The Natural Language of Finance

forthcoming in *Foundations and Trends in Finance*

Gerard Hoberg and Asaf Manela[*]

June 12, 2025

Abstract

We summarize the wide array of natural language processing (NLP) tools used in financial economics research. These tools empower researchers to incorporate rich but subjective textual data into advanced empirical analysis. NLP tools have pros and cons, and some are better suited to certain research agendas. Research using these tools has exploded in prevalence over the past ten years, and we document the major contributions in corporate finance, asset pricing, and beyond. These tools offer the flexibility to test hypotheses that were not possible before their advent, while also offering improvements in the clarity of identification and the ability to separate hypotheses that purport to explain a set of findings. Finally, we identify challenges and directions for future work.

[*] University of Southern California Marshall School of Business and Washington University Olin Business School, respectively. Hoberg can be reached at hoberg@marshall.usc.edu. Manela can be reached at amanela@wustl.edu. We thank the editors (Sheridan Titman, Chester Spatt and Josef Zechner), the Publisher (Zac Rolnik), and the reviewer (Harry Mamaysky) for excellent guidance and advice. We also thank Federica Ansbacher and Will Cong for excellent comments.

1 Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) and linguistics focused on the interaction between computers and human languages. It involves designing algorithms and systems that enable machines to understand, interpret, generate, and respond to human language in a meaningful way. From an economist's perspective (our focus), NLP facilitates a mapping of subjective content to objective quantitative measures, which in turn facilitates rigorous econometric analysis and inferences not possible before its advent.

The recent launch of large language models such as OpenAI's ChatGPT and Anthropic's Claude awakened large swaths of the population to the possibility that artificial intelligence (AI) will dramatically change many aspects of society including research. Yet for a subset of researchers in finance and economics (and beyond), this was already a foregone conclusion, as many of the same views that led pioneering developers at these companies to develop these novel technologies were already influencing research for quite some time. What truly came as a surprise even to more seasoned researchers is just how fast these new technologies suddenly moved the frontier.

The confluence of these statements is that cutting-edge AI tools such as natural language processing already have a well-developed history with established precedents, but also that the forward-looking pace of adoption will likely far exceed that seen in the past. The goal of this study is to fully document the evolution of these technologies and to document the core strengths and weaknesses of NLP tools in this area, and also to provide some more speculative content regarding where more fertile grounds for research might lie. This article is meant to be relevant not only to academic researchers, but also to students and practitioners interested in utilizing these tools for practical applications.

A core goal of this study is to help scholars less experienced with these tools to understand if these tools are appropriate for their project, to understand the core weaknesses still to be resolved, and to guide researchers regarding how to choose specific NLP tools to address different types of research problems. The high level view is that many projects can benefit significantly or might only be possible with the help of these tools. Yet these tools are not a panacea, as some limitations often misunderstood by some researchers still exist. Finally, we note that there is no one-size-fits-all NLP tool that is "best" for all problems.

Rather, developing NLP based research pipelines is perhaps 60% an art given the widespread heterogeneity of research objectives in financial economics, and 40% weighted on knowledge of which tools are most “cutting edge” in computer science.

Our review begins with a section documenting the core advantages of NLP tools, which are many. These include (A) the ability to measure economic variables that were not possible in the past, (B) the ability to measure variables with greater precision than in the past, (C) the ability to improve the rigor and identification when multiple interpretations are not excluded, and (D) to improve the clarity of interpretation of any result using specialized mechanism tests. Any one of these contributions alone would be a significant contribution. Put together, the benefits to many projects are large.

Yet we are careful to remind readers that these tools are still evolving, and for good reason. They are not yet a panacea for all research agendas and practical applications as we note a number of limitations that future researchers might aim to address. These include (A) some textual corpora are not informative regarding a researcher’s key questions and thus there is too little statistical power, (B) some textual corpora are generated through a voluntary disclosure process and lack consistent reporting, which can lead to false inferences, (C) some agendas are exposed to the “black box critique” and do not allow the researcher to learn why an AI-based model is generating strong findings, (D) less informed researchers might choose AI models that are a poor fit, resulting in inferences with limited value, and (E) many newer LLMs are trained using recent data and can expose researchers to look-ahead bias in many settings where historical predictions are needed.

Establishing a balanced knowledge of both costs and benefits is crucial for the first major decision researchers must make: to pursue a research project or not, and if the project is a go, whether AI tools should be considered. For projects passing this initial assessment, this article next provides guidance regarding which of the many NLP tools is likely a best fit for a given research project. No one tool is optimal for all projects, and hence we develop a researcher-objective-based framework and note that projects come in three flavors. These include (A) targeted objectives that require measurement of a highly specialized economic item in textual corpora or specific document features such as positive or negative tone, (B) objectives that require using all content in a corpus holistically to maximize signal strength to predict an economic variable, and (C) comparative objectives that require large numbers of comparisons between documents assessing their degree of relatedness. We refer to these three as "Research Objective Categories" (ROCs). Researchers looking to choose optimal methodologies should first identify which of the three categories best matches their research objectives.

We next illustrate that the set of candidate optimal NLP tools to consider for a given project is greatly narrowed once the researchers have identified their project's ROC. We then note two additional considerations that help to further narrow the set closer to one, which are the researcher's access to costly resources and the researcher's preference for optimizing signal strength versus clarity of interpretation. Many academic projects, for example, might favor interpretation over signal, but some applied projects might be focused more on signal strength.

Our study then documents the history of several specific NLP tools in the literature while noting which tools best pair with which ROC objectives. These tools include early methods such as basic word lists, more intermediate methods such as thematic topic models and precise anchor phrase queries, to more advanced and cutting-edge techniques such as large language models. We annotate these discussions with views regarding which methods are likely to fall out of favor and how the more cutting-edge LLMs are likely (or not) to replace some tools.

We end our review with a literature summary of key sub-disciplines in financial economics to document how NLP tools have revolutionized the study of key areas. Regarding corporate finance, we examine the study of industrial organization, disclosure and information environments, corporate investments, financial constraints, innovation, and the role of AI in corporate finance. Analogously, in asset pricing, we cover short-term return predictability, the gradual price adjustment to information, risk and uncertainty, institutional investors, social media, and high-frequency return predictability. We also briefly summarize additional areas such as the study of climate finance and banking. We enrich these discussions with notes regarding areas of financial economics where the use of NLP is rapidly growing, and note where fertile ground for continued discoveries might lie. We acknowledge that our summary of sub-disciplines is necessarily incomplete due to space constraints. Our focus is therefore more on building intuition for how researchers successfully used the existing tools and to illustrate how this research can generate particularly successful outcomes.

A complete treatment is impossible. We therefore highlight a few review articles that summarize major developments. These include studies that summarize methods from various perspectives (Gentzkow et al. (2019), Ash and Hansen (2023), Loughran and McDonald (2020), Cong et al. (2020), and Cong et al. (2019)). Goldstein et al. (2021) summarizes key conclusions from a pioneering special issue published in the *Review of Financial Studies*, and Kelly and Xiu (2023) summarizes the use of machine learning in finance in this same journal. We note that machine learning is another tool often adopted in agendas using NLP technologies or AI tools. Our focus on NLP tools is different, although complementary to their focus on machine learning.

This article is organized as follows. Section 2 provides big picture perspectives on the benefits and limitations of NLP tools in finance and is core to project selection and whether a given project is a good fit for using NLP tools. Section 3 provides a framework for researcher objectives, which can guide researchers on methodological selection. Coding and methods are then discussed in Section 4. Section 5 summarizes advances in corporate finance that have benefited from NLP tools, and Section 6 does the same for asset pricing. Section 7 notes impact on other areas in financial economics. Section 8 notes emerging research areas and challenges that methodological researchers have yet to solve, and Section 9 concludes.

2 Comparative Advantages of Text in Finance

2.1 Core Strengths

Textual data and advanced NLP methods offer researchers significant advantages and are a complement rather than a substitute to numerical databases. For example, stock returns are best measured as numerical data, but NLP technologies can improve our ability to predict returns while facilitating the identification of interpretable channels explaining the predictability. Other complementarities arise when numerical databases exist with limitations, and NLP tools can enhance their measurement for specific applications. A typical example is measuring constructs in a more forward-looking way or with a finer-grained classification. For example, NLP can measure corporate investment with greater future predictability than is possible with numerical accounting data, and NLP can add granularity to measure types of investment not observable in accounting data. We now summarize core features that make research projects most amenable to benefits from NLP methods.

