N	427	2385	Caterpillar Inc
FCX.N	401	1798	Freeport-McMoRan Inc
VZ.N	400	1773	Verizon Communications Inc
TGT.N	383	1715	Target Corp
BMY.N	375	1955	Bristol-Ourers Squibb Company
LUV.N	361	1857	Southwest Airlines Co
KO.N	348	1499	Coca-Cola Co
UTX.N	337	1597	United Technologies Corp
UPS.N	336	1582	United Parcel Service Inc
IBM.N	334	1598	International Business Machines Corporation
MDT.N	317	1774	Medtronic PLC
BK.N	309	1227	Bank of New York Mellon Corp
RTN.N	302	1274	Restaurant Group PLC
Total	**32,571**	**97,102**	

Table 4-2 List of 44 companies with at least 30 news-days each year from 2014 to 2019.

%News of positive, negative, and neutral sentiment			
	Positive	Neutral	Negative
Alerts	23.5	62.06	13.9
Article (title)	15.3	70.5	14.2

Article (body)	15.0	77.2	7.8

Table 4-3 Distribution of sentiment scores in sample

We have a total of 97,102 news alerts and 32,571 news articles. We have predominately neutral sentiment labels for both news alerts and articles as shown in Table 4-3.

错误!未找到引用源。 shows contemporaneous correlation between RSJ and return, and correlation between today's RSJ and next-day return for each of the 44 companies. We first note that there's strong positive correlation between contemporaneous RSJ and same-day return, but the correlation reverses when considering today's RSJ and tomorrow's return. Different companies exhibit very different correlations of contemporaneous RSJ and close-to-open return. Most correlations are very strong with United Airlines being the strongest at 0.755, and the weakest also exhibit a moderately strong correlation of 0.455 for PG & E. The average correlation between each day's RSJ and return remains high at 0.569. When considering the correlation between today's RSJ and tomorrow's return, however, the pattern is less clear. Some stocks exhibit positive correlations, while others exhibit negative correlations. The magnitudes are consistently low. While there appears to be no ambiguity involved in contemporaneous correlations between RSJ and stock return, the reversal in correlation between today's RSJ and tomorrow's RSJ needs further investigation, and we hypothesis that news of certain types induces more negative and predictable RSJ (volatility jump) reversal. Therefore, it is critical to find a way to predict next-day RSJ. By doing so, ex-ante, we should achieve higher portfolio performance compared with Bollerslev et al (2020). As events trigger price and volatility jumps, and firm events are captured by firm news, we hence believe that using firm news would help predict next-day volatility jumps, which can be sued to form portfolios today.

		News heterogeneity			
		length of article body	Number of audiences	Number of Subjects	Percentile of RV
alerts	Negative		3.7	23.3	73.2
	Neutral		3.6	23.9	64.5
	Positive		3.3	21.8	71.3
	F-Stat		**174.2**	**203.1**	**697.1**
article body	Negative	191	8.6	32.5	70.8
	Neutral	175	7.9	31.8	56.8
	Positive	187	7.5	28.1	76.0

	F-stat	**43.35**	**90.6**	**84.3**	**965.2**
article title	Negative		7.6	34.1	66.2
	Neutral		8.1	31.2	57.6
	Positive		7.6	29.3	70.4
F-stat			**32.8**	**48.3**	**433.2**

Table 4-4. News heterogeneity.
This table shows the length of news articles' body contents, news' number of audiences, news' number of subjects, and the percentile of each news-day's stock by considering 20 days before and 20 days after each day.

We start by asking if there is heterogeneity in positive, negative, and neutral sentiment news. We table the differences in length of article body, number of subjects, and number of audiences for alerts, news articles' headlines, and news articles' bodies together with their F-statistics in Table 4-4[11]. All p-values are lower than 0.00 where I refrain from showing in the table to be parsimonious. We note that news does appear to be different in their length, number of subjects, and number of audiences the news is tailored to. All days with news come with more volatile stock-days. '*Audience'* in MRN database technically contains the products the investor subscribes to. It is defined as: '*Product codes identify which desktop news product(s) the news item belongs to. They are typically tailored to specific audiences. Examples: "M" for Money International News Service and "FB" for the French General News Service*'[12].

We also check if they differ in volatility. To do so, we compute 5-minute realized volatility, semivariances, and realized signed jumps (the difference between positive and negative semivariance scaled by total realized variance) for each stock-day and ask the percentile of the volatility measure from 20 days before and 20 days after a specific date. This hence can be considered a 'local percentile' of volatility. We see that there is also heterogeneity in volatility percentiles. It hence appears that positive, neutral, and negative news are intrinsically different and warrant investigation on their own. Previous studies which consider only sentiment-carrying news, i.e., positive and negative news only, potentially gives up valuable information. While the results on realized volatility is probably intuitive, it's probably surprising that there are statistically different

[11] This is the same table as Table 3.8; we reproduce here for reader's convenience.

[12] This is taken from the official user guide of *Refinitiv Real-Time News: Feed and Archive*, version 2.15. This document is available on request with Refinitiv.

length of article, number of news audience, and number of subjects in positive, neutral, and negative news. In my working paper, we have shown that portfolios selected from news article sentiment gives much higher portfolio performance than portfolios selected from news alerts sentiment, suggesting that news articles are more informative than news alerts. Here, we separately consider the effects of news alerts and news articles by also forming portfolios based on news alerts and news articles.

4.2.3 Topics

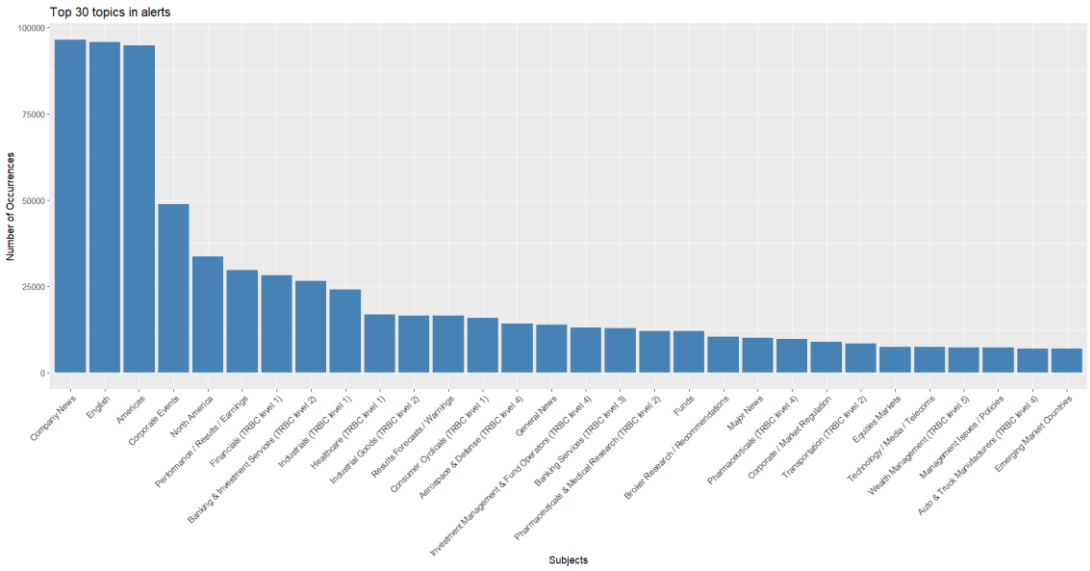
We note that not all the 2,400 topic codes are covered in the sample. There are a total of 1,268 topic codes in news alerts and 1,398 topic codes in news articles. While the number of topic codes are too many to show, we show the categories the subject codes belong to in Table 4-5.

Subject category	Number of subjects under this category
Business Sector	1034
Geography	380
Commodity	143
Health	117
Asset Class / Property	99
Sport	95
Event Type	88
Organization	85
Genre	44
Government / Politics / International Affairs	42
Language	38
Science	33
Indicator Type	25
Legacy News Topic	21
Society / Social Issues	21
Broad News Topic	18
Corporate	17
Environment / Accident / Disaster	17
Physical Asset Type	17
News Flag / Status	16
Crime	12
Sporting Event	11
Arts / Culture / Entertainment	9
Money / Finance	9
Sports-related	7

Religion	6
Sport Combined with Geography	5
Legal	1

Table 4-5 Distribution of topics under each category

While the subject codes fall into 28 categories, the topics they cover are very different, with overwhelmingly 1,034 topic codes dealing with the business sector, most of which cover which industry the company's business belongs to. This is followed by geography, which shows which area, state, or other geographic locations the company operates in, or where product/service is sold to. Going down the table, we see a large range of topics covered by the topic codes, although the coverage of topic codes is very unbalanced, with 6 categories having less than 10 relevant topic codes.



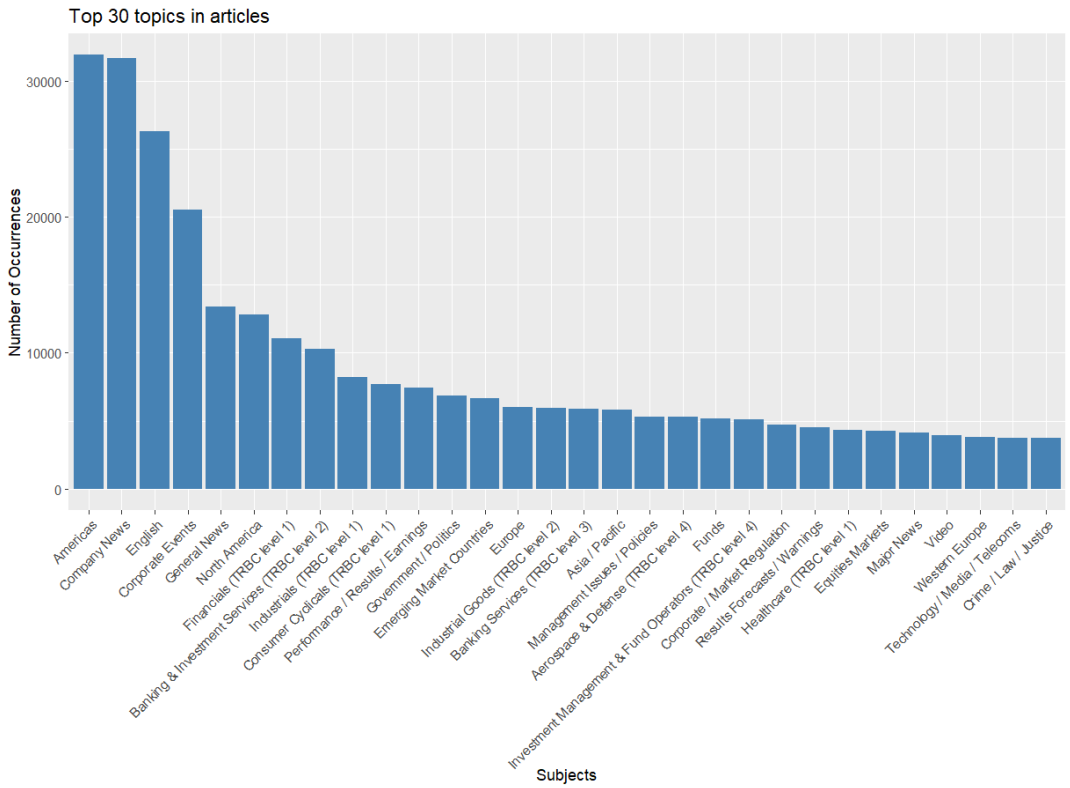


Figure 4-3 Distribution of topics in articles

The top 30 topics covered in alerts and articles are in Figure 4-3. While there are a large variety of topics covered, apparently not all topics are relevant. For example, almost all news is about '*Americas*' '*Company News*' and they are written in '*English news*'[13], which do not carry meaningful inferences. In terms of feature engineering, as the number of topic codes is too huge to investigate individually, we include all variables and allow XGBoost to select the most useful ones[14]. We may consider XGBoost a 'automatic feature engineering' as it only selects the most useful variables to include in its model.

[13] We note there are a small amount of news mis-labeled. For example, all news in the dataset is in English and is about US companies, but not all of them carry labels of '*Company News*', '*English*', and '*Americas*'.

[14] For readers less familiar with tree models: XGBoost selects the variables that give the largest improvement in objective function. It hence will only select the most '*useful*' or '*relevant*' variables.

4.3 Methodology and Results

In this section, we present the methodologies we employ in this study. After each methodology section, we immediately present results.

4.3.1 Realized semi-variance, signed jump variation, and realized signed jump

We first note that the terms we use in this study is consistent with Patton and Sheppard (2015). While different studies use different terms, they only differ by how their terms are named.