- [**New Economic Concepts Need Measurement**] Perhaps the most obvious advantage of NLP tools is to facilitate measurement of economic concepts that are not available using existing numerical databases. An example is the state of a firm's products in the product life cycle (see Hoberg and Maksimovic (2022)). Research objectives that are theoretically motivated but under-researched due to measurement limitations are particularly fertile for NLP-based research.

- [**Existence of Large Informative Textual Corpus**] Also straightforward, NLP methods have leveraged benefits when the research agenda has a relevant large corpus with high coverage of the concepts needing measurement. An example is the 10-K corpus has great coverage of publicly traded firms and devotes a significant fraction of its content to discussing product markets (Hoberg and Phillips (2016)). Benefits rise dramatically with corpus size as NLP tools are automated and scalable.

- [**Signal Strength Maximization for an Objective**] Some research agendas, such as predicting stock returns, often put high weight on maximizing predictive signals. Historical context indicates why this fits well with NLP. Most NLP methods are developed by experts in computer science and linguistics who often focus almost exclusively on signal strength as the primary criterion for publication. For financial economists, these benefits thus come "for free" when using these methods.

- [**Interpretability of Results is Important**] Although older methods were limited, many newer NLP methods are designed to ensure a high level of transparency and interpretability. For example, topic models are designed to provide a full taxonomy of the interpretable content in a corpus. Embedding models and generative AI offer even more opportunities to improve interpretability.

- [**The Researcher Needs to Improve Identification**] As we document in the next subsection, NLP tools offer a number of channels to improve the identification of a given result.

The items above indicate widespread gains and it is easy to conclude that every project can gain from these tools. We thus end this section with important notes on limitations.

We list a set of features that are most frequently misunderstood about the benefits of NLP techniques.

- [**Numerical Constructs Already Exist**] Many research projects rely on predictions for which high-quality numerical measures already exist. For example, insider trading can be measured using SEC filings. A researcher trying to improve this resource using text is unlikely to succeed, although NLP can be used to test deeper theories of insider trading.

- [**Textual Corpus Exists but is Small**] Many researchers have approached the authors of this study asking how NLP tools can be used to measure interesting economic constructs, but later revealed that only a small corpus was available. For example, one such conversation related to a corpus with just 80 documents, each being 4 pages. When the corpus is small, researchers should more straightforwardly consider hand-classification.

- [**Voluntary Disclosure Too Big a Factor**] Much NLP research is based on content in verbal corpora that is not subject to mandatory reporting. Press releases are one example, as the existence of content in this corpus arises from a series of endogenous decisions. In other cases such as SEC filings, much content is mandatory and uniformly present. When a corpus is primarily generated by endogenous choice, the researcher needs to evaluate the tradeoffs of added measurement flexibility (favors use of NLP) versus the extent of confounders (favors caution). Of course, if the endogenous choices themselves are left-hand-side dependent variables rather than presumed exogenous determinants, then this concern vanishes. We discuss this issue in greater detail in Section 2.4.

- [**Resource Considerations**] NLP tools require computing power, coding expertise, and financial resources. Fortunately, ongoing work in computer science and the convergence around the use of easy-to-use Python packages continues to level the playing field. Additionally, generative AI can increasingly write code for a researcher with fairly high accuracy (although current models are not yet perfect and still require oversight by an expert coder).

Overall, the strengths and positive use cases of advanced technologies far outweigh the weaknesses, illustrating why the use of these techniques in finance and economics is rapidly increasing. We next discuss the more specific issues of identification, transparency, and voluntary versus mandatory content.

2.2 Identification

It is obvious that advanced technologies can facilitate the creation of new measures of concepts previously unmeasurable. It is less obvious how these technologies can improve a researcher's ability to conduct causal analysis. We explore this possibility and present an intellectual framework for guiding researchers regarding how these technologies can be used in this way, and how important arguments can be constructed regarding causal conclusions.

The simple act of measuring a new concept does not directly address causality. Yet this kind of step is nevertheless important as it expands the research frontier in a topical sense. Also, basic regressions can establish important associations between new economic variables. These statements do not yield causality, but are an important factual starting point.

In order to establish a causal relationship, researchers need to (A) illustrate that adequate power exists to make a significant inference (for example, powerful instrument tests and F-tests) and (B) the researcher needs to argue that the inferences satisfy the exclusion restriction. As requirements for the first condition are econometric, we will not discuss this condition further here except to note that reducing measurement error (a core contribution of NLP technologies) can make progress on instrument power. We will instead focus on the exclusion restriction and propose an intellectual framework based on two principles.

1. **Principle 1: [Full Exclusion Requires a Sole Interpretation]:** The exclusion restriction is defended by making economic arguments that there is one and only one channel through which a key explanatory variable can influence an outcome variable. If the given explanatory variable is a true random treatment, this argument is easy to make. If the variable is an accurately measured economic variable using text, causal conclusions do not necessarily obtain directly.

2. **Principle 2: [Partial Exclusion and Techniques that Exclude More Interpretations]** We propose that some research results can admit inferences that are "closer

to causal" than others, and this concept of distance can be quantified as the number of plausible economic channels that remain unexcluded as plausible explanations of a given finding. A perfect random experiment has only one unexcluded explanation and fully satisfies the exclusion requirement. A less ideal instrument might be consistent with two explanations, but not more. A lower quality instrument might have 5 unexcluded explanations. Instruments with fewer unexcluded interpretations offer superior "identification". In turn, the random experiment is closest to causality, the second instrument is next closest, and the third is most distant.

Many researchers agree that perfect random experiments are not available to test many of the most important economic questions researchers need to address. Because it is better to produce some high quality research on these topics rather than producing no research due to the absence of perfection, reasonable scholars might consider that an article that reduces the number of unexcluded explanations for a given important finding makes an important contribution. We propose that advanced technologies such as NLP offer researchers the opportunity to improve the quality of exclusion on four dimensions, all of which motivate more research using these tools.

1. **Improvement Channel 1: [Using NLP to construct instruments that plausibly satisfy the exclusion requirement**] Most quasi natural experiments used in the literature (such as unexpected changes in specific laws) are easily measured without advanced technologies. Yet some sources of quasi exogenous variation might have difficult to measure impact on each company in a sample. An example is the Alice vs CLS Bank Supreme Court Decision that suddenly reduced protection for a large number of patents generally relating to algorithms and methods. Unfortunately, no numerical indicator of treatment exists for each patent and hence it is not possible to compute a "treatment effect" that can be used in a panel or a difference-in-difference test using standard numerical data. However, a reasonably large subsample exists indicating a set of submitted patents that were rejected by patent reviewers specifically due to the Supreme Court Decision and another subset that were not. As a result, Acikalin et al. (2022) trains large language models using this ground truth sample to identify exposure to the Supreme Court decision as a function of each patent's detailed

textual descriptions. The result is a predicted susceptibility score that can be easily aggregated to each firm-year, thus creating a computed treatment effect at the firm level and therefore a formal difference-in-difference analysis. We note this improvement channel first because using NLP to construct instruments that plausibly satisfy the exclusion requirement directly is the most unambiguous case regarding how new technologies can assist in causal analysis.

2. **Improvement Channel 2:** [**Using Technology to Reduce Measurement Error of Key Inference Variables**] Many researchers agree that using NLP tools can reduce measurement error in measured quantities. In the context of the two principles noted above, measurement error is an economic problem that can admit an increased number of unexcluded explanations for a given finding. For example, a researcher measuring a new type of investment using straight word lists might inadvertently measure all forms of investment as the word investment itself would be in the word list. Hence a statistically significant finding using this variable might be due to the proposed new type of investment, or it might manifest through picking up signals about other types of investment. Thus at least two unexcluded explanations. In contrast, a precise approach to measuring the specific type of investment (for example using precise anchor-phrase queries with validation), should have very little measurement error. An example is Calomiris et al. (2024b) who identify regulatory sentences in earnings calls and classify them as indicative of increasing or decreasing regulation while using ChatGPT to validate. Such a method thus excludes more explanations as the measure does not pick up other types of investment. It follows that technologies that reduce measurement error generate improved identification. We note measurement error second due to the high potential of advanced technologies to make significant headway on measurement error.