Consider a continuous-time stochastic process for log-prices, p_t which consists of a continuous component and a pure jump component:

$$p_t = \int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} (\Delta p_s)^2$$

Where $\Delta p_s = p_s - p_{s-}$ represents price jump, if any.

Realized variance proposed in Andersen et al (2001) is commonly used in high frequency data calculated as the sum of squared returns. Realized semivariance breaks down realised variance stemming from positive and negative part of return and is defined as:

$$RSV^- = \sum_{i=1}^{n} r_i^2 I_{r_i<0}$$

$$RSV^+ = \sum_{i=1}^{n} r_i^2 I_{r_i>0}$$

And apparently, $RV = RSV^+ + RSV^-$. Barndorff-Nielsen et al (2008) shows that in the limit, realised semivariance can be written as sum of: 1) a continuous part of price process and 2) a jump component, and positive and negative parts of semivariance simply converge to:

$$RSV^+ \to \frac{1}{2}\int_0^t \sigma_s^2 ds + \sum_{0\leq s\leq t} \Delta p_s^2 I_{\Delta p_s>0}$$

$$RSV^- \to \frac{1}{2}\int_0^t \sigma_s^2 ds + \sum_{0\leq s\leq t} \Delta p_s^2 I_{\Delta p_s<0}$$

We can hence easily remove the continuous part and obtain the price jump part of realised variance.

$$\Delta J^2 \equiv RSV^+ - RSV^- \to \sum_{0\le s\le t} \Delta p_s^2 I_{\Delta \mathrm{p_s}>0} - \sum_{0\le s\le t} \Delta p_s^2 I_{\Delta \mathrm{p_s}<0}$$

This is called *signed jump variation (SJV)* in Patton and Sheppard (2015). As different stocks by their nature differ drastically, to allow comparison between different stocks' SJVs, Bollerslev et al (2020) defines *relative signed jump variation (RSJ)* by normalizing SJV by total realised variation, which is bounded between -1 and 1.

$$RSJ_t = \frac{SJV_t}{RV_t}$$

Patton and Sheppard (2015) give a detailed implementation procedure to calculate SJV using high frequency data. We use the same implementation strategy in this paper. The same procedure is also used in Bollerslev et al (2020). Let log tick prices of a stock on a trading day be $p_0, p_1, \dots, p_n$ between 9:30:00 am and 4:00:00 pm and on each trading day, there are $n+1$ tick updates. Then to sample every 5 minutes, on average, we need 78 samples. Realised variance calculated uniformly during trading time starting from the j^{th} observation is then:

$$RV^j = \sum_{i=1}^{78} \left(p_{\lfloor ik+j\delta \rfloor} - p_{\lfloor (i-1)k+j\delta \rfloor}\right)^2$$

Where $k = \frac{n}{78},\ \delta = \frac{n}{780},\ \lfloor . \rfloor$ rounds down to the next integer.

Consistent with literature, Patton and Sheppard (2015) uses event time (business time) sampling instead of calendar time sampling per 5 minutes. The authors note that sampling in event time produces estimators with superior statistical properties and this is common in literature, such as Bollerslev and Todorov (2011) and Barndorff-Nielsen et al (2008). They use the first and last observation of the day as their first and last datapoint, and sample evenly during the day for the rest 77 observations.

4.3.2 XGBoost regression

Because of its technicalities, interested readers may refer to Appendix 4.6.1for technical details on XGBoost and its estimation. Here, we simply note that XGBoost is one of the most powerful and popular algorithms for modelling structured tabular data. The authors of XGBoost have shown that XGBoost's performance is much better than competing algorithms; it's also well-suited for sparse

dataset, such as in our case, where we have around 1,400 topic codes but on average, each piece of news only have around 30 topic codes, hence leaving a sparsity of around 98%. More technical details the authors' comparisons with other algorithms can be found in the original paper of Chen and Guestrin (2016).

4.3.3 Baseline portfolios

Before running XGBoost regression, we first test two baseline portfolios by forming portfolios based on 1) today's SJV only and 2) today's news sentiment only. Specifically, we form a zero-cost long-short portfolio by going long (short) stocks in the bottom (top) 10 percentile of RSJ for the total sample and going long (short) stocks in the top (bottom) 10 percentile of sentiment. We test sentiment portfolio separately for headline and article sentiments. This baseline portfolio achieves an average daily excess log-return of 0.0003, standard deviation of 0.0078, and annualized Sharpe ratio of *0.619* when sorting only on *RSJ*. When sorting only on *sentiment*, article portfolio achieves an average daily excess log-return of 0.0019, standard deviation of 0.015, and annualized Sharpe ratio of *1.94,* and alert portfolio achieves an average daily excess log-return of -0.001, standard deviation of 0.023, and annualized Sharpe ratio of *-0.708*. The difference in alert and article sentiment portfolio again suggests different informational contents of alerts and articles, and probably surprisingly, using stocks from large companies based solely on their sentiment is not a profitable business. This suggests that previous research which rely on only journal article (such as Wall Street Journal), or news headlines are potentially sub-optimal; and their results can potentially improve when they use the information in news body contents instead.

4.3.4 Predictive XGBoost regression using topic codes

Now, we test the predictability of RSJ and use predicted RSJ to form portfolio. We run a predictive XGBoost regression using next-day *RSJ* as target variable (regressand) and all the topic codes as features (regressors), controlling for the stock's realized volatility, number of subjects of news, number of audiences of news, today's positive and negative part of realized volatility, and today's sentiment scores of each stock. We experiment different hyperparameters and find the best-performing model that achieves bias-variance trade-off[15]. Specifically, we use a customized two-

[15] For readers who are unfamiliar with machine learning and boosting algorithm: XGBoost is extremely good at fitting to

stage model choice procedure: we first run XGBoost regression using RMSE of test set as training termination criteria, then find the model's corresponding portfolio performance in the test set. We decide on the model that gives the highest Sharpe ratio in the test set. We perform analysis separately for articles and alerts.

We note that a very large number of topic codes appear only in a very small amount of news. Excluding topic codes for *United States, Americas, Company News, Corporate Events, and English* (the largest topic groups which also tend to carry no meaningful implications), we plot the distribution of all topic codes' frequencies in train, test, and validation sets and tabulate their summary statistics. See Figure 4-4 and Table 4-6.

Dataset	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Train	0	2	10	201.4	70.7	14,382
Test	0	0	1	31.1	11.0	2,229
Validation	0	1	3	82.3	29.0	6,976

Table 4-6 Number of Times Each Topic Appear in Train, Test, and Validation Set

training set, but it fits so well that when the model gets complex, it will look for the purely noisy and random part of the data, this is what we call the algorithm 'memorizes' the dataset and overfits, and the resulting model will not work well for models it has not seen. We avoid this overfitting problem by controlling model complexity which makes the model works less well on training set but much better on validation and test sets. This is called bias-variance trade-off.

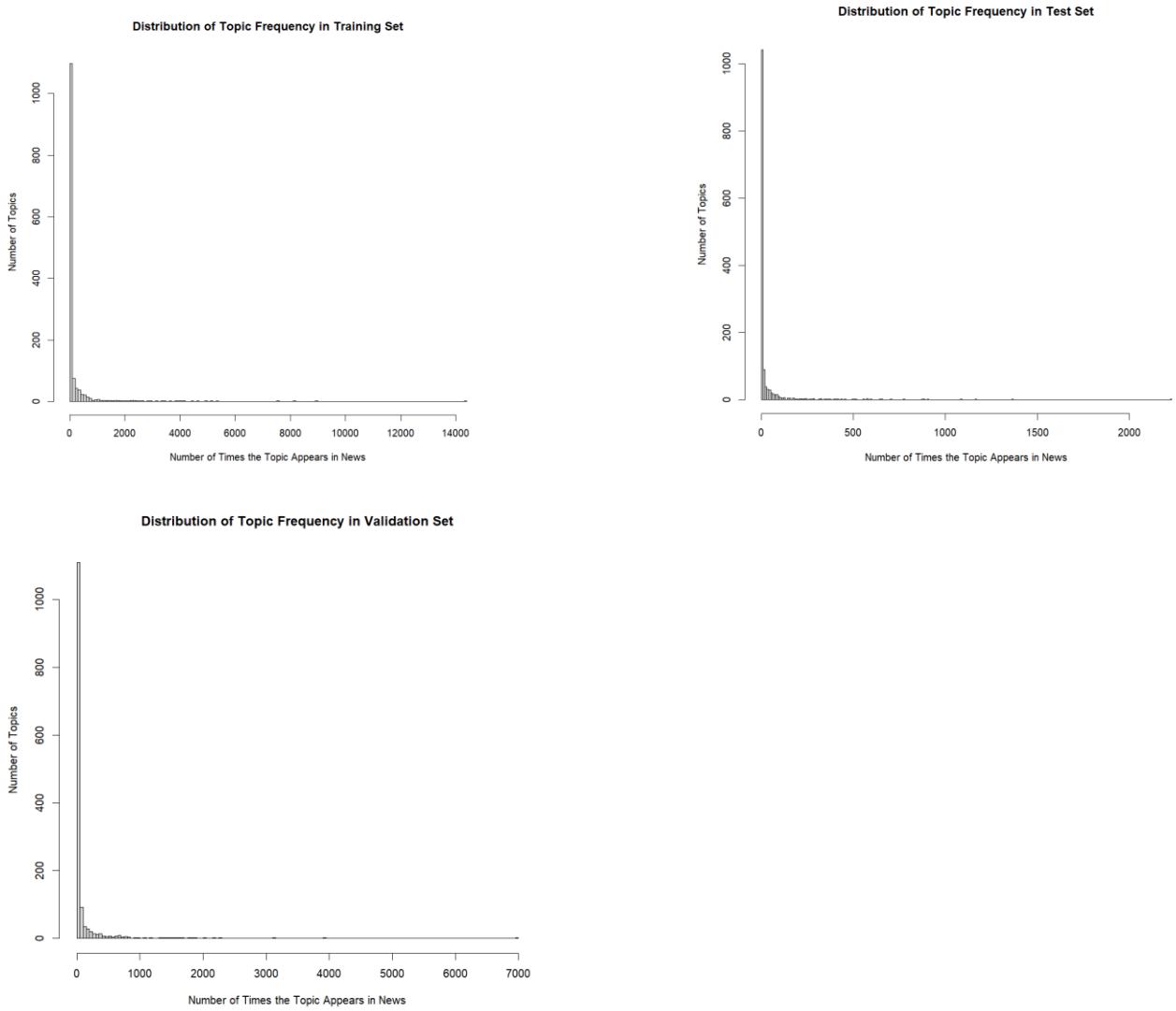


Figure 4-4 Distribution of topics in train, test, and validation sets

Where the largest topic code represents news on *corporate events* and is again not carrying useful implications. We note that the distribution of topic coverage is extremely unbalanced, with 50% of topics covering under 10 news pieces. We acknowledge that data limitation is an issue here and with sufficiently large dataset (for example, with at least 20 data points[16] per topic code), our analysis

[16] This choice of number of observations per data point is subjective. The idea is to have a 'sufficiently large' dataset which covers each topic.

performance would greatly improve, and likely, portfolio performance would further improve.

We split the total sample of 2014-2019 (inclusive) to two parts: 2014-2018 (inclusive) for data modelling, and 2019 for validation. We further split the dataset of 2014-2018 to two sets: training set (80% of total sample) and test set (20% of total sample). We train the model using training set, fine-tune using test set to choose the best model to get the best hyperparameters and validate the model using the validation set. As the validation set of 2019 data is something the model has not seen in both training and fine-tune phase, it provides a direct indicator of model goodness (or badness).

For alerts, we use XGBoost with L1 regularization[17] to control for model complexity. We carefully fine-tune the model to avoid overfitting. The resulting portfolios achieves an average daily excess log-return of 0.0013 (0.0004) , standard deviation of 0.013 (0.0323), and annualized Sharpe ratio of 1.63 (0.1944) for equal weighting (value weighting). Total number of days covered is 118 trading days for financial year 2019. We have only 2.067 stocks on average in the portfolio. RMSE of predicted next-day RSJ in training (test) set is 0.280 (0.282). When rebalancing the EW portfolio positions weekly, we make a loss where annualised Sharpe ration is -1.09. This is from negative return where we make a loss of -0.015 with a standard deviation of 0.010. We note that only 233 variables (including controls) are selected into the model. We note that there's significant improvement over portfolio selection based only on news alerts' sentiment scores and RSJ brought about mainly by higher return.