3. **Improvement Channel 3:** [**Using Technology to Construct Highly Specific Mechanism Tests**] In many cases, the researcher is fully aware that a given result might be attributable to more than one unexcluded explanation. Here, NLP tools can help because text is exceedingly rich and voluminous, thus permitting researchers to design highly targeted additional tests of competing mechanisms. For example, the researcher would hypothesize that an auxiliary finding must be visible in the data if

one explanation is the truth, but another finding should obtain if a second explanation is the truth. One example is the finding that vertical integration is more prevalent among innovative firms that have mature realized technologies than among those with immature (not yet patented) technologies as in Frésard et al. (2020). The authors propose contract incompleteness for earlier stage firms as a candidate explanation, as innovation incentives are diluted under integration. Another explanation could be that some technologies take longer to patent or are difficult to patent. The authors use 10-Ks to generate measures of contracting frictions and legal challenges regarding technology. The authors illustrate that the main results strongly associate with scenarios where these types of legal issues are prevalent as firms disclose these issues in their 10-Ks more often than control firms. These types of tests move the overall conclusions closer to a causal inference as specialized evidence favors one interpretation. At a minimum, this evidence of a proposed mechanism increases the likelihood of a proposed explanation as being supported, especially when tests are particularly targeted and specific to one explanation. Yet this approach has limitations as high quality mechanism tests might not exist or available tests might be imperfect.

4. **Improvement Channel 4:** [**Casusal NLP Techniques**] Work in computer science including Athey and Imbens (2015) and Veitch et al. (2020) suggest that causal conclusions using NLP are possible under some (albeit strong) assumptions about textual data quality. Under the assumption that a textual corpus is all-informative regarding a key outcome variable, causal effects could be estimated through a "strong control" approach. For example, suppose that the text of patents is all-informative about the expected value of future patent cites and patent quality. Then one can control for the fundamental quality of the patent and separately identify a secondary effect such as the impact of gender or race. By exploring counterfactual models (for example considering the role of gender or race), it is possible to identify the magnitude of such potentially causal effects. Representative studies illustrating this approach include Hochberg et al. (2023) and Cong and Yang (2024), who examine biases in patent citation and USPTO patent applications, respectively. The former takes a traditional approach and estimates two separate neural networks for female and male inventors and then estimates counterfactuals by examining predicted cites under gender reassignment. The latter

extends the causal NLP framework to study race and other features, and incorporates data on human impressions of names, which can be juxtaposed against actual race data. Their novel approach, in turn, can achieve causal identification with materially weaker assumptions. Overall, Causal NLP is still nascent and the general assumption of all-informative controls is still difficult to assess in many settings. Yet we believe this approach has promise even if currently making only partial progress on identification. We also believe additional work strengthening this approach is likely to continue.

5. **Improvement Channel 5: [Using Technology to Add Important Controls]** In some cases, a researcher might be worried about a specific alternative that is difficult to observe. Yet econometrics would indicate that adding a control for the alternative variable would be a solution as this issue is a problem of omitted variables. In this case, the researcher might use new technologies to measure the important control variable. The researcher also should focus on reducing measurement error in the control. An example is Bowen et al. (2023), who measure technological breadth from patent text as an important control when identifying the effects from rapidly emerging technologies (a different text-based measure). If the original results are robust to including the new control, the finding would once again move the inferences closer to a causal conclusion as another explanation is excluded. Even if the distance is not closed fully, this moves the literature forward relative to what precedes it.

Overall, causal inference is one of the greatest challenges of modern research in financial economics. We argue that research making continued progress toward causality, that moves toward closing the gap and excluding additional explanations for key results, can make important contributions. Moreover, each contribution can set the stage for new researchers to further reduce the number of alternative explanations. In time, some of the more difficult and important questions researchers need to answer might become understood with more clarity and rigor.

We conclude this section by noting that text can also be used to generate and test hypotheses (see Ludwig and Mullainathan (2024); Manning et al. (2024)). In the context of identification, this for example can guide the researcher regarding what the set of non-excluded hypotheses might be.

2.3 Transparency: Avoiding the “Black Box Critique”

The “Black Box Critique” has a long history in finance with roots at least as far back as the failure of Long-Term Capital Management in the 1990s. It is a significant concern, and researchers using NLP tools need to be cognizant of its risks, especially regarding agendas with policy relevance or with practical applications where economic agents might make important decisions based on the results of the research. As the term has some roots in the popular press, we introduce the concept through with a brief (fictional) narrative.

Narrative: *A superstar statistician presents a hedge fund with a software application that takes large amounts of stock price data as input, and outputs portfolio weights to invest in for the next week. Back-tested portfolios achieve an incredible Sharpe ratio of 7.0 even using liquid assets. As this performance far exceeds any strategy currently used by the fund, they follow the algorithm’s advice. As the profits pile up, the fund increases its leverage on the strategy. But one day a major event in South America shifts returns dramatically and the fund loses 90% of its assets under management before they could exit their leveraged positions. Fund managers lament listening to the black box and have no clue where they went wrong.*

The most fundamental lessons from the story are that decision makers need to know why a given model is predicting outcomes in the way it does. Models trained to maximize one economic outcome (such as the first moment of the stock return) might also inadvertently maximize another (such as exposure to systematic risk). A predictive model with no interpretability is therefore potentially dangerous. On the other hand, if the black box could provide interpretability, the decision maker could enhance their strategy with risk management to more safely optimize their objectives.

Concerns about the black box critique (henceforth BBC) have now become entrenched in how well-trained financial economists evaluate research, how policies are evaluated, and how practical decisions are made. The authors of this study are supportive of this transparency movement and we note another feature of the current academic system that exacerbates the concern. The issue is that financial economists borrow advanced technologies from other fields including computer science, linguistics and data science. The publication requirements in these fields often hinge on success in predicting a first moment outcome. In contrast, proving that the algorithm also has inbuilt risk-management protections is not a central publication requirement, leaving the crucial task to financial economists using the methods. At issue is whether or not an algorithm can document why a prediction is being made and what other objectives might be (unintentionally) maximized.

The lack of attention to interpretability among algorithm developers in these other disciplines has been changing rapidly as demand for interpretability is rising among scholars that use NLP (including financial economists). As of today, researchers in financial economics should be able to provide both predictability and interpretation with a much improved level of granularity. This makes the Black Box Critique more an issue of the 1990s and 2000s and less an issue for the 2020s and beyond (provided researchers are trained to use the interpretability tools and these tools are high quality). Yet the BBC has two features: (1) interpretation of a key finding and (2) unforeseen risk exposures unseen in the primary channel. The second issue of unmapped risks is more difficult to fully address. We thus encourage researchers to invest more efforts into using new technologies to improve risk management and to develop tests to assess the ability of various technological solutions to weather unexpected events.

We now briefly note which NLP tools provide notable interpretability tools, while noting that achieving interpretability is a major accomplishment and needs project-specific thinking. This matter ultimately is more an art than it is a simple python script, and hence it is discussed and highlighted in many different ways beyond this list throughout this full article.

- **[Early Attempts]:** Earlier methods that had elements of transparency include word lists, as a researcher knows exactly which words the calculations are loading on. Basic document cosine similarities facilitate reporting overlapping words most important in documents being compared. Yet both attempts, while helpful to some extent, do not fully inform what truly drives a predictive signal. Word lists can show the mechanics of a calculation but do not reveal which words actually matter, and which words might be correlated with the words on the list but are not themselves on the list (confounders). The most daunting issue is that any given word that matters might matter for many reasons as the calculation does not give a clear causal explanation (omitted correlated variables being focal). Regarding cosine similarities, the issue is different and is that reporting lists as examples does not clearly indicate what matters overall for a sample with millions of observations as the examples might not be representative. These examples illustrate good intent but the authors of this study view them as only suggestive by 2020s standards, particularly regarding research addressing policy domains and practical applications where human well-being is at stake.

- **[Topic Models (A Big Idea)]:** The advent of topic models such as Latent Dirichlet Allocation (Blei et al. (2003)) came relatively early in 2003 and was seen as a major leap forward on interpretability. Akin to factor models in asset pricing, when documents in a corpus are highly thematic and have a strong "verbal factor structure", topic models offer a full decomposition of a corpus into its systematically present topics. Remarkably, LDA is a methodology that can generate this topic structure (known as a taxonomy) with very little researcher input. Generally, researchers need only specify an integer indicating the granularity of the taxonomy they would like LDA to generate.[1] As we will discuss in the next section on methodological selection, topic models such as LDA, while revolutionary by many definitions, have not achieved the originally envisioned success as the factors generated by LDA are not as clearly interpretable as hoped in many settings. This of course is a matter of opinion and the authors of this study see some successes but nevertheless conclude that scholars in linguistics and computer science need to innovate further in this space before success can be declared.

- **[Anchor-Phrase (Limited in Scope but Precise)]:** The anchor-phrase methodology is an advanced and highly targeted tool that identifies specific types of content in a corpus with unambiguous interpretations. We discuss the details of this approach in the next section. A consequence is this approach is directly interpretable with significantly reduced ambiguity. This method enhances interpretability but is only applicable for projects requiring a measure of highly specialized content. If a researcher aims to understand all content in a corpus and needs a taxonomy, the anchor-phrase approach is not helpful.

- **[Embeddings and large language models]:** Embedding models in their modern form rose to prominence with Mikolov et al. (2013), building on earlier ideas such as

[1] Although we note there are some techniques in computer science can automate this selection. See Newman and Girvan (2004), Newman (2006), and Newman et al. (2010) for computer science methods, and Mamaysky (2024) for an example in finance.

Latent Semantic Analysis (LSA) from the 1990s. These models indicate a significant leap beyond basic topic models such as LDA noted above. Unlike LDA, which is a bag of words model that does not consider the proximity of words in the text and thus lacks power, embedding models consider context and can internalize the importance of words that co-appear in near-proximity to a focal term. Two studies, Hanley and Hoberg (2019) and Cong et al. (2019) notably show that combining the logic of topic models explicitly or implicitly with the context of embedding models can generate material improvements in the ability to attribute results to interpretable themes. Larger language models such as ChatGPT and Anthropic's Claude further leverage these benefits as additional prompting can reveal increasingly compelling interpretations of economic predictions.

- [**Automated Hypothesis development**]: Although this area of work is nascent and any predictions are thus speculative, we note that the possibility of using NLP tools to auto-generate hypotheses explaining a particular finding can be fruitful regarding improving interpretability. Ludwig and Mullainathan (2024) is a notable example and suggests that computer-generated hypotheses can identify a set of explanations for a prediction even when the economist running the analysis is unable. Intuitively, understanding the wider range of potential explanations seems crucial in informing economists regarding both how to improve identification and manage risk.

2.4 Voluntary versus Mandatory Reported Content

One of the more significant challenges affecting some research using NLP methods is whether the verbal content being analyzed is the result of a mandatory or voluntary disclosure framework, and whether the content is subject to strategic disclosure and might not be truthful. For example, documents like 10-Ks are mandatory for public firms and are enforced by regulation. On the other hand, discussions in social media or discussions about some specific topics even within 10-Ks are less monitored by the SEC and can be voluntary. An example is discussions of future acquisition plans, which firms might or might not disclose as part of their corporate strategy. The distinction is important because, when disclosure of content is not required (voluntary), the economic inferences the researcher obtains might be confounded by the incentives firms have to disclose the key information or not. For example, a researcher might find that statements about future mergers are negatively related to stock returns. Yet this in itself might not indicate that future mergers cause lower returns as it might be the case that firms with the best merger opportunities do not disclose such statements to maintain a proprietary advantage. Hence such disclosures might predict returns due to a selection effect and not mergers themselves.

This type of voluntary content is often referred to as strategic disclosure, and the extent to which it is a problem is related to the likelihood that firms disclose or do not disclose the focal content. Some information is universally disclosed such as a description of a firm's business. But other information including more speculative forward looking statements might be less commonly disclosed. Generally 10-Ks are less exposed to this issue than are many other corpora because the 10-K itself is mandatory among public firms, and much key information is thus widely disclosed. A notable addendum to this discussion is that strategic disclosure itself can be studied as a dependent variable. In this case, endogeneity of this sort is desirable and not an empirical issue of concern. Understanding strategic disclosure is a fruitful area for future research.

When faced with more endogenous voluntary disclosure regimes, researchers seeking clear identification have some tools they might consider to mitigate the endogeneity concerns. First, they might explore subsamples where disclosure is more prevalent or where it is more likely to be mandatory. Second, if available, the researcher might consider exogenous instruments that shift the focal economic attributes of the content being researched. Third, the researcher might explore more powerful NLP tools that increase coverage or that examine multiple documents.

Finally, researchers can embrace the endogeneity using a structural model. For example, Kelly et al. (2021a) develop an economically motivated high-dimensional selection model that can improve learning from text (and from sparse counts data more generally). They model separately the choice to include a phrase versus the choice of how frequently to repeat it. Importantly, their model allows for parallel estimation, making it computationally scalable. A limitation of their model is that it treats text data as a bag-of-words or bag-of-phrases, disregarding the order of phrases in a sequence of text. A promising avenue for future work would be to incorporate some of the large language model innovations we discuss below into more sophisticated text selection models.

Practically, many studies address the selection concern by showing a body of evidence in many settings supporting a theory, and showing non-consistency with alternative hypotheses. Please see the earlier Section 2.2 on how NLP tools can be used to address endogeneity as that content is directly relevant to this discussion. We will thus keep this section parsimonious given that the remedies are largely the same. We conclude this section by noting that voluntary versus mandatory disclosure, and the propensity of document writers to be fully truthful are big and important questions and future research in this area remains fruitful. We encourage future researchers and econometricians to develop a wider toolkit illustrating new ways this challenge can be addressed.

2.5 Generated Regressor Issues

Natural language data like text allows researchers to generate novel empirical measures of interesting social phenomena. A sometimes overlooked byproduct of this process, however, is that a regression of an economic outcome on text-based measures requires generated regressor adjustments when assessing the parameter uncertainty associated with its estimates (Pagan, 1984).

If a researcher estimates a text-based measure in a first stage and then regresses an economic outcome on the text-based measure in a second stage regression, the confidence intervals around the second-stage coefficients can be adjusted using the Murphy and Topel (2002) approach. For example, Manela and Moreira (2017) reports standard errors that account for the fact that its main regressor, news-implied volatility (NVIX), is estimated in a first-stage text regression.

Battaglia et al. (2024) shows that such inference issues extend to AI- and ML-generated variables more broadly and can lead to biased estimates and invalid inference. The paper proposes two methods to correct bias and perform valid inference in such settings. As the importance of NLP tools grows, we expect the importance of such inference adjustments to play an increasingly important role in careful empirical analysis.

3 Research Objectives

3.1 Intellectual Foundation for Optimizing Methodology

The typical financial economist or finance practitioner at the time of this writing has experienced years of training in disciplines most often spanning business, economics, and mathematics. Some, but not the majority, have had significant exposure to computer science. Among those who have, fewer have had this exposure at a high level in the past 5-10 years. Hence a large number of active researchers and practitioners know little about the wide ranging set of tools that are available in the natural language processing (NLP) toolkit, and even fewer likely have a high degree of confidence in selecting the best tools to use for a particular research agenda. This section aims to close this gap. As the potential contributions that can be established using the growing volumes of textual, audio and visual data is tremendous in scope, improved methodological selection could greatly accelerate progress and the quality of research in the field.

By a wide margin, the most important starting perspective is there is currently no "one-size fits all" NLP methodology that best implements all (or even most) research agendas.[2] We propose that methodological selection is best implemented using a simple "research objective" classification system coupled with refinements based on a couple of key "researcher considerations." We present this framework below. The resulting system typically reduces the optimal methodological choice for a specific project down to just one (or at most 2-3) candidate methodologies to pursue.