For articles, we use XGBoost with L1 regularisation and again carefully fine-tune the model. The resulting portfolio achieves an average daily excess log-return of 0.0014, standard deviation of 0.0105, and annualized Sharpe ratio of 2.06. The drastic improvement in Sharpe ratio is mainly from lower variance. It is an excellent improvement over news alerts models as we can achieve a higher return with lower volatility. There are only 5.1 stocks on average in our portfolio over 247 trading days where we form portfolios in. Because of limitation in data availability, the small number of news and hence stocks to select into portfolio suggests that we essentially have a risky position. With sufficiently large dataset, we believe the portfolio performance should further improve. RMSE

[17] L1 regularisation is similar to Lasso regression in OLS. It's used to reduce model complexity.

of predicted next-day RSJ in test set is 0.325. The higher portfolio return of article models again suggest that articles are more informative and previous studies which rely on just headlines are sub-optimal. The portfolio is profitable in 57.6% trading days. The portfolio's long leg sees a much lower return but also much lower volatility compared with the short leg: while the long leg has a daily return of 0.00054 and standard deviation of 0.011, the short leg sees a daily return of 0.003 and standard deviation of 0.020. Overall, the short leg appears to be more profitable in terms of return-risk trade-off. Raising holding periods and using value-weighting on daily holding do not increase performance, where daily return, standard deviation, and annualised Sharpe ratios are -0.024 (0.0017), 0.149 (0.0378), and -1.17 (0.7259).

We now look at what topics (and control variables of volatility and sentiment measures) are important for predicting next-day return.

We report here only features that appear at least 10 times in the training set. Interested readers may refer to Appendix C for results covering all features and for models using news alerts. Considering only topics that appear at least 10 times, for news alerts and news articles, the most important topics and corresponding importance scores are presented in Table 4-7 and Figure 4-5.

Top Predictive Variables for Next-Day SJV (News Alerts)

Order	Gain	Cover	Frequency	Description	Type
1	0.179	0.026	0.026	Mortgage Application Data	Indicator Type
2	0.085	0.013	0.013	Residential Mortgage-Backed Securities	Asset Class / Property
3	0.038	0.015	0.015	Pipelines	Physical Asset Type
4	0.022	0.015	0.015	Property & Casualty Insurance (NEC) (TRBC level 5)	Business Sector
5	0.020	0.019	0.019	Negative Realised Signed Jump	Business Sector
6	0.019	0.018	0.018	Pension Funds (TRBC level 5)	Business Sector
7	0.019	0.017	0.017	Investment Banking (TRBC level 5)	Business Sector
8	0.014	0.012	0.012	Nuclear Armaments / Nuclear Proliferation	Government / Politics / International Affairs
9	0.014	0.015	0.015	Healthcare Facilities & Services (TRBC level 4)	Business Sector
10	0.012	0.011	0.011	Capital Movement Data	Indicator Type
11	0.012	0.010	0.010	Islam	Religion
12	0.012	0.015	0.015	Banking Capital and Liquidity Requirements	Event Type
13	0.010	0.013	0.013	Mongolia	Geography
14	0.010	0.005	0.005	Syria	Geography
15	0.009	0.010	0.010	Government Borrowing Requirement	Indicator Type
16	0.009	0.009	0.009	Middle East (Energy)	Geography
17	0.009	0.015	0.015	Precious Metals	Commodity
18	0.009	0.014	0.014	Cocoa	Commodity
19	0.008	0.011	0.011	Philippines	Geography

Top Predictive Variables for Next-Day SJV (News Articles)

Order	Gain	Cover	Frequency	Description	Type
1	0.004	0.004	0.004	Workforce	Event Type
2	0.004	0.004	0.003	Products / Services	Event Type
3	0.003	0.003	0.003	Key Personnel Changes	Event Type
4	0.003	0.003	0.002	Miscellaneous Specialty Retailers (TRBC level 4)	Business Sector
5	0.003	0.003	0.003	Wealth Management (TRBC level 5)	Business Sector
6	0.003	0.003	0.005	Management Issues / Policies	Event Type
7	0.003	0.003	0.002	Organizational Restructuring	Event Type
8	0.003	0.003	0.003	South America / Central America	Geography
9	0.003	0.003	0.002	Insurance (TRBC level 3)	Business Sector
10	0.003	0.002	0.005	Corporate / Market Regulation	Event Type
11	0.003	0.002	0.003	Transportation (TRBC level 2)	Business Sector
12	0.003	0.002	0.004	Funds	Asset Class / Property
13	0.003	0.002	0.003	Debt / Fixed Income Markets	Asset Class / Property
14	0.003	0.002	0.004	Europe	Geography
15	0.003	0.003	0.003	Mergers / Acquisitions / Takeovers	Event Type
16	0.003	0.002	0.004	Performance / Results / Earnings	Event Type
17	0.003	0.003	0.002	Company News	Broad News Topic
18	0.003	0.002	0.004	Consumer Cyclicals (TRBC level 1)	Business Sector
19	0.003	0.002	0.004	Banking Services (TRBC level 3)	Business Sector

Table 4-7 The most important topics in predicting next-day RSJ

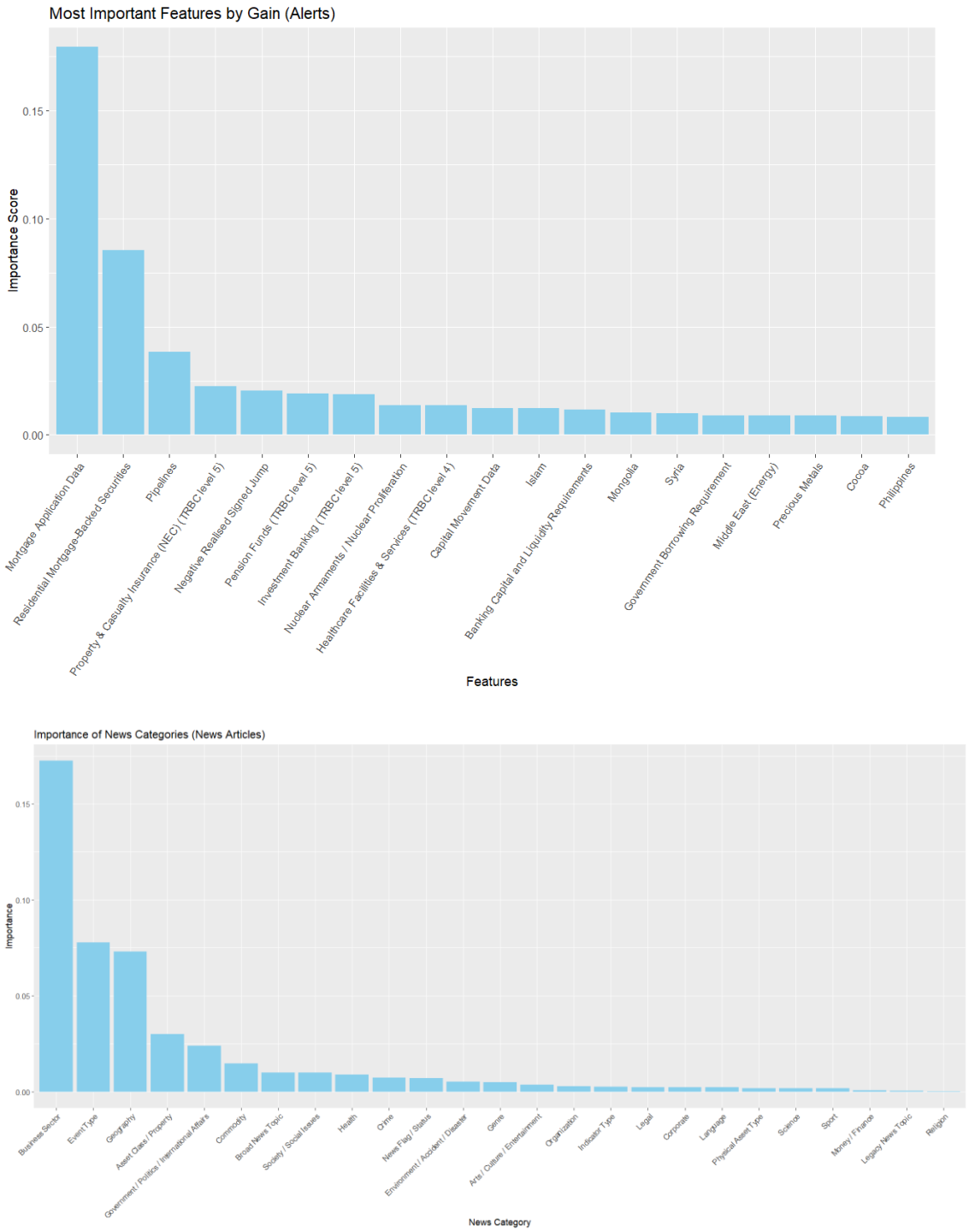


Figure 4-5 Most important news categories under XGBoost regression for predicting next-day RSJ

We rely on Gain, which is reduction in objective function when using each feature versus not including it in model, to assess the importance of the variable. We also report Cover and Frequency

for interested readers. A large variety of topics are covered in the top topics including which geographic location the company operates in or where product is tailored to, what business sector the company or product relates to, disease, some prevailing events, et cetera. We do not attempt to identify the exact topics that are relevant, but instead, we emphasize that, some topics are relevant, and some are irrelevant, for predicting tomorrow's volatility jumps. XGBoost provides a way of automatically selecting from the huge number of topics the relevant ones and using the predicted model to predict tomorrow's volatility jump, which provides a basis of portfolio selection, is highly profitable.

We note that importance score is a relative measure of feature importance as it compares model performance when including against not including the feature into the model, hence importance score can only be considered in relative terms. Also, the variables' economic significance (magnitude of parameter) is not summarized by the importance score.

Coupled with the fact that each topic can be further grouped into one of the 28 groups, the impact of volatility and today's negative part of realized variance is even less significant.

Importance of News Categories in Descending Order of Gain (News Articles)				
Order	News Category	Gain	Cover	Frequency
1	Arts / Culture / Entertainment	0.004	0.004	0.0030
2	Asset Class / Property	0.030	0.032	0.0261
3	Broad News Topic	0.010	0.009	0.0124
4	Business Sector	0.172	0.174	0.1403
5	Commodity	0.015	0.018	0.0109
6	Corporate	0.002	0.003	0.0013
7	Crime	0.007	0.007	0.0061
8	Environment / Accident / Disaster	0.005	0.006	0.0037
9	Event Type	0.078	0.083	0.0726
10	Genre	0.005	0.007	0.0033
11	Geography	0.073	0.082	0.0603
12	Government / Politics / International Affairs	0.024	0.026	0.0217
13	Health	0.009	0.010	0.0057
14	Indicator Type	0.002	0.004	0.0016
15	Language	0.002	0.002	0.0040
16	Legacy News Topic	0.000	0.001	0.0001
17	Legal	0.002	0.002	0.0026
18	Money / Finance	0.001	0.001	0.0004

19	News Flag / Status	0.007	0.007	0.0082
20	Organization	0.003	0.003	0.0017
21	Physical Asset Type	0.002	0.003	0.0010
22	Religion	0.000	0.000	0.0001
23	Science	0.002	0.002	0.0012
24	Society / Social Issues	0.010	0.011	0.0078
25	Sport	0.002	0.003	0.0012

Table 4-8 Most important features by news category

We now group the topic codes into the 28 categories and show in Table 4-8. We do not have 28 news categories because only around 12% of all topics covered in the news are selected into the model, which do not cover all the categories, suggesting that some categories in the sample are not useful for predicting tomorrow's RSJ. We note that this implies two possibilities: 1) the excluded categories are indeed irrelevant, or 2) we just do not have a large enough dataset to explore their effects. Future research may attempt to combine Refinitiv MRN database with other news databases for a more comprehensive dataset. Business sector is the largest source of predictability of next-day RSJ, followed by Health, Geography, and Commodity. Our results suggest that by considering what happened in the market, and hence what topics are covered in news, we can greatly improve Our ability to predict the next-day volatility jump, RSJ, and this helps form portfolio.

4.4. Potential explanations

We now attempt to shed some light on why and how we achieve the outstanding portfolio performance when exploring topic information. The focus in this section is on the model using news articles.

4.4.1 Information and asset returns

There have been several theoretical and empirical studies on how news affect asset prices through information channel. Veldkamp (2006) is a theoretical work on media frenzies and information in the financial market. The author shows that media frenzies, which is supplying large amount of information, raises asset prices by reducing the uncertainty about the asset's payoffs. Therefore, when the asset valuation is high, media frenzy would arise, and would facilitate the asset prices to induce higher asset returns.