Our objectives-based framework is meant to support research using textual corpora. We separately discuss audio and visual research in a later subsection. Our discussion will henceforth take the perspective of a researcher looking to answer a given research question using a given textual corpus. As examples, we will discuss two popular corpora in financial economics research: a corpus of 10-Ks from the SEC Edgar website and a corpus of newspaper articles. Although we take the perspective of an academic researcher, the same logic applies to a hypothetical practitioner looking to build a practical solution to a business problem or to build software that fills a void in the profession.

[2] Yet the authors note that, in coming years (perhaps a 5-10 year horizon), more advanced artificial intelligence tools with high quality prompting might move us closer but not fully to a one-size fits all approach.

3.1.1 Research Objective Categories (ROCs)

We propose that essentially all NLP-based research questions in financial economics fall into one of three major "research objective categories."

- **ROC-Targeted.** This category includes research objectives that require the extraction of a highly targeted and specific concepts from a corpus. An example is identifying which companies mention that they hedge risk using derivatives in a corpus of 10-Ks. ROC-Targeted also includes projects that score text on positive or negative tone, readability, uncertainty or other themes such as political slant.

- **ROC-Holistic.** This category includes research objectives that require extracting a maximally informed aggregate signal based on all content in an entire corpus to predict or explain an economic variable. An example is identifying the "total informativeness" of 10-Ks to explain corporate valuation ratios or using all information in newspaper articles to predict stock returns. This category also includes projects aimed at developing taxonomies modeling the complete diversity of content that exists in a corpus.

- **ROC-Comparative.** This category includes research objectives that require (a potentially large number) of comparisons between documents in the corpus. An example is building an industry classification from the business descriptions in 10-Ks (see Hoberg and Phillips (2016)), which requires comparing how similar documents are for every pair of companies in a given sample to identify which firms have similar text and thus are likely operating in the same industry. Another example is evaluating how intensively documents in a time series are revised (a comparison of versions of the same document) to model the arrival rate of new information (see Hanley and Hoberg (2012) and Cohen et al. (2020b)). Finally, Goetzmann et al. (2022) is an example of ROC-comparative methods to assess the similarity of media articles to market crash narratives following the 1929 and 1987 market crashes.

The first step in selecting a potentially optimal NLP tool for a given research question is to classify the given research objective using the above system. This reduces the set of candidate methods to a manageable level. Before indicating the set of objective-method pairs, we note two important researcher considerations that will further refine which methods are likely optimal for a given project.

- [**Signal Strength versus Interpretability Consideration:**] The researcher should decide if they are more interested in maximizing signal strength and predictability, or in maximizing the interpretability of the results. Optimizing both would be ideal, but existing tools often require trading off one for the other.
- [**Budgetary and Programming Resources Considerations:**] Does the researcher have access to significant resources in the form of programming capacity and budget? Alternatively, is a speedier and lower cost approach needed? More advanced methods are typically more challenging and costly to implement although this hurdle is gradually declining in magnitude. Low cost methods, on the other hand, typically are less accurate.

Finally we note that generative AI LLMs such as ChatGPT, Claude or Gemini can also help with selecting potentially optimal NLP tools. For example, asking these models directly as in the following example is typically quite informative: "If I want to find instances of mentions of lawsuits in earnings calls, what is the best NLP tool to use?"

3.1.2 From Objectives to Methodological Selection

We next briefly summarize which methods pair with each ROC class, and we provide guidance on how the two above considerations further reduce the candidate methods. A researcher can limit their attention to the most relevant subsections below to fully assess implementation.

ROC-Targeted: ROC-Targeted projects either need tools that can identify instances of highly-specific economic statements or tools to measure document features such as tone/sentiment, readability or uncertainty.

Regarding highly specific economic statements (such as statements indicating hedging activity or that a firm is financially constrained), this objective is challenging due to the sheer size of many textual corpora. A consequence is the desired content is generally rare in the corpus and document writers often express business concepts in different ways and using different writing styles. The overarching measure of success is to minimize both false positives and false negatives while providing adequate flexibility and transparency for validation and replication. From least to most advanced methods, a researcher might consider: word lists, embedding-model-informed word lists, precise anchor-phrase queries, or generative AI. Regarding the two considerations, researchers focused on interpretability and identification might favor anchor-phrase queries due to their clear interpretation, flexibility and replicability. On the other hand, a limitation is that the risk of false negatives can be non-trivial when complexity is high. A researcher more focused on signal strength might instead prefer embedding-model-informed word lists. This approach will likely reduce false negatives but expose the researcher more to false positives. Regarding resource considerations, a researcher with a larger budget and more coding sophistication would likely achieve maximum performance using generative AI, although with some reductions in replicability. Interpretability is not automatic with generative AI, but it is possible to generate interpretability through the development of additional prompts to the model.

Regarding projects measuring specific document features such as tone or uncertainty, researchers should consider (from least to most sophisticated) tools such as word lists, embedding-model informed word lists, text regression, generative AI, or application-specific methodologies derived by other researchers. This class of research projects was perhaps the first that researchers in financial economics considered in earlier years. As a result, many seminal studies based on sentiment word lists laid a foundation for much subsequent research aimed at improving these methods. Significant methodological variation now exists. Methodologies developed by other researchers can be a compelling option in many cases. Examples of such researcher-proposed methods include word lists constructed by training machine learning technologies (e.g., Frankel et al., 2022; García et al., 2023), or formulas derived to measure readability or "fog" (e.g., Loughran and McDonald, 2014).

The simplest approach to ROC-Targeted research, basic word lists, is generally losing favor due to the increased flexibility and accuracy of more advanced methods. Yet this method can remain relevant for researchers looking for preliminary estimates of a concept's relevance or that face limited resources. Another benefit of basic word lists is that manual curation of terms can benefit from a given researcher's domain expertise. Yet we note that domain expertise can be applied perhaps more impactfully if the domain expert guides the construction of precise anchor-phrase queries rather than simple word lists. Additionally, embedding models can further inform the domain expert of related content they might not have been aware of.

Finally, mixed methods can improve scalability and lower cost while preserving accuracy. For example, anchor phrase queries (fast and inexpensive) can prune a large corpus down to relevant candidate paragraphs, and generative AI models (slower and more costly) can then validate the small set of selected paragraphs with high accuracy (see Calomiris et al. (2024b) for example). Overall, while a few methods might initially seem relevant for ROC-Targeted projects, each researcher's views on the additional considerations above, and the specificity of measuring some concepts such as document tone, typically refines the set to just 1-2 likely best-practice models for a proposed research project.

ROC-Holistic: ROC-Holistic projects need tools that can model the full richness of content in a corpus and then use the resulting data structures to predict or explain a given economic variable. A researcher might also be interested in building a taxonomy of the overall content in the corpus as this can increase interpretability. This is ambitious and current methods are good, but they are also imperfect and methods will likely continue to be developed in this area.

This can be seen as diametric opposite from the goals of the targeted research objectives noted above. A targeted research question seeks to filter out the overwhelming majority of the content in the corpus in order to focus on rare mentions of a specific concept. In contrast, a holistic research problem generally loads very little on any one specific concept. An ROC-Holistic researcher should consider tools such as thematic topic models, embedding technologies, and generative AI. These methods automate the process of categorizing documents into their main content-subcategories with varying levels of sophistication based on their size and the depth of their use of context in training.

Early methods such as word lists are impractical for holistic research as the number of topics in many corpora is large and a researcher would have trouble self-identifying all available topics. Regarding researcher considerations, the more advanced methods noted above tend to dominate on most dimensions except for the resources consideration. Generative AI models come with dimensionality reduction tools akin to embeddings, but typically require more budget and programming particularly when interpretability is important. Topic models and simpler embeddings, in contrast, can be generated at lower cost. Topic models such as Latent Dirichlet Allocation can quickly generate holistic representations with little input, while providing basic (albeit often noisy) interpretations. Embeddings, on the other hand, require additional programming and additional tools to ensure interpretability. These models can be significantly more interpretable than topic models such as LDA (which use little context as it is a "bag of words" approach).[3]

ROC-Comparative: ROC-Comparative projects need tools that can compare documents accurately regarding their relatedness or similarity, typically while preserving scalability to process large numbers of documents. The most common utilizations include evaluating the similarity of pairs of documents to identify peers or evaluating the extent to which documents are revised when they are available with multiple versions revised over time (such as IPO prospectuses). An ROC-Comparative researcher should consider (from least to most sophisticated) tools such as cosine document similarity, embedding technologies, and generative AI (but rather than prompting, they would use the less-well-known spatial modeling functionality of these technologies).