In another paper, Veldkamp (2006) introduces markets for information that generate high price covariance within a rational expectations framework. This paper provides a rational explanation for

the observed excess covariance of asset prices: asset prices' covariances are too high relative to covariances between their fundamentals. There had been irrational, behavioral explanations to this issue (see, for example, Barberis et al, 2005). In this paper, Veldkamp (2006) argues that when different investors observe the same information, such as macroeconomic data, that affects different assets' prices, prices, and returns of these assets would comove even if these assets' fundamentals are not correlated. When analyst coverage is higher, however, we should observe lower co-movement because investors can rely on the more accurate and relevant information for each company.

4.4.2 Investor attention and asset prices

Today, we receive a huge amount of information through news from various sources. Apparently, it's impossible to respond to all of them because of our limited time, effort, and cognitive abilities. Subsequently, investors need to make a choice as to what information to read and react to, and what information not to respond. In other words, news will naturally receive different levels of attention (Kahneman, 1973). This hinders information from being immediately incorporated into asset prices, shedding doubt on traditional finance models which require asset prices to instantaneously reflect all new information. Hirshleifer et al (2009) shows that investors get distracted and when there are more firms making announcements, price reactions are smaller, and prices take longer times to reach equilibrium.

Following Kahneman's pioneer inter-discipline work, many researchers have attempted to investigate how investor attention (and inattention) causes investors' differential information processing, trading behaviors, and hence return and volatility dynamics. Unsurprisingly, many studies find connections between investor inattention and slow diffusion of information. See, for example, Hendershott et al, 2013; Da et al, 2014; Peng and Xiong, 2006.

The literature of rational (in-)attention states that investors pay attention only when it is profitable to do so. The cost involved in financial attention is not just our cognitive capacities; it also includes but is not limited to time dedicated, opportunity costs, monetary costs involved in data analysis, et cetera. For theoretical frameworks, see, for example, Sims (2003). However, individuals selectively pay attention to information (even when it is costless to them), even though having more information would improve their decision-making (Oster et al, 2013; Thornton, 2008; Caplin and Eliaz, 2003; Ehrlich et al, 1957; Sicherman et al, 2017). One explanation is the ostrich effect of

Karlsson et al (2009) which states that investors should pay more attention to their finances after good news than bad news, and subsequently, investors pay more attention to their portfolios after market goes up. Another possible explanation is realization effect, which states that trading, which leads to realized gains and losses instead of unrealized paper gains and losses, cause higher changes in utility. Both are psychological, behavioral explanations to investors being rationally inattentive.

Da et al (2011) is one of the first attempts of using Google search volume as a direct measure of investor attention. Previously, investors have used indirect measures, including but is not limited to 1) extreme returns (Barber and Odean, 2008); 2) trading volume (Gervais et al, 2001; Hou et al, 2009, among others); 3) price limited (Seasholes and Wu, 2007); 4) cost of advertising (Lou, 2014; Grullon et al, 2004). These measures typically make the same assumption: if there is a shock (or unexpectedly high increase) in any of the variables, investors must have paid attention to the stock. However, by their nature, this assumption is weak, because there are many other factors affecting asset prices, and even if some news appears on media, we do not have guarantee investors read it (Da et al, 2011). This leads to their attempt to propose a direct measure of investor attention of the public by asking often do investors search for stock ticker symbols on Google. As investors purposefully search for companies they are interested in, it also avoids the issue of whether investors read the news (or information). The authors focus on large stocks in the Russel 3000 index. Using weekly data, the authors find that their abnormal Google Search Volume Index (SVI) (log-SVI during the current week minus the log-median SVI during the previous eight weeks) leads other indirect measures of investor attention (such as extreme returns), indicating that investors may start to pay attention to a stock in anticipation of a news event. The authors also find that their attention measure and news sentiment measure are not correlated, and that the SVI mainly measures retail investors' attention.

Building upon previous research, Sicherman et al (2017) provides recent evidence on investor attention. The authors use online account logins to directly measure investors' attention to their portfolios. Having access to a large amount of data, the authors first document some features of investor attention and inattention: investors pay much more attention to their portfolios than they trade, consistent with theoretical frameworks that investors get hedonic utility from merely paying attention to or avoiding information (Loewenstein, 1987; Brunnermeier and Parker, 2005); investors' behaviors are consistent with the ostrich effect at daily, weekly, and monthly return

windows; there is a negative relationship between VIX and investor attention. Demographically, male and more wealthy investors are more prone to the ostrich effect, but they also pay more attention on average; while investors holding more bonds are less prone to the ostrich effect (because their holdings are less risky), when holding only bonds, investors would pay more attention when the market is going down. Additionally, ostricity appears to be a personal character. Terming *conditional trading* as investors who trade after logging into their accounts (i.e., trading after they are paying attention), they authors find that trading is positively correlated with news media stock market coverage. As conditional trading is also negatively correlated with VIX, the net effect of trading is ambiguous. Overall, the authors suggest behavioral explanations to trading: the observed increase in conditional trading when market is going down is consistent with stop-loss and bargain-hunting; when investors experience positive returns from their past trading, they would subsequently trade more because of they are paying attention to their portfolio returns, instead of because of conditional trading. Specifically, after controlling for attention, trading and prior returns are negatively correlated. Therefore, following profits, investor attentions would be higher while trading net of attention is lower.

Ben-Rephael et al (2017) investigates how institutional investor attention affects and facilitates price discovery. In this study, the authors propose a new measure of institutional investor attention by using the news searching and news reading activity for specific stocks on Bloomberg terminals, which are mainly used by institutional investors. The authors' proposed measure is based on a feature of Bloomberg: Bloomberg records how many times each article is read, and how many times readers actively search for each stock's news. These investor attention scores are then ranked against the same stock's numbers with the last 30 days to normalize. The authors then define abnormal institutional attention (AIA) which takes a value of 1 if there has been a spike in investor's attention for a given day. Interestingly, institutional investors also show a decrease in attention from Monday to Friday, consistent with retail investor attention in Liu and Peng (2015). The authors then consider earnings announcements as sources of news and how this information is incorporated into asset prices. Specifically, the authors show that institutional attention responds more quickly to major news events, leads retail attention, and facilitates permanent price adjustment. Using a full data of around 2,000 firms, the authors further show that we observe post-earnings announcement drifts in stock prices when institutional investors fail to pay attention to the announcements.

While some find that investor attention resolves uncertainty and leads to more efficient prices, others find that they would trigger higher behavioral biases, leading to a comprehensive study of Jiang et al (2022). In this study, the authors investigate the relationship between investor attention and 17 asset pricing anomalies, most of which are firm-level accounting ratios, such as net operating asset, idiosyncratic volatility, earnings-to-price ratio, net stock issues, gross profitability, et cetera. The authors find that the anomaly arbitrage portfolios generate significantly higher Fama–French three-factor abnormal returns following high-attention days than following low-attention days. The authors further show that large traders trade on anomalies more aggressively than small traders to realize the arbitrage profits more quickly.

4.4.3 Trading volume

Gervais et al (2001) is one of the first few systematic studies on high-volume premium. In this research, the authors find that stocks with unexpectedly high (low) weekly trading volume would appreciate (depreciate) during the following month. The authors attempt to explain this trend by arguing that positive shocks in a stock's trading would increase its visibility, hence increasing its demand in the due course, vice versa. The authors attempted to explain the return using return correlations, firm announcements, market risk, and liquidity. However, none of them are significant. In developing their stock visibility explanation, the authors rely on results from Miller (1977) and Mayshar (1983), which state that holders of a stock are most optimistic about its prospects on average. Their argument is largely based on the assumption that investors can only sell what they already hold, but everyone can buy what they deem to be a good buy. In other words, limits to short sell play an important role in the author's proposed mechanism. The authors then show that the high-volume premium would exist even when there is little price movement.

Empirically, we have much evidence that even sophisticated investors fail to allocate enough attention to stocks. Notably, trading volume has been used extensive as a measure of investor attention. The intuition is straightforward: when investors pay more attention to a stock, they will trade more intensely and timelier when new information arrives, hence higher trading volume should indicate higher investor attention. Another reason for its popularity is the readily availability of trading volume data. For example, Hou et al (2009) uses trading volumes to measure investor attention and show that while stocks with higher investor attention would see their stock prices adjusting quicker to the new information, they also tend to over-react. Overall, the authors find that

price momentum profits are higher among high volume stocks and in up markets, while earnings momentum profits are higher among low volume stocks and in down markets. Loh (2010) finds that low-attention stocks react less to stock recommendations than high-attention stocks, and they subsequently take longer to reach equilibrium prices. Boehmer and Wu (2013) shows that with more active short sellers, information is incorporated into asset prices timelier and more efficiently, facilitating price discovery. While studies have different opinions on the informativeness of short sellers, most agree that short sellers are not noise traders, and they trade on information. See, for example, Diamond and Verrecchia (1987), which shows that short sellers move asset prices to fundamentals by trading on their private information; Diether et al (2009), which shows that short sellers trade on short-term overreactions of stock prices and earns significant returns by correcting the overreactions; Christophe et al (2004), which shows that, using data for the five days before earnings announcement, short selling had already begun, suggesting that investors who had private information had already begun to exploit their private information. The authors also show that short sellers are more active in stocks with low book-to-market ratios. While the exact data frequencies differ, they tend to be low frequency, allowing time for investors to react on their information. For example, Asquith et al (2005) uses monthly data, Diether et al (2009) uses five-day returns, and Boehmer et al (2008) forms portfolios by holding portfolios for 20 trading days. The authors do not explicitly justify their choice of data frequency and portfolio window; generally, they follow what's the popular choice in literature and allow for enough time for investors, and hence asset prices, to react.

This strand of literature recognizes that stocks experiencing positive volume shocks would receive positive returns in the near future (see, for example, Gervais et al, 2001; Kaniel et al, 2012). Exactly how high-volume premium attracts higher stock return has also attracted substantial research attention. Explanations include: 1) Investor recognition hypothesis of Merton (1987). Positive trading shocks increase investor visibility and subsequently, investors purchase and drive up the stock because of its higher visibility and lower cost of capital. See, for example, Lerman et al, 2010; Kaniel et al, 2012. 2) Mispricing explanation. For example, Statman et al (2006) states that investor overconfidence would lead them to purchase more the stock, hence higher future return. 3) Risk-related explanations. In a recent study, Wang (2021) examines if macroeconomic variables help explain the volume premium. The author finds that increase in high-volume premium predicts a decrease in industrial production next-month, and that in the cross-section, industrial production

helps explain high-volume premium. However, the author does acknowledge that incorporating liquidity and industry production into Fama-French five-factor model explains only 1/3 of the total high-volume premium.

Israeli et al (2022) is another recent study on this issue. In this study, the authors examine whether firms increase their investment activity as a potential real consequence of decline in cost of capital. In this sense, it serves as a test of investor recognition hypothesis. As reductions in cost of capital is typically associated with increases in investment projects' net present values, the authors (as they claim) are the first to explore this link between trading volume and a company's subsequent investment activity. The authors confirm that a one standard deviation increase in unexpected trading volume increases annual investment expenditures by 1.4%. As it takes times for the effect of reductions in cost of capital to set in, the authors consider two to four quarters after we observe an increase in the stock's abnormal trading volume. To establish the link, the authors first test and confirm that there is a positive relationship between abnormal trading volume and future financing cash flows, suggesting that the company is raising additional funds following positive trading volume shocks. The authors then check and find that following trading volume shocks and that this effect is more significant for financially constrained companies, supporting their claim that volume shocks lead to reductions in cost of capital.

4.4.4 Investor recognition hypothesis

The investor recognition hypothesis was first advanced by Merton (1987). It states that if information is not perfect among investors, investors would avoid stocks they are not familiar with. The Number of investors who know about a security is termed the degree of investor recognition. In a pioneer and very influential study, Fang and Peress (2009) finds that stocks with *no media coverage* earn higher returns than stocks with media coverage. Specifically, risk-adjusted portfolio returns formed using stocks with no media coverage earn 'no media premium' of 8%-12% per annum. By reaching out to a larger audience, hence helping disseminating information to a wider public, Fang and Peress (2009) argues that stocks with higher media coverage earns lower returns. Other studies include, for example, Lehavy and Sloan (2008), Bodnaruk and Ostberg (2005), Chen, Hong, and Stein (2002), et cetera.

Literature generally supports the investor recognition theory. If this is the case, then as the number of audiences grows (or more precisely, higher number of types of investors), the news is

communicated to more investors, and they exploit the profitable opportunities, hence profitability should be lower. To test, we check if mean number of audiences has a negative effect on portfolio return. The investor recognition effect of Merton (1987) is sometimes called investor-base broadening (see, for example, Wang, 2007). We note an important feature of this hypothesis: even if investors already have all the information, news stories may still cause stock price reactions when the form of the news captures the attention of more investors.

We note, however, not all studies favor investor recognition hypothesis. For example, Chen, Hong, and Stein (2002) find a positive relationship between the change in the number of institutional holders and future stock returns. Using a similar method, Bodnaruk and Ostberg (2005) finds evidence favoring investor recognition hypothesis. Lehavy and Sloan (2008) shows that such contradicting results can be reconciled, and that after controlling for the positive autocorrelation in investor recognition, investor recognition should reduce future stock return, and the positive relationship appears to be from not controlling for the autocorrelation in investor recognition measure.

The measure of investor recognition varies in literature. Arbel, Carvell, and Strebel (1983) uses the number of institutional investor holdings of stocks as a measure of investor recognition and finds a negative relationship between investor recognition and future stock return. Fand and Peress (2009) uses the number of newspaper articles on LexisNexis database about a company to proxy for the stocks' media exposure. Lehavy and Sloan (2008) uses changes in investor holdings as a proxy for investor recognition. None of the measures are similar to *naudience* we use in this study. In this sense, we also contribute to existing literature with a new media coverage measure. This is because the types of subscribers (or type of investors, but not to be confused with 'type' as referring to institutional/ retail investors) is unavailable in existing measures.

4.4.5 News and volatility jumps

Traditionally, there has been a huge literature on how news affect price levels and returns. Recently, literature has begun to study macroeconomic news effect on price jumps. See, for example, Rangel (2011), Lahaye et al (2011), and Huang (2018). More recently, literature began to explore how news affect financial markets through volatility jumps. This is a relatively new literature and evolves from traditional financial research in volatility. Notably, this strand of literature (especially the early ones) develops that price and volatility jumps tend to co-occur and are often event-driven. See, for

example, Bandi and Reno (2016), Broadie et al (2007), and Eraker et al (2003). This suggests that jumps are a critical part of financial markets following the large number of events that hit the market. However, jumps in price and volatility tends to be neglected, especially in traditional models. In traditional asset pricing models, return is continuous, and investors only need to consider small, continuous changes in prices when making investment decisions. However, event-driven volatility jumps are more and more common in the high-frequency world, hence it is critical to also consider the jump components in price and volatility. As highlighted by Liu et al (2003), major events often trigger jumps in stock prices and volatility, and identifying such jumps is critical for portfolio allocation decisions. For other theoretical models involving volatility jumps, see, for example, Fulop et al (2015).

So far, studies in volatility jumps and news events tend to be macroeconomic focused. One reason may be the easy availability of market-wide data (such as stock index) and macroeconomic data (such as announcements of macroeconomic variables). Huang (2018) and Chan and Gray (2018) are two recent examples. In Chan and Gray (2018), the authors focus on the announcements of four macroeconomic variables: the Federal Open Market Committee's (FOMCs) announcement of the target Federal Funds rate, non-farm payroll (NFP), the unemployment rate, and the producer price index (PPI). The authors' choice follows from previous studies which show that such macroeconomic variable announcements lead to price jumps (Rangel, 2011; Huang, 2018; Chan and Gray, 2017). They find that (probably unsurprisingly) macroeconomic news announcements do influence volatility jumps of government bonds and the market (as proxied by the S&P 500 index). The larger the information content, the higher the jump. Bad news would lead to large volatility jumps than positive. Traditionally, it is a tedious job to separate volatility jumps from price jumps. The authors follow Andersen et al (2012) to use intraday high-frequency futures data to construct an estimate of the daily integrated variance, which is robust to price jumps, if any. The authors then use stochastic processes to identify volatility jumps.

In this paper, we use RSJ, a recent development in econometrics, to conveniently (compared with existing literature) measure volatility jumps for each stock. We also differ drastically from existing literature because we focus on firm-specific news. Different from literature, which often relies on news announcements of macroeconomic variables to identify relevant events, and hence have limited dataset, we use Refinitiv MRN which contain a much larger amount of firm-specific news.

4.4.6 Fama-French model and beyond

We present Fama-French models and return distributions in Table 4-9 and Figure 4-6. Common portfolio performance measures are presented in Table 4-9. A look at the return histogram shows that the portfolio gives higher positive returns than negative on average, with a longer right tail than left tail.

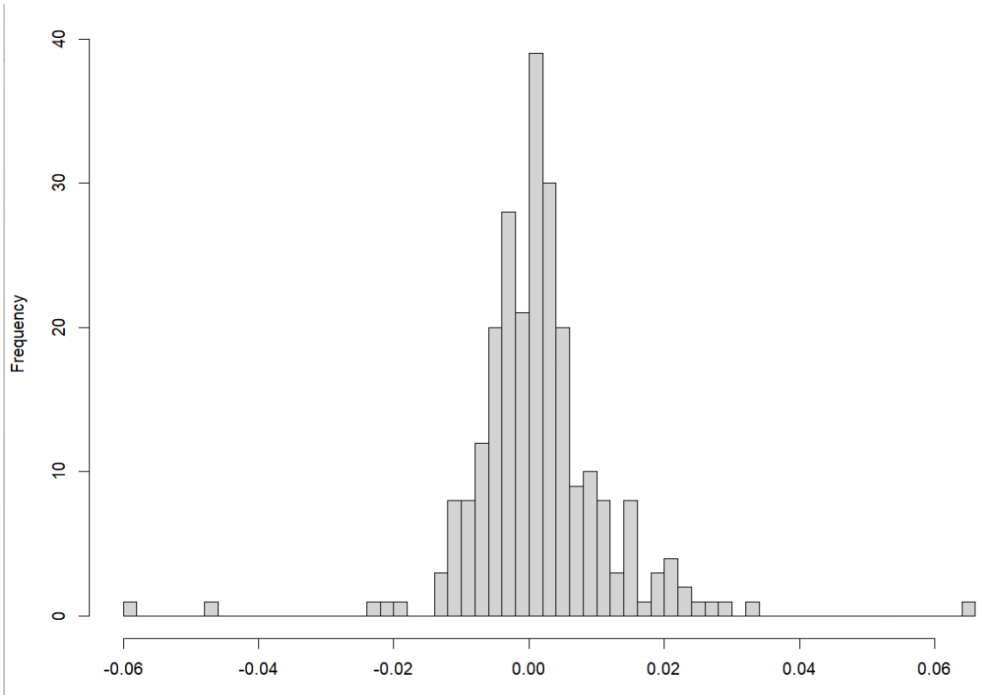


Figure 4-6 Distribution of daily portfolio return (with news articles)

Fama-French 5 with Number of Audiences and Volume Shocks

	Portfolio	**Portfolio**	**Short Leg**	**Long Leg**
(Intercept)	0.0013.	0.0066*	0.0033	0.0005
	(0.0007)	(0.0027)	(0.0040)	(0.0019)
RM	0.0014	0.0019	0.0057	0.0018
	(0.0030)	(0.0030)	(0.0060)	(0.0032)
SMB	0.0008	0.0010	0.0010	-0.0006
	(0.0009)	(0.0009)	(0.0018)	(0.0010)
HML	0.0009	0.0012	0.0022	0.0019
	(0.0015)	(0.0015)	(0.0029)	(0.0016)
RMW	-0.0019	-0.0021	-0.0043	0.0019
	(0.0015)	(0.0015)	(0.0029)	(0.0016)

Fama-French 5 with Number of Audiences and Volume Shocks

	Portfolio	**Portfolio**	**Short Leg**	**Long Leg**
CMA	-0.0018	-0.0016	-0.0033	-0.0003
	(0.0022)	(0.0022)	(0.0044)	(0.0024)
n-audience		-0.0006.	0.0000	-0.0000
		(0.0003)	(0.0004)	(0.0002)
Volume shocks		0.0031.	0.0043	-0.0006
		(0.0017)	(0.0030)	(0.0015)
R^2	0.0183	0.0448	0.0235	0.0318
Adj. R^2	-0.0020	0.0168	-0.0070	0.0018
Num. obs.	247	247	232	234

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4-9 Fama-French 5-factor model on article portfolio

Mean r	sd	Sharpe Ratio	Max Drawdown	Sortino Ratio
0.0014	0.0105	2.06	0.0897	0.1702711

Table 4-10 Common portfolio measures for XGBoost article model

Having established related literature in finance, we establish their links to this study and how they can be used to explain the observed high performance of our portfolio. Following this strand of literature, the database in our study, Refinitiv MRN, provides a way of testing for investor attention of our portfolio. As investors pay a large sum[18] to subscribe to this professional news platform, it is likely investors would subsequently read the news. However, we still have the problem that we have no guarantee investors will read the news timely, if at all. We have an indirect measure of investor attention and media coverage using the number of audiences in this sense. Assuming ceteris paribus, the higher the number of audiences, the more investors the news is communicated to, and the higher investor attention. This is similar to the indirect measures of investor attention, and in this sense, we also contribute to the literature another way of measuring investor attention. As investors only pay a premium subscription fee when they are after a certain industry, asset, et cetera, we may also argue

[18] The exact amount each investor is paying, however, is not available. Similarly, we do not have data on the charge of each Refinitiv News product.

that this is indeed a direct measure of investor attention, because they *choose* to subscribe to the news topic. We then use each news' average number of topics to measure to what extent the news appears in media and/or captures investor attention.

Trading volume is easily obtainable. To test if trading volume, an indirect measure of investor attention, explains the portfolio return, we use unexpected trading volume as a regressor to check if it captures and explains portfolio returns.

To set the benchmark, we first regress daily portfolio returns on the Fama-French 5 factor model. We find that our portfolio achieves significant daily alpha in excess of Fama-French 5 factor model where the daily alpha is 0.0013. None of the Fama-French five factors are significant. We note that adjusted-R^2 is negative, indicating that the Fama-French five factor model is very poor at explaining the portfolio returns. In adding additional explanatory variables to the Fama-French 5-factor model, we also seek to test where the profitability comes from, and if the average effect of investor attention is indeed through investor purchasing the stocks significantly outweighing investors selling and short selling. Therefore, in addition to daily portfolio returns, we also consider the positive and negative legs of the portfolio separately in running our explanatory regressions. We do not seek to explain high-volume premium. Instead, we check if high-volume premium helps explain the observed portfolio return. To do so, we first fit ARIMA model to the 44 stocks' daily volume. The choices of AR lag, MA lag, and order of integration are selected by AIC where MA and AR lags are between 0-60, and order of integration of 0 and 1. We then predict each stock's daily volume using the ARIMA model fitted for each stock. Finally, we obtain the residuals for each stock's daily volume series and the residual of the stocks' daily trading volume represents daily volume shock. To standardize, we convert the daily volume shock as a percentage of each day's trading volume.

If the media coverage and investor attention effects are true, then as the news is known to more investors and more types of investors (as measured by *naudience*), we should expect a lower return on average. This is because investors are paying attention to the stocks covered in these news and profitable opportunities would quickly disappear. Our next-day portfolio hence would see lower return. We test the hypothesis by adding the mean number of audiences of news stories that are used in forming portfolio. We note that the number of audiences is statistically significant at 1% level with a negative effect. Specifically, as the news is known to one more type of audience (and hence more investors on average), we expect portfolio's daily return to decrease by 0.0006, on average.

When adding the mean number of audiences and volume shocks, the daily portfolio return's alpha would increase to 0.007 and all pricing factors in Fama-French five factor model remains insignificant.

In the daily portfolio return, we add mean volume shocks of the stocks used in the long and short legs of portfolio in addition to mean number of audiences. Volume shock is significantly positive at 0.0031. When breaking down portfolio by long and short legs, however, neither volume shocks nor number of audiences are significant, and the overall *alpha* of the long and short legs stop being significant.

Overall, number of audiences and volume shocks, which represent proxies for media coverage and high-volume premium, both have some explanatory power to the portfolio's daily return. However, neither maintains significance in both the daily return and each leg of the portfolio. Future research may attempt to further explain the observed portfolio return with other risk and behavioral explanations.

4.5 Conclusion

In this paper, we use XGBoost to predict next-day RSJ and use the predicted RSJ to form one-day portfolio for the next day. We document that controlling for news sentiment and volatility measures, next-day RSJ is reasonably predictable where RMSE of predicted RSJ (bounded between -1 and 1) is around 0.34. Using the predicted RSJ to form one day portfolio achieves outstanding portfolio performance where Sharpe ratio is 2.06 with only 44 stocks available, effectively suggesting a risky portfolio. The portfolio performance is significantly higher than portfolio selection based on both today's RSJ and today's news sentiment. Investor recognition and attention potentially explain the results.