While basic cosine similarities were prevalent in earlier years before the advent of embedding models, they have become less preferred in recent years as they produce weaker signals and yet offer little advantage regarding the resource utilization consideration. Both cosines and embeddings typically require similar efforts to prepare documents, both are typically coded in Python, and both can be implemented using pre-written code packages. A recent development is the use of embedding functionality that is built into generative AI. This approach has little track record, but intuition would suggest that the primary tradeoff is they will produce improved accuracy but this will come with a financial cost for larger processing jobs (as well as reduced replicability and concerns about look-ahead bias).[4]

[3] Bag of words (BOW) is a label given to models that only condition on existence and frequency of words in a document, and hence that do not condition on the relative position of words in a document and how that informs interpretation. Generally BOW models are becoming less prevalent due to their obvious limitations.

[4] Replicability can be improved by using open source models, or models that allow setting a temperature parameter to zero to reduce randomness. Look ahead bias can be mitigated by using masking techniques or models trained on data pre-dating the researchers prediction sample.

4 Methods

We next develop intuition and discuss specific methods in detail, summarizing their historical advantages and limitations, providing noteworthy examples, and noting each method's likely potential going forward.

4.1 Using Generative AI to Code NLP

Although the primary goal of our study is to provide readers with a framework for assessing their project's match with NLP methods, how to select the right NLP tools, and to understand core issues such as identification and the history of NLP in financial economics, we briefly note in this section some insights regarding how financial economists (who are not trained programmers) can quickly learn techniques for fast and accurate coding at low cost. The most significant insight on this is that generative AI LLMs have inbuilt capabilities to generate high-quality Python code or even debug existing code. They can also be used in a conversational way to explain to an untrained researcher how to get started using a Unix terminal or how to run scripts, and how to fix them when various errors are returned by the terminal.

Here is an example of a prompt that was handed to Claude, Anthropic's generative AI model for a ROC-comparative doc2vec embedding project:

> **Prompt to Claude:** *I have a folder* `c:\corpora\` *containing one subfolder for each year from 1999 to 2018 (for example* `c:\corpora\1999\`*). Each of these yearly folders has a large number of documents. Can you create a Python program that does the following separately for each year: (1) detects all files in the given year's folder, (2) imports each file as a separate document in a corpus, (3) then does any suggested preprocessing and cleaning to prepare it for use in a doc2vec model (in my application, I especially need to remove numbers or any special characters but preserve all words), (4) then trains a doc2vec model on this corpus (use standard settings but dimension needs to be 300), and (5) then extracts the 300 element vector representation for each document, and (6) then runs a loop to compute pairwise cosine similarities for each pair of documents in the corpus, and (7) exports the pairwise similarity matrix as a 3*

> *column tab delimited file with the three columns being the name of the first document in the pair, the name of the second document in the pair and then the cosine similarity. This file can have a header row of "doc1" "doc2" and "score". The file should be saved in* `c:\corpora\output\`*. The name of the file should be "pairsYYYY.txt" where YYYY is the year you are running based on the above loop.*

The result is that Claude provided a complete Python script as an attachment, which could be downloaded and run directly on a Unix terminal. The result was nearly bug-free and was generated in seconds. We also note that bug rates can be further reduced by first asking Claude if he has any clarification questions to ask before generating the code as the models have some ability to detect vague instructions. Any bugs that were present in the example were quickly fixed by prompting Claude with the specific error message received (for example, using a prompt of this form "I received this error <<<copy paste error message here>>>. Can you fix the code or suggest what I should do?"). The debugging process was resolved quickly. One tip is that debugging will be less needed and faster if the original prompt provides more detailed steps.

The narrative response from Claude after the original prompt also told the user which packages they will need to install in Python to run the script it provided (specifically it indicated they need to run "pip install gensim numpy pandas scikit-learn"). Claude's excellent narrative summary also indicated any judgment calls it made in producing the final program. For example, the above prompt did not indicate how many epochs to train doc2vec using, and Claude indicated in its narrative that it selected 20 epochs as its choice. In this example, the expert's experience was that 50 epochs was more accurate in their past projects, and hence they asked Claude a follow-on request "Can you edit the code to change the number of doc2vec training epochs to 50" and they immediately had an updated script. An aside is that an untrained scholar might not know what epochs are, but once again, this knowledge is quickly acquired simply by asking Claude to explain what they are and what considerations a researcher might consider in selecting the number of epochs. Generally developing code through a conversational approach such as in this example has many benefits over and above the old model of simply outsourcing all coding to a third party, as the old model often leaves the financial economist uninformed regarding important judgment calls or potential issues in the code. Yet there are still benefits to sharing the resulting code and conversation with a computer scientist as a safety check or for implementation. This is particularly true for novices who are less trained on issues such as model parameters, alternative models or how to use high performance terminals.

Claude's narrative also indicated that the python script will change all letters to lowercase (a good practice as meaning of a term should not be different in many settings based on case). Finally, the narrative indicated that the script included built-in error-handling as Claude is aware of potentially problematic situations and it can alert the researcher to them during processing. In reading the narrative, for example, the expert realized that a few of the documents in their corpus are degenerate and only have a tiny amount of text. In this case, they asked Claude the follow-on prompt "can you edit the code to have it ignore any documents in the corpus that have fewer than 25 words?". Claude immediately updated the script.

The objective of this section is to indicate to less trained NLP experts that python code is easy to develop using a conversational approach with LLMs such as Claude or ChatGPT and the quality is generally quite high. There are other LLMs on the web that are designed specifically for coding although stock LLMs such as those referenced here are already quite good. To properly manage this process, the expert should prompt the LLM with a sequence of things that need to be done in the code as was done above. The more detail provided, the less follow-on requests need to be made. Yet the system is flexible enough that some experts might develop their ideal program over a longer conversation with the LLM.

This practice significantly reduces the barriers to entry for financial economists to develop code for their projects even if they are not trained coders. Moreover, the LLM's code tends to be optimized for speed, clarity, and efficiency given its deep training. The LLM also provides comments in the code it generates so it is clear to a third party reviewer (something human coders often forget to do). Going beyond the above examples, scholars have used conversations with LLMs to learn how to use Unix to run Python scripts, how to install python on a new machine, how to create virtual environments, how to write batch processing code, and how to handle parallel processing. This LLM approach can also be used to generate database management code in SAS or econometric modeling code in Stata that directly writes output in the form of Latex tables as other examples. One author of this study had his entire faculty website's html code drafted by LLMs in roughly 30 minutes with style and performance exceeding their old site. The result also had a much more professional look and feel.

As we summarize specific NLP tools below, we briefly mention some of the python packages that are typically used for each method, but we do not provide detailed coding practices due to space constraints. This is also because we suggest that readers consider using LLMs to draft their code directly or to work with research assistants to do so, making it even less relevant for us to provide this content. The rapid progress of code generating models has lowered the barriers to applying NLP in finance.

A word of caution, though, about relying on LLMs to generate code. We encourage researchers to only rely on LLMs to generate output that the researchers can verify themselves. The LLM can save researchers time spent locating the right function or syntax to use in a piece of code. But eventually, the researcher must understand the end result and be confident in its correctness. Like research assistants, AI-based research assistants can make mistakes, and such mistakes are the responsibility of the researcher. One practice experts use to ensure quality is to save small samples of the data before and after significant calculations. The researcher would hand-examine these outputs to ensure they match expectations. If they do not, they can reassess the foundation or converse further with the LLM to make corrections. It is also important that the researcher understands the structure of the text corpus used. For example, the researcher may need to decide which articles should be included in the analysis, how to deal with revisions, and which firms an article relates to. The LLM can help, but asking it to process the entire corpus may be computationally infeasible. In general, a deep knowledge of the data, the tools and an understanding of what the code does are important for successful research.