We also note that daily portfolios constructed using news articles achieve significantly higher Sharpe ratios than using news alerts, suggesting that news articles are more informative. Future researchers may wish to use the body contents of news instead of just headlines whenever possible.

Our results suggest that the specific topics talked about in the news have strong implications to tomorrow's volatility jump, and using this predicted volatility jump to form portfolio is highly profitable. This is probably because there have been limited attempts to explore the topic information in news; for those who do attempt to explore how topic information may assist in

investment decisions, their approaches may not be adequate. For example, existing studies are often forced to use LDA to classify the topic information of texts, and typically there are only a handful of topics possible. This is far from the reality, where we have thousands of topic codes in our dataset. Instead of estimating the topics, we have detailed pre-labels which give us accurate representations of news topics. To our knowledge, we are the first to explore how such pre-labelled, detailed topics may help in predicting next-day volatility jumps, and hence portfolio construction with financial news. Future research may explore how such information may be used in other fields of financial economics.

Not all topics are relevant, though. Of the 1,400 topics, only 173 help form the predictive model. We note that it would be practically impossible to manually investigate all the topic codes to select the relevant ones, and letting the algorithm automatically select features is a more practical (and proven useful) strategy. We also acknowledge that because of dataset limits, we keep many features that occur under 10 times in the training set to at least retain the inter-connections of different topics that XGBoost can pick up. With more comprehensive dataset, we can potentially improve our model and portfolio performance.

One possible direction for future research is to extend dataset. With sufficiently large dataset, one should achieve better portfolio performance because one can effectively hedge idiosyncratic risks. To do so, we can either combine Refinitiv MRN with other databases, such as Dow Jones Newswire, or label their news sources by different topics. We note, however, trying to give news labels would not be straightforward, because one piece of news typically covers different topics.

4.6Appendix

4.6.1 Appendix A. XGBoost and its estimation

This section provides a technical note on XGBoost and its estimation methods for technical readers. Readers should be comfortable with tree ensemble models and its basics, which is one of the most basic and popular architecture in machine learning.

XGBoost is invented by Chen and Guestrin (2016). Being a tree-ensemble model, it is designed for structured tabular data. Assuming we have a total of K trees, the estimated value of the dependent variable is simply the sum of outputs from each tree, taking the dependent variables as inputs.

$$\hat{y}_i = \sum_{k}^{K} f_k(x_i)$$

Like all tree ensembles, XGBoost seeks to minimize an objective function. XGBoost is designed with regularization in mind such that the objective function already includes a penalization term for model complexity.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

The first term $l(y_i, \hat{y}_i)$ approximates training loss and effectively measures how good (or bad) the model fits to the training set. The second term $\Omega(f_k)$ is a measure of model complexity and prevents overfitting. As a model gets more complex, it will overfit to training data by 'fitting' the purely noisy and random part, lowering its accuracy out-of-sample. The L2 regularization term penalizes complex models that's prone to overfitting.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

T is the number of leaves and $\sum_{j=1}^{T} w_j^2$ is *L2-norm* of leaf scores. We start with:

$$\hat{y}_i^{\,0} = 0$$

And then:

$$\hat{y}_i^1 = f_1(x_i) = \hat{y}_i^{\,0} + f_1(x_i)$$

$$\hat{y}_i^2 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)$$

...

$$\hat{y}^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i)$$

For regression problems where the dependent variable is continuous, loss function is often approximated by squared error. If we consider the t^{th} tree, we then have:

$$\begin{aligned} Obj^t &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i) + \Omega(f_t) + constant \\ &= \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{t-1} + f_t(x_i)\right)\right)^2 + \Omega(f_t) + constant \\ &= \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{t-1} - y_i\right) f_t(x_i) + f_t(x_i)^2\right]^2 + \Omega(f_t) + constant \end{aligned}$$

For numeric approximation, the authors apply second order Taylor series expansion on the objective function. To simplify things, the authors further define $g_i \; and \; h_i$. We still assume squared loss function.

$$g_i = \partial_{\hat{y}^{t-1}} (\hat{y}^{t-1} - y_i)^2$$
$$h_i = \partial^2_{\hat{y}^{t-1}} (y_i - \hat{y}^{t-1})^2 = 2$$

And similarly:

$$G_j = \sum_{i \in I_j} g_i$$
$$H_j = \sum_{i \in I_j} h_i$$

For optimization purpose, Chen and Guestrin (2016) introduces a tweak from here where the objective function is changed from iteration through the trees to iteration through the leaves. Now we can write the objective function in terms of the leaves.

$$Obj = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w^2 \right] + \gamma T$$

The optimal weight and objective function can now be simply derived.

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

$$W_j^* = -\frac{G_j}{H_j + \lambda}$$

4.6.2 Appendix B. Notes on data cleaning

We obtain each company's intraday, open, and close prices from Refinitiv (Thomson Reuters) Tick History (TRTH) database. When a piece of news relates to specific companies, the news would be tagged with the companies' Reuters Instrument Code (RIC), and we use RICs to match and identify companies. Companies' RICs are in the format of "Ticker.Exchange", where the suffix identifies the exchange. For example, "AAPL.O" stands for Apple traded on NASDAQ. The dataset we obtained contains news from assets and companies around the world. To identify and filter for US stocks (i.e., find the 'Exchange' part of the RICs), we use the latest US exchange suffix provided by Refinitiv. Unfortunately, Refinitiv does not maintain a historical list of US exchange suffix, and we note that we would lose a small number of exchanges that existed some time during the sample and disappeared at some point in time. There are a total of 4,807,623 observations from 14,214 companies, where 4,389 firms still exist today, and 7,864 companies are historical and were delisted at some point in time. 1,961 companies are not found in TRTH, but they only account for 17,518 pieces (or 0.36%) of news. Such news may include failed IPOs, data error, Refinitiv service alerts and maintenance, et cetera. After excluding companies not identified in TRTH, we have a total sample of 4,790,104 pieces of news from 12,253 companies over a 24-year period.

4.6.3 Appendix C XGBoost results covering all features

For article models, the most important features are listed below.

	Topic	Class	Gain	Cover	Frequency	n_train	%training	n_test	%test	n_vali	%vali
1	Workforce	Event Type	0.0041	0.0039	0.0043	1525	0.0694	215	0.0646	363	0.0499
2	Products / Services	Event Type	0.0036	0.0039	0.0026	691	0.0315	87	0.0262	252	0.0346
3	Miscellaneous Specialty Retailers (TRBC level 4)	Business Sector	0.0034	0.0032	0.0017	464	0.0211	79	0.0238	1	0.0001
4	Key Personnel Changes	Event Type	0.0032	0.0031	0.0036	1334	0.0607	198	0.0595	316	0.0434
5	Organizational Restructuring	Event Type	0.0032	0.0036	0.0024	573	0.0261	99	0.0298	100	0.0137
6	Debt / Fixed Income Markets	Asset Class / Property	0.0031	0.0027	0.0038	2395	0.1090	323	0.0971	662	0.0910
7	Wealth Management (TRBC level 5)	Business Sector	0.0031	0.0026	0.0033	2265	0.1031	329	0.0989	394	0.0542
8	Management Issues / Policies	Event Type	0.0030	0.0030	0.0055	3933	0.1790	569	0.1711	786	0.1080
9	Crime / Law / Justice	Broad News Topic	0.0030	0.0022	0.0029	2658	0.1210	402	0.1209	669	0.0919
10	Broker Research / Recommendations	Event Type	0.0029	0.0030	0.0019	522	0.0238	58	0.0174	211	0.0290
11	Economic Indicators	Event Type	0.0029	0.0027	0.0025	796	0.0362	147	0.0442	275	0.0378
12	Funds	Asset Class / Property	0.0029	0.0022	0.0041	3837	0.1747	566	0.1702	757	0.1040
13	South America / Central America	Geography	0.0028	0.0026	0.0028	980	0.0446	139	0.0418	255	0.0350
14	Europe	Geography	0.0028	0.0023	0.0048	4008	0.1825	599	0.1801	1377	0.1893

15	Securities & Commodity Exchanges (TRBC level 5)	Business Sector	0.0028	0.0028	0.0018	431	0.0196	48	0.0144	48	0.0066
16	Insurance (TRBC level 3)	Business Sector	0.0028	0.0028	0.0018	519	0.0236	81	0.0244	101	0.0139
17	Technology / Media / Telecoms	Business Sector	0.0027	0.0023	0.0032	2050	0.0933	356	0.1070	1339	0.1840
18	Technology Equipment (TRBC level 2)	Business Sector	0.0027	0.0028	0.0018	502	0.0229	77	0.0232	139	0.0191
19	Equities Markets	Asset Class / Property	0.0027	0.0023	0.0038	2357	0.1073	427	0.1284	1473	0.2024
20	Performance / Results / Earnings	Event Type	0.0027	0.0022	0.0051	4823	0.2196	990	0.2977	1597	0.2195
21	Deals	Event Type	0.0027	0.0024	0.0042	2172	0.0989	253	0.0761	313	0.0430
22	Canada	Geography	0.0027	0.0028	0.0022	630	0.0287	84	0.0253	190	0.0261
23	Corporate / Market Regulation	Event Type	0.0027	0.0024	0.0053	3524	0.1604	486	0.1461	695	0.0955
24	Technology (TRBC level 1)	Business Sector	0.0026	0.0021	0.0026	2251	0.1025	381	0.1146	969	0.1332
25	Healthcare (TRBC level 1)	Business Sector	0.0026	0.0022	0.0030	2941	0.1339	415	0.1248	933	0.1282
26	Mergers / Acquisitions / Takeovers	Event Type	0.0026	0.0025	0.0031	1301	0.0592	153	0.0460	135	0.0186
27	Telecommunications Services (TRBC level 3)	Business Sector	0.0025	0.0025	0.0021	729	0.0332	99	0.0298	210	0.0289
28	Company News	Broad News Topic	0.0025	0.0026	0.0018	21414	0.9748	3260	0.9802	6969	0.9578
29	Investment Banking & Investment Services (TRBC level 3)	Business Sector	0.0025	0.0021	0.0021	1089	0.0496	186	0.0559	1183	0.1626

Chapter 5 Conclusions

Today, the use of AI and computer science technologies in inter-disciplinary research is popular. Among the efforts, a notably large (and fast growing) literate is text-as-data following the huge success of Transformer. This thesis contributes to this literature by applying the model to Refinitiv Machine Readable News database, where typical computer science researchers do not have access to due to its cost. While computer science literature also seeks to apply the latest algorithms to application in finance and economics problems, a large drawback is that they typically do not seek to answer the 'why' and 'how', which are of utmost importance to financial economists. By bridging financial economics and computer science literature, I hope this thesis would be of help to current and future research in business schools and computer science.

Specifically, in this thesis, we used state-of-the-art econometrics and machine learning techniques to reconsider price discovery, sentiment analysis, and volatility predictions. In doing so, we show how finance practitioners and academics may benefit from new techniques and new databases. We try to link the way traditional financial economics tackles the same topics and show new insights using the new technologies and a new database. Despite the effort, we know this only serves as a small step towards a better understanding of the financial market, and there are obviously more to be done.

5.1 Common knowledge in financial markets

Our understanding and interpretation of price discovery has not changed for decades. That is, price discovery comes from somewhere, and must somewhere. However, one does not need expert financial economics knowledge to know that a huge part of information is known to all. This is probably because 'price discovery' and its measure in market microstructure literature comes (almost unambiguously, for market microstructure researchers) the classical Hasbrouck framework (1990; 1994). There has been little effort in challenging this interpretation.

We show that analytically, the common part of several price series can be estimated, and that empirically, the vast majority of information is common to all markets. This is intuitive: most of the time, market participants react to the same set of information that is made available to everyone. It is unreasonable to expect a market to be more '*informative'* just because users of the specific venue

react faster. This is probably a better interpretation of *price discovery* and trading venue's *informativeness*: the part of information that is not known to all determines how *informative* the trading venue is.

Our analysis, of course, is far from perfect. While we show that empirically, common information matters, and this can be estimated in an easy way (at least for two-market and three-market cases), we are yet to use the model to its full potential. Future research may consider why common information matters and how the model can be used for economic implications in their specific field of research.

5.2 Revisiting sentiment analysis

The idea that investor sentiment explains and predicts market movements, and that texts provides a valuable source for sentiment extraction, is not new. It may surprise us how early the first attempts were made when Cowles (1993) tried to do one of the earliest studies on financial sentiment analysis. For almost a century, however, sentiment analysis with text data fails to find its way to financial economists' toolbox. With the hugely successful ChatGPT, text-as-data draws attention of almost all financial economists, and many research topics are revisited.

With new technology, sentiment analysis should imply higher portfolio performance. This is what we found, and this is no surprise. Thanks to Refinitiv Machine Readable News database, we are able to gain more insights with sentiment analysis. Our analysis shows that being more information-rich, investors should probably prefer news articles over news alerts. This helps improve their portfolio performance. When news is more complex, it is more difficult to infer the true informational contents, and investors would find it hard to explore the profit opportunities fully and timely.

We note, however, it is probably inaccurate to call this method 'sentiment', as one cannot effectively distinguish between sentiment and informational contents in the news. This is, however, the common term used by researchers and practitioners. Future research may attempt to find a way to distinguish between sentiment and information in texts.

5.3 News and volatility jumps

News, unless fully expected, will cause price and volatility jumps as market participants adjust their expectations of asset prices. Being directly observable, price changes are easier to deal with and investigate. Volatility, however, can only be estimated, and usually involves complicated time series

and stochastic models. With recent, easy-to-use developments in volatility jump measures, we use realised signed jump variation to demonstrate its predictability. We further show that portfolio construction with predicted volatility jumps is highly profitable. The high predictability and high profit are probably due to limited attempts to explore profitability opportunities via topic prediction. Why have there been limited attempts? While we cannot say for sure, limits in data availability appear to be the driving force. Detailed topic information is only available in expensive proprietary databases, but research by computer scientists typically relies on small (and typically free) databases to demonstrate the algorithms work better than previous models instead of directly tackling financial economics problems. We further bridge financial economics and computer science research by showing how the profitability may be linked to and explained by factors advanced in financial economics literature.

An obvious (and natural) extension to this work is to explore higher moments with news, such as kurtosis and skewness. However, it is less obvious how textual news causes higher moment dynamics of stock returns, and establishing such causal channels may be challenging. Broadly, applying the methodologies to other markets other than the US stock market – if one has reasons to believe other markets would yield different results.

References

Ait-Sahalia, Y. (2004). Disentangling Diffusion from Jumps. Journal of Financial Economics.

Aït-Sahalia, Y., Cacho-Diaz, J., & Hurd, T. R. (2009). Portfolio choice with jumps: A closed-form solution.

Al-Thaqeb, S. A., & Algharabali, B. G. (2019). Economic policy uncertainty: A literature review. The Journal of Economic Asymmetries, 20, e00133.

Anand, A., & Subrahmanyam, A. (2008). Information and the intermediary: Are market intermediaries informed traders in electronic markets? Journal of Financial and Quantitative Analysis, 1–28.

Andersen, T., Benzoni, L., and Lund, J., 2002. An Empirical Investigation of Continuous-Time Equity Return Models. Journal of Finance, Vol. 57, Issue 3, pp. 1239-1284.

Andersen, T., Bollerslev, T., and Diebold, F., 2007. Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. Review of Economics and Statistics, Vol. 89, Issue 4, pp. 701-720.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Volatility and correlation forecasting. Handbook of Economic Forecasting, 1, 777–878.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2001). The distribution of realized exchange rate volatility. Journal of the American Statistical Association, 96(453), 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Vega, C. (2003). Micro effects of macro announcements: Real-time price discovery in foreign exchange. American Economic Review, 93(1), 38–62.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259–1294.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. International Journal of Forecasting, 36(2), 334–357.

Baillie, R. T., Booth, G. G., Tse, Y., & Zabotina, T. (2002). Price discovery and common factor models. Journal of Financial Markets, 5(3), 309–321.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. The Journal of Finance, 61(4), 1645–1680.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129–152.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131(4), 1593–1636.

Baker, S. R., Bloom, N., Davis, S. J., & Kost, K. J. (2019). Policy news and stock market volatility (Issue w25720)). National Bureau of Economic Research.

Bandi, F. M., & Reno, R. (2016). Price and volatility co-jumps. Journal of Financial Economics, 119(1), 107–146.

Barberis, N., Shleifer, A., & Vishny, R. W. (2005). A model of investor sentiment. Princeton University Press.

Barclay, M. J., & Litzenberger, R. H. (1988). Announcement effects of new equity issues and the use of intraday price data. Journal of Financial Economics, 21(1), 71–99.

Barndorff-Nielsen, O. E., & Shephard, N. (2005). Variation, jumps, market frictions and high frequency data in financial econometrics.

Barndorff-Nielsen, O. and Shephard, N., (2004). Power and Bipower Variation with Stochastic

Volatility and Jumps. Journal of Financial Econometrics, Vol. 2, pp. 1-37.

Bessembinder, H. (2003). Quote-based competition and trade execution costs in NYSE-listed stocks. Journal of Financial Economics, 70(3), 385–422.

Beveridge, S., & Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle.' Journal of Monetary Economics, 7(2), 151–174.

Biais, B., Foucault, T., & Moinas, S. (2015). Equilibrium fast trading. Journal of Financial economics, 116(2), 292–313.

Black, F. (1976). Studies of Stock Price Volatility Changes. Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, 177–181.

Black, F. (1986). Noise. The Journal of Finance, 41(3), 528–543.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1–8.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). ARCH models. Handbook of Econometrics, 4, 2959–3038.

Bollerslev, T., Law, T. H., & Tauchen, G. (2008). Risk, jumps, and diversification. Journal of Econometrics, 144(1), 234–256.

Bollerslev, T., Li, S. Z., & Zhao, B. (2020). Good volatility, bad volatility, and the cross section of stock returns. Journal of Financial and Quantitative Analysis, 55(3), 751–781.

Bollerslev, T., & Todorov, V. (2011). Tails, fears, and risk premia. The Journal of Finance, 66(6), 2165–2211.

Bonaime, A., Gulen, H., & Ion, M. (2018). Does policy uncertainty affect mergers and acquisitions? Journal of Financial Economics, 129(3), 531–558.

Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? The Journal of Finance, 40(3), 793–805.

Booth, G. G., Lin, J. C., Martikainen, T., & Tse, Y. (2002). Trading and pricing in upstairs and downstairs stock markets. The Review of Financial Studies, 15(4), 1111–1135.

Booth, G. G., So, R. W., & Tse, Y. (1999). Price discovery in the German equity index derivatives markets. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 19(6), 619–643.

Bouchaud, J. P., Kockelkoren, J., & Potters, M. (2006). Random walks, liquidity molasses and critical response in financial markets. Quantitative finance, 6(02), 115-123.

Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. The Review of Financial Studies, 32(3), 992–1033.

Brogaard, J., Hendershott, T., & Riordan, R. (2014a). High-frequency trading and price discovery.

The Review of Financial Studies, 27(8), 2267–2306.

Brogaard, J., Hendershott, T., & Riordan, R. (2014b). High-frequency trading and price discovery. The Review of Financial Studies, 27(8), 2267–2306.

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. Journal of Empirical Finance, 11(1), 1–27.

Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation. The Journal of Business, 78(2), 405–440.

Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. The Quarterly Journal of Economics, 130(4), 1547–1621.

Campbell, J. Y., & Hentschel, L. (1992). No news is good news: An asymmetric model of changing volatility in stock returns. Journal of Financial Economics, 31(3), 281–318.

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., & Whitelaw, R. F. (1998). The econometrics of financial markets. Macroeconomic Dynamics, 2(4), 559–562.

Campbell, J. Y., & Mankiw, N. G. (1987). Permanent and transitory components in macroeconomic fluctuations (Issue w2169)). National Bureau of Economic Research.

Can stock market forecasters forecast? (1933). Econometrica: Journal of the Econometric Society, 309–324.

Cao, C., Hansch, O., & Wang, X. (2009). The information content of an open limit-order book. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 29(1), 16–41.

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61, 304–323.

Charoenrook, A. (2005). Does sentiment matter. Unpublished working paper. Vanderbilt University.

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2013). Customers as advisors: The role of social media in financial markets.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Chen, X., & Ghysels, E. (2011). News—Good or bad—And its impact on volatility predictions over

multiple horizons. The Review of Financial Studies, 24(1), 46–81.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cheung, Y. W., & Ng, L. K. (1992). Stock price dynamics and firm size: An empirical investigation. The Journal of Finance, 47(5), 1985–1997.

Choi, J. J., Laibson, D., & Metrick, A. (2002). How does the Internet affect trading? Evidence from investor behavior in 401 (k) plans. Journal of Financial Economics, 64(3), 397–421.

Chordia, T., Roll, R., & Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. Journal of Financial Economics, 101(2), 243–263.

Chou, R. Y. (1988). Volatility persistence and stock valuations: Some empirical evidence using GARCH. Journal of Applied Econometrics, 279–294.

Chu, Q. C., Hsieh, W. L. G., & Tse, Y. (1999). Price discovery on the S&P 500 index markets: An analysis of spot index, index futures, and SPDRs. International Review of Financial Analysis, 8(1), 21–34.

Clark-Joseph, A. D., Ye, M., & Zi, C. (2017). Designated market makers still matter: Evidence from two natural experiments. Journal of Financial Economics, 126(3), 652–667.

Çolak, G., Durnev, A., & Qian, Y. (2017). Political uncertainty and IPO activity: Evidence from US gubernatorial elections. Journal of Financial and Quantitative Analysis, 52(6), 2523–2564.

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

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1989). What Moves Stock Prices? Journal of Portfolio Management, 15, 4–12.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461–1499.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. The Journal of Finance, 53(6), 1839–1885.

Das, S. R., & Chen, M. Y. (2007). Yahoo! For Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), 1375–1388.

Deb, F. H., McInish, T. H., Shoesmith, G. L., & Wood, R. A. (1995). Cointegration, error correction, and price discovery on informationally linked security markets. Journal of Financial and Quantitative Analysis, 563–579.

Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. (2008). In Econometrica (Vol. 76, Issue 6, pp. 1481–1726).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dolan, R. J. (2002). Emotion, cognition, and behavior. Science, 298(5596), 1191–1194.

Du, S., & Zhu, H. (2017). What is the optimal trading frequency in financial markets? The Review of Economic Studies, 84(4), 1606–1651.

Dungey, M., Henry, O., & Mckenzie, M. (2013). Modeling trade duration in US Treasury markets. Quantitative Finance, 13(9), 1431–1442.

Easley, D., O'Hara, M., & Yang, L. (2016). Differential access to price information in financial markets. Journal of Financial and Quantitative Analysis, 51(4), 1071–1110.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3), 1184–1216.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. The Journal of Finance, 48(5), 1749–1778.

Fama, E. F. (1965). The behavior of stock-market prices. The Journal of Business, 38(1), 34–105.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The journal of Finance, 25(2), 383-417.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. Journal of financial economics, 49(3), 283–306.

information. International Economic Review, 10(1), 1–21.

Fan, J., Xue, L., & Zhou, Y. (2021). How much can machines learn finance from Chinese text data? Available at SSRN 3765862.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies, 15(4), 915–953.

Fleming, M. J., Mizrach, B., & Nguyen, G. (2018). The microstructure of a US Treasury ECN: The BrokerTec platform. Journal of Financial Markets, 40, 2–22.

Foucault, T., Hombert, J., & Roşu, I. (2016). News trading and speed. The Journal of Finance, 71(1), 335–382.

Foucault, T., Kozhan, R., & Tham, W. W. (2017). Toxic arbitrage. The Review of Financial Studies, 30(4), 1053–1094.

Frazzini, A., & Lamont, O. A. (2007). The earnings announcement premium and trading volume. NBER working paper, (w13090).

French, K. R., & Roll, R. (1986). Stock return variances: The arrival of information and the reaction of traders. Journal of Financial Economics, 17(1), 5–26.

Friedman, M., & Friedman, M. (1953). Essays in positive economics. University of Chicago press.

Garbade, K. D., & Silber, W. L. (1983). Price movements and price discovery in futures and cash markets. The Review of Economics and Statistics, 289–297.

Garcia, D. (2013). Sentiment during recessions. The Journal of Finance, 68(3), 1267–1300.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16), 6266–6282.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance, 48(5), 177.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with

heterogeneously informed traders. Journal of Financial Economics, 14(1), 71–100.

Gonzalo, J., & Granger, C. (1995). Estimation of common long-memory components in cointegrated systems. Journal of Business & Economic Statistics, 13(1), 27–35.

Greenwood, R., & Nagel, S. (2009). Inexperienced investors and bubbles. Journal of Financial Economics, 93(2), 239–258.

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance, 18(2), 321–340.

Gulen, H., & Ion, M. (2016). Policy uncertainty and corporate investment. The Review of Financial Studies, 29(3), 523-564 9-1801.

Hanousek, J., Kočenda, E., & Novotný, J. (2012). The identification of price jumps.