4.2 An Intuition-Building Textual Sample

To add intuition before we compare different methods, we next develop the following imaginary business outlook passage from a corporate 10-K filing. We use this as a running example when discussing how various methods work:

> Looking ahead to fiscal 2024, we anticipate strong growth opportunities in our cloud computing segment, particularly in AI-enabled enterprise solutions. While we face in-

> creasing competition from both established technology firms and emerging startups, our R&D investments of $850 million in 2023 have strengthened our market position. We expect our operating margins to remain under pressure in the first half due to ongoing supply chain constraints and rising labor costs, especially in semiconductor manufacturing. However, our recent acquisition of SecureNet Technologies for $275 million should enhance our cybersecurity offerings and create cross-selling opportunities. The current macroeconomic environment, including potential interest rate changes and geopolitical tensions, may impact customer spending patterns. Nevertheless, our diversified revenue streams and robust balance sheet with $2.1 billion in cash provide financial flexibility. We plan to continue expanding our presence in emerging markets, particularly in Southeast Asia where we see significant untapped demand for our digital transformation solutions.

While this example is not uncommon, aspects of the text illustrate many standard challenges in NLP research. The starting point is the question: how should a researcher represent this short piece of natural language, especially if they seek to link such text to numerical databases for econometric analysis? The challenge is that even this simple 150-word paragraph is a surprisingly high-dimensional object. A word like "strong," which appears twice, hints at a positive outlook. But context matters, as the meaning of each word depends on the words that precede or follow it in the sequence. Although a word like "emerging" sounds positive in isolation, it hints of trouble in the context of "increasing competition from [...] emerging startups." To reduce its dimensionality for use in subsequent analysis, researchers often ask a more specific question such as "Will this firm outperform its competitors in 2024?" From this narrower perspective, a negative outlook seems appropriate given that the firm faces "increasing competition" and "operating margins" that are "under pressure." Overall, the broader the research question posed, the more challenging the dimension reduction problem becomes.

A more data-driven approach might collect a sample of such business outlooks from many firms and time periods and try to run a regression of future performance on the text. But how would the text be represented in such numerical models? Would the researcher use simple word counts? To get more of the context around each word, they might consider 2, 3, or n word sequences (n-grams). As n grows, the ability to capture context improves, but at the same time, the dimensionality of the explanatory variables grows quickly and the density of observed n-grams declines quickly.

An increasingly common approach as we write this article in 2025 is to prompt generative AI models for predictions. But how do such models come up with their answers and is there bias?

The natural language processing literature has gradually developed better tools to address these questions and to generate low-dimensional representations of natural language. Skilled researchers have successfully used these tools on a wide array of specific research questions. We summarize many of the most widely used methods and how they match with various researcher objectives.

4.3 Basic Word Lists

Word lists represent the first major methodological innovation in computational textual analysis in finance. Their introduction enabled systematic analysis of qualitative information at scale, giving rise to several seminal contributions that established textual analysis as a mainstream methodology in financial economics.

The word list approach reduces our example text to simple frequency counts across predefined categories:

$$\begin{aligned}
\text{Positive} &= 5 = \mathit{sum}\left(\text{strong} : 1, \text{opportunities} : 2, \text{enhance} : 1, \text{robust} : 1\right) \\
\text{Negative} &= 3 = \mathit{sum}\left(\text{constraints} : 1, \text{pressure} : 1, \text{tensions} : 1\right) \\
\text{Uncertainty} &= 3 = \mathit{sum}\left(\text{anticipate} : 1, \text{expect} : 1, \text{may} : 1\right)
\end{aligned}$$

The word list approach offers computational simplicity and clear interpretability, making them particularly appealing for preliminary assessments in textual analysis. However, this approach suffers from significant limitations when additional rigor is needed: it ignores word context and word orders in text, fails to capture complex semantic relationships, and it cannot account for negation or qualification of key terms. These limitations often lead researchers to employ this method primarily as a preliminary baseline or in combination with more sophisticated approaches. Modern technologies offer improved power and accuracy, particularly regarding specific interpretation rather than broad content assessment.

Early influential work by Antweiler and Frank (2004) examines message board content and stock market activity, finding that posting volume predicts return volatility and trading volume. Tetlock (2007) provides foundational evidence that media pessimism predicts lower market returns followed by a reversion to fundamentals, leveraging the General Inquirer's Harvard IV-4 psychosocial dictionary to measure pessimism. Building on this, Tetlock et al. (2008) show that negative words in firm-specific news predict low firm earnings and that stock prices briefly underreact to this information. Das and Chen (2007) examines stock-related message board posts, introduces techniques for extracting sentiment in the presence of negating words, and demonstrates that overall sentiment from these boards is linked to market prices, trading volumes, and volatility.

An important contribution by Loughran and McDonald (2011) shows that general-purpose dictionaries significantly misclassify common words in financial contexts—nearly three-quarters of words deemed negative by the Harvard Dictionary are not typically negative in financial applications. This insight led to the development of specialized financial word lists that leverage deep domain expertise through careful manual curation. While such expert-driven approaches benefit from accumulated practitioner knowledge, they may also introduce biases stemming from the curators' experiences and perspectives. Nevertheless, the Loughran-McDonald word lists have become a cornerstone resource in financial textual analysis, demonstrating the value of domain-specific expertise in methodology development.

Word lists have proven particularly valuable for measuring document features (ROC-Targeted objectives) and examining risk-related content. Hassan et al. (2019) use the method to construct firm-level political risk measures from earnings calls, while Jiang et al. (2024) employ NIST word lists to assess cybersecurity risk exposure through 10-K filings. The approach has also proven effective for ROC-Targeted objectives in corporate finance. Li et al. (2013) develop a simple yet effective measure of competition by counting mentions of competition-related words in 10-K filings, while Bourveau et al. (2020) examine cartel behavior by tracking specific product-related terms.

However, word lists have important limitations. They ignore context, cannot account for negation, and may misclassify words with multiple meanings. These limitations are particularly acute for ROC-Holistic objectives, where understanding the full context of discussion is crucial. Contemporary research increasingly favors more sophisticated methods that can better capture linguistic nuance and context, thus reducing measurement error, though word lists remain valuable as a transparent first-pass methodology for many applications.

The simplicity of word lists corresponds naturally to keyword search methods, which have become increasingly prevalent in corporate finance research as noted by Bowen et al. (2017). This basic framework continues to provide a foundation for more sophisticated methodologies while offering a useful baseline for validation of more complex approaches.

To implement these methods, many researchers rely on standard Python libraries such as `nltk` (for tokenization, stopword removal) and `re` (for advanced string operations). The `pandas` library helps with large-scale data manipulation. We also note that generative AI models including ChatGPT and Claude can, in fact, write the python code for the researcher. Such code tends to be accurate with few bugs when using careful and highly detailed prompting regarding what the code needs to accomplish.

4.4 Cosine Similarities

Cosine similarity is a method for computing the extent to which two spatial vectors are pointing in the same direction. The application of this method to compute similarity between documents has a long history in computer science (see Sebastiani (2002)) and dates back to Salton et al. (1975), who develop a vector space model for automatic indexing. In financial economics, it is the tool of choice for ROC-Comparative projects.

The method is flexible and can be implemented in many ways. We will discuss both rudimentary methods as well as more cutting-edge approaches using dimensionality reduction and embeddings. Although a complete understanding of these approaches is important, we note that more advanced methods based on dimensionality reduction are likely to become dominant going forward. This is likely because the cost of implementing these more advanced methods is not much higher than that for the more rudimentary methods, and the advantages of dimensionality reduction are significant.

All cosine similarity methods first require the researcher to transform documents into vectors V. Once the vectors are derived, they are normalized and similarity between any pair of documents in the corpus is assessed using the dot product of the two document vectors. We note three key approaches to deriving document vectors.

- **[Word-count Vectors]** Where N is the number of unique words in a corpus, this approach requires mapping each document in the corpus to a vector of length N. Intuitively, each element of the vector corresponds to one of the N words. For a given document, elements of its N-vector are populated with integers indicating the number of times each word appears in a given document. A binary variation is to populate N-vectors with dummies equal to one if the given word appears in the given document and a zero otherwise.