Hansen, P. R., & Lunde, A. (2006). Realized variance and market microstructure noise. Journal of Business & Economic Statistics, 24(2), 127–161.

Harris, F. H. D., McInish, T. H., & Wood, R. A. (2002). Security price adjustment across exchanges: An investigation of common factor components for Dow stocks. Journal of Financial Markets, 5(3), 277–308.

Harris, L. E., & Panchapagesan, V. (2005). The information content of the limit order book: Evidence from NYSE specialist trading decisions. Journal of Financial Markets, 8(1), 25–67.

Hasbrouck, J. (1988). Trades, quotes, inventories, and information. Journal of financial economics, 22(2), 229–252.

Hasbrouck, J. (1991a). Measuring the information content of stock trades. The Journal of Finance, 46(1), 179–207.

Hasbrouck, J. (1991b). Measuring the information content of stock trades. The Journal of Finance, 46(1), 179–207.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. The Journal of Finance, 50(4), 1175–1199.

Hasbrouck, J. (2002). Stalking the “efficient price” in market microstructure specifications: An

overview. Journal of Financial Markets, 5(3), 329–339.

Hasbrouck, J. (2021). Price discovery in high resolution. Journal of Financial Econometrics, 19(3), 395-430.

Hasbrouck, J., & Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. Journal of financial Economics, 59(3), 383–411.

Hasbrouck, J., & Sofianos, G. (1993). The trades of market makers: An empirical analysis of NYSE specialists. The Journal of Finance, 48(5), 1565–1593.

Heston, S. L., & Sinha, N. R. (2017). News vs. Sentiment: Predicting stock returns from news stories. Financial Analysts Journal, 73(3), 67–83.

Hillert, A., Jacobs, H., & Müller, S. (2014). Media makes momentum. The Review of Financial Studies, 27(12), 3467–3501.

Hoberg, G., & Phillips, G. M. (2018). Text-based industry momentum. Journal of Financial and Quantitative Analysis, 53(6), 2355–2388.

Hong, H., & Stein, J. C. (2003). Differences of opinion, short-sales constraints, and market crashes. The Review of Financial Studies, 16(2), 487–525.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168–177.

Huang, A. G., Tan, H., & Wermers, R. (2020). Institutional trading around corporate news: Evidence from textual analysis. The Review of Financial Studies, 33(10), 4627–4675.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), 806-841.

Huang, R. (2000). Price discovery by ECNs and Nasdaq market makers. Owen School of Management, Vanderbilt University.

Huberman, G., & Regev, T. (2001). Contagious speculation and a cure for cancer: A non-event that made stock prices soar. The Journal of Finance, 56(1), 387–396.

Hutto, C., & 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)).

Ingram, R. W., & Frazier, K. B. (1980). Environmental performance and corporate disclosure. Journal of Accounting Research, 614–622.

Ito, T., Lyons, R. K., & Melvin, M. T. (1998). Is there private information in the FX market? The Tokyo experiment. The Journal of Finance, 53(3), 1111–1130.

Jacod, J., & Todorov, V. (2009). Testing for common arrivals of jumps for discretely observed multidimensional processes.

Jens, C. E. (2017). Political uncertainty and investment: Causal evidence from US gubernatorial elections. Journal of Financial Economics, 124(3), 563–579.

Jeon, Y., McCurdy, T. H., & Zhao, X. (2022). News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies. Journal of Financial Economics, 145(2), 1-17.

Jiang, G. J., & Oomen, R. C. (2008). Testing for jumps when asset prices are observed with noise–a "swap variance" approach. Journal of Econometrics, 144(2), 352-370.

Jiang, L., Wu, K., & Zhou, G. (2018). Asymmetry in stock comovements: An entropy approach. Journal of Fin Jiang, Jingwen and Kelly, Bryan T. and Xiu, Dacheng, Expected Returns and Large Language Models (November 22, 2022). Available at SSRN: https://ssrn.com/abstract=4416687ancial and Quantitative Analysis, 53(4), 1479–1507.

KAI-INEMAN, D. A. N. I. E. L., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 363–391.

Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2003). Winter blues: A SAD stock market cycle. American Economic Review, 93(1), 324–343.

Kant, N., Puri, R., Yakovenko, N., & Catanzaro, B. (2018). Practical text classification with large pre-trained language models. arXiv preprint arXiv:1812.01207.

Ke, Z. T., Kelly, B. T., & Xiu, D. (2019). Predicting returns with text data (Issue w26186)). National Bureau of Economic Research.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models.

International Review of Financial Analysis, 33, 171–185.

Kelly, B., Pástor, Ľ., & Veronesi, P. (2016). The price of political uncertainty: Theory and evidence from the option market. The Journal of Finance, 71(5), 2417–2480.

Keynes, J. M. (1936). The general theory of employment, interest, and money. Springer.

Kim, S. H., & Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior & Organization, 107, 708–729.

Klibanoff, P., Lamont, O., & Wizman, T. A. (1998). Investor reaction to salient news in closed-end country funds. The Journal of Finance, 53(2), 673–699.

Kyle, A. S. (1985). Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society, 1315–1335.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2), 495–540.

Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment. Journal of Banking & Finance, 26(12), 2277–2299.

Lehmann, B. N. (2002). Some desiderata for the measurement of price discovery across markets. Journal of Financial Markets, 5(3), 259–276.

Lemmon, M., & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. The Review of Financial Studies, 19(4), 1499–1529.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports?

Li, X. (2021). Does Chinese investor sentiment predict Asia-pacific stock markets? Evidence from a nonparametric causality-in-quantiles test. Finance Research Letters, 38, 101395.

Liang, C., Tang, L., Li, Y., & Wei, Y. (2020). Which sentiment index is more informative to

forecast stock market volatility? Evidence from China. International Review of Financial Analysis, 71, 101552.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990a). Noise trader risk in financial markets. Journal of political Economy, 98(4), 703–738.

Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990b). Noise trader risk in financial markets. Journal of political Economy, 98(4), 703–738.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187–1230.

Loughran, T., & McDonald, B. (2020). Textual analysis in finance. Annual Review of Financial Economics, 12, 357–375.

MacKinlay, A. C. (1997). Event studies in economics and finance. Journal of Economic Literature, 35(1), 13–39.

Maia, M., Freitas, A., & Handschuh, S. (2018). Finsslx: A sentiment analysis model for the financial domain using text simplification. 2018 IEEE 12th International Conference on Semantic Computing (ICSC, 318–319.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782–796.

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. Journal of Financial Economics, 123(1), 137–162.

Marcus, G. (2018). Deep learning: A critical appraisal.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors.

Menkhoff, L., & Rebitzky, R. R. (2008). Investor sentiment in the US-dollar: Longer-term, non-linear orientation on PPP. Journal of Empirical Finance, 15(3), 455–467.

Menkveld, A. J. (2013). High frequency trading and the new market makers. Journal of Financial Markets, 16(4), 712–740.

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. Annual Review of Financial Economics, 8, 1–24.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. Journal of Financial Economics, 3(1–2), 125–144.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111–3119).

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. The Journal of Finance, 32(4), 1151–1168.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. IEEE Access, 8, 131662–131682.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica: Journal of the Econometric Society, 347–370.

Nofsinger, J. R. (2005). Social mood and financial economics. The Journal of Behavioral Finance, 6(3), 144–160.

Noise: A Swap Variance Approach. (n.d.). Journal of Econometrics, 144(2), 352–370.

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. Journal of Financial Economics, 144(1), 273-297.

O'Hara, M. (2003). Presidential address: Liquidity and price discovery. The Journal of Finance, 58(4), 1335–1354.

O'Hara, M. (2015). High frequency market microstructure. Journal of Financial Economics, 116(2), 257–270.

Ozturk, S. R., Wel, M., & Dijk, D. (2017). Intraday price discovery in fragmented markets. Journal of Financial Markets, 32, 28–48.

Pagan, A. R., & Schwert, G. W. (1990). Alternative models for conditional stock volatility. Journal of Econometrics, 45(1–2), 267–290.

Park, K.F., Shapira, Z. (2017). Risk and Uncertainty. In: Augier, M., Teece, D. (eds) The Palgrave Encyclopedia of Strategic Management. Palgrave Macmillan, London. https://doi.org/10.1057/978-1-349-94848-2_250-1

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2017). Do funds make more when they trade more? The Journal of Finance, 72(4), 1483–1528.

Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics, 97(3), 683–697.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP, 1532–1543.

Pontiff, J. (1996). Costly arbitrage: Evidence from closed-end funds. The Quarterly Journal of Economics, 111(4), 1135–1151.

Putniņš, T. J. (2013). What do price discovery metrics really measure? Journal of Empirical Finance, 23, 68–83.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. PloS One, 10(9), 0138441.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications, 117, 139–147.

Santosh, S. (2016). The speed of price discovery: trade time vs clock time. SSRN Scholarly Paper ID, 2567486, 43.

Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and speculative bubbles. Journal of political Economy, 111(6), 1183–1220.

Schwert, G. W. (1981). The adjustment of stock prices to information about inflation. The Journal of Finance, 36(1), 15–29.

Schwert, G. W. (1989). Why does stock market volatility change over time? The Journal of Finance, 44(5), 1115–1172.

Shapiro, A. H., & Wilson, D. J. (2022). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. The Review of Economic Studies, 89(5), 2768-2805.

Shapiro, J. E. (1993). Recent competitive developments in US equity markets (No. 93). New York Stock Exchange.

Shastri, K., Thirumalai, R. S., & Zutter, C. J. (2008). Information revelation in the futures market: Evidence from single stock futures. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 28(4), 335–353.

Shiller, R. J. (1980). Do stock prices move too much to be justified by subsequent changes in dividends?

Shiller, R. J. (2000). Irrational Exuberance. Princeton University Press.

Shiller, R. J., & Princeton. Taleb, N. (2005). Irrational Exuberance. Princeton University Press.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. The Journal of Finance, 52(1), 35–55.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers, 24–29.

Smith, V. L., & revSummers, L. H. (2003). Constructivist and ecological rationality in economics. American Economic, 41(3), 591-601.

Solt, M. E., & Statman, M. (1988). How useful is the sentiment index? Financial Analysts Journal, 44(5), 45–55.

Solution, F. (n.d.). Annals of Applied Probability (Vol. 19, pp. 556–584).

Souma, W., Vodenska, I., & Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. Journal of Computational Social Science, 2(1), 33–46.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. Journal of Financial Economics, 104(2), 288–302.

Stock, J. H., & Watson, M. W. (1988). Testing for common trends. Journal of the American Statistical Association, 83(404), 1097–1107.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? The Review of Financial Studies, 24(5), 1481–1512.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance, 63(3), 1437–1467.

The financial/economic dichotomy in social behavioral dynamics: The socionomic perspective. (2007). The Journal of Behavioral Finance, 8(2), 84–108.

Todorov, V., & Tauchen, G. (2011). Volatility jumps. Journal of Business & Economic Statistics, 29(3), 356–371.

Tse, Y. (2000). Further examination of price discovery on the NYSE and regional exchanges. Journal of Financial Research, 23(3), 331–351.

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. Financial Analysts Journal, 57(3), 41–51.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141–188.

Uhl, M. W. (2014). Reuters sentiment and stock returns. Journal of Behavioral Finance, 15(4), 287–298.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., & Zhao, B. Y. (2015). Crowds on wall street: Extracting value from collaborative investing platforms. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 17–30.

Wang, Y. H., Keswani, A., & Taylor, S. J. (2006). The relationships between sentiment, returns and volatility. International Journal of Forecasting, 22(1), 109–123.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 606–615.

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. University of Michigan Business School Working Paper.

Yan, B., & Zivot, E. (2010). A structural analysis of price discovery measures. Journal of Financial Markets, 13(1), 1–19.

Yan, B., & Zivot, E. (2007, February). The dynamics of price discovery. AFA 2005 Philadelphia Meetings.

Yu, J., & Yuan, Y. (2011). Investor sentiment and the mean–variance relation. Journal of Financial Economics, 100(2), 367–381.

Zakoian, J. M. (1994). Threshold heteroskedastic models. Journal of Economic Dynamics and control, 18(5), 931–955.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4), 1857–1863.

Zhang, F. (2010). High-frequency trading, stock volatility, and price discovery.

Zhang, L., Ourkland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association, 100(472), 1394–1411.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC pharmacology*, *10*(1), 1-11.

Anderson-Sprecher, R. (1994). Model Comparisons and R2. The American Statistician, 48(2), 113–117.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. biometrika, 78(3), 691-692.

Kvålseth, T. O. (1985). Cautionary note about R 2. The American Statistician, 39(4), 279-285.

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of econometrics, 77(2), 329-342.