- **[tf-idf Vectors]** The tf-idf (Term Frequency-Inverse Document Frequency) method weights vectors to emphasize words that are highly prevalent in a document but less prevalent in the corpus in general. For a given document "d" and term "t", the vector is populated by the product of the total relative frequency multiplied by the inverse document frequency as follows:

$$\text{tfidf weights} = \frac{\text{Number of times term t appears in document d}}{\text{Total number of terms in document d}} \; x \qquad (1)$$
$$\frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

  The tf-idf weighting scheme is widely believed to improve signal strength as it reduces the impact of common words (which are often uninformative) while amplifying the importance of words that are highly represented in a given document relative to their overall usage in the corpus. This approach dates back to Salton et al. (1975), and support for the method was developed through experimentation rather than via theory. While the method performs well in most contexts, we note that it is not uniformly signal-improving as Hoberg and Phillips (2016) found a degradation of signal using tf-idf as compared to the binary approach for raw text vectors noted above regarding the specific task of industry classification. This suggests the simple appearance of an industry-identifying term is more informative than the overall or relative frequency.

- **[Embedding Vectors]** The final approach is to use dimensionality reduction to generate document vectors. Although the true dimensionality of the corpus is N, the number of unique words, embedding facilitates a mapping of each document to lower dimensional vectors $M < N$. The content of the M-vectors is determined automatically by the embedding algorithms, which take the documents as input. The key reason why this approach is so powerful relative to the above methods is that reducing dimensions allows highly related words such as "couch" and "sofa" to be scored as being very similar. This is not possible using raw document or tf-idf vectors (these rudimentary methods treat them as entirely unrelated). Although we explain embedding models fully in Section 4.8, we note that commonly used embedding models in the context of cosine similarities include doc2vec, BERT, RoBERTa or models related to ChatGPT.

Regardless of which method is used to construct document-specific vectors V above, we now discuss how to compute cosine similarities between any pair of documents. The vectors are first normalized to unit length. Cosine similarity between two documents $d1$ and $d2$ then is simply the dot product of the normalized vectors (V_{d1} and V_{d2}):

$$\text{cosine similarity} = \frac{V_{d1}}{\sqrt{V_{d1} \cdot V_{d1}}} \cdot \frac{V_{d2}}{\sqrt{V_{d2} \cdot V_{d2}}}. \tag{2}$$

A high cosine similarity indicates that the pair of documents were written using mostly the same vocabulary. A low value indicates that the documents contain unrelated content. Regardless of which method is applied, the cosine similarity approach has many excellent properties that motivate its widespread use.

In the context of our running example text described above, using raw word-count vectors, we would represent the example text as a sparse vector in $\mathbb{R}^N$, where N is the vocabulary size:

$$[\text{cloud} : 1, \text{computing} : 1, \text{solutions} : 2, \text{market} : 1, \ldots]$$

If another firm's 10-k were to mention cloud computing solutions as well, its cosine similarity with the example firm would be relatively high (closer to 1), while an annual report of a firm that uses different words to describe itself would have a low cosine similarity score (closer to zero).

We summarize this method's strengths below.

- **[Visualizations]** The normalization of the vectors to unit length gives that every document has a spatial location on a unit sphere, where vectors pointing in roughly the same direction are highly related documents. Vectors pointing in very different directions are little related. This permits intuitive visualizations of the data that can also benefit from clustering the resulting network observations.

- [**Low Correlation with Document Length**] The normalizations reduce the correlation between measured similarities and document length. Avoiding such correlations is a virtue among all methods as document length can confound inferences given that it has a complex interpretation.
- [**Bounded and Interpretable**] Cosine similarities are bounded in $[0, 1]$ and are not sensitive to outlier problems. They are intuitive to interpret as values near unity indicate perfectly similar documents and values near zero indicate entirely unrelated documents.

We also note the core benefits of using cutting-edge embeddings instead of raw document vectors or tf-idf weights.

- [**Synonym Recognition**] The raw document methods treat different words as being completely unrelated as they are represented in different elements of the document vectors. Hence a document discussing couches might have low similarity to a document discussing sofas. Embedding models are trained on large corpora and can recognize that terms such as couch and sofa are, in fact, very similar. Embedding models would score these documents as being highly similar.
- [**Words Appearing Close Together in the Corpus**] The raw document cosine similarities do not take word position within the document into consideration. For example, the terms "leather" and "sofa" would impact the similarity calculation in precisely the same way if they appeared side by side "leather sofa" or if they appeared far apart within a document. Typically, words appearing close to one another in a document provide context to a human reader, and this indicates additional meaning that will be entirely missed by basic raw document cosine similarities. This limitation is addressed quite profoundly by the more advanced embedding models as such is their primary objective.

Many research papers in financial economics use the cosine similarity method. Representative examples construct relatedness networks including Hoberg and Phillips (2016) on horizontal industry relatedness and Frésard et al. (2020) on vertical relatedness. Examples focused on document revisions in time series include Cohen et al. (2020b) (year-over-year revisions to 10-K sections) and Hanley and Hoberg (2012) (revisions to IPO prospectuses) and Hoberg et al. (2014) (compares each company's business description to the pace of aggregate dynamic evolution). Examples of specialized content analysis using cosine similarity include Kalmenovitz (2023) (regulatory intensity) and Florackis et al. (2023) (cybersecurity risk).

Python modules such as `scikit-learn`'s `pairwise` can help or one can compute dot products in `numpy`. `gensim` is also a popular choice for vector analysis requiring cosine similarities. For large corpora, sparse matrix operations in `scipy.sparse` can improve efficiency.

4.5 Precise Anchor Phrase Queries

The method of anchor-phrase is a large-step forward in addressing a key weakness of the word-list methodology and is most relevant for ROC-Targeted research projects. Key is that specific economic concepts such as financial constraints or hedging strategies are difficult to uniquely identify using single words or bigrams that typically appear in word lists. Moreover, the atomistic terms that appear in word lists often have many other meanings in their actual usage, which can add confounders. A key example based on Hoberg and Maksimovic (2015) is that the concept of financial constraints requires the joint appearance of at least two separate economic concepts: (1) the existence of an investment that a firm would like to fund and (2) evidence that the firm is having difficulties funding it. Building a word list around the term "investment" alone can confound inference because the list would load on discussions of investment strategies in general, making the result a measure of potential growth moreso than a measure of constraints. One can add terms like delay to the word list as an indicator of inability to conduct the investment, but then the word list would additionally load on other forms of delay such as supply chain disruptions.

Anchor phrase techniques were historically used by practitioners perhaps moreso than academics and its true beginnings are difficult to identify. Regarding the above example, its solution to avoid the confounder problem is to require that both concepts (investment and delay) have to appear in close proximity within a document (in the same paragraph or within 15 words). If, in addition, both concepts appear in a section of the 10-K discussing financial liquidity, the only meaningful interpretation of any "hits" is that a company is discussing its need to delay investments due to issues with financial liquidity (i.e., it is financially constrained).

The resulting quality of hits using this approach is excellent and false positives are rare when anchor phrase queries are well-constructed. However, significant care is needed from domain experts to construct the anchor phrase lists as they must be thorough. Not doing so can leave a significant mass of false negatives. One approach to mitigating false negatives is to first identify all terms that appear in proximity to core seed terms and sort by descending frequency. A domain expert would then build the anchor phrase lists from the most frequent terms by adding only those that are consistent with the desired interpretation. High speed query tools discussed below can make this process efficient and fast if budget permits, although custom coded Python scripts can also be used.

The implementation of the anchor-phrase method first requires the researcher to develop multiple word lists (typically two to five) depending on the sophistication of the economic concept being measured. Using the above example, one list would be a list of terms that uniquely indicate that forms of investment are being discussed such as "research and development" and "capital expenditures". A second list would be indications of delay such as "delay", "delaying", "postpone" or even "termination". Actual anchor phrase lists are longer than these simple examples.

Once the lists are complete, the researcher would run a query seeking "hits" where at least one word from each list appear close together within each document. Close appearance typically means that they appear within a narrow window of words (such as a 15 word window) or they appear in the same paragraph or even the same sentence. The width of the window can be calibrated using hand-checking a test sample or by using a known "ground truth" sample to observe which approach achieves the best balance between false positives and negatives. Window expansion should not go beyond a point where false positives increase rapidly. Generally, the window would be widened to reduce false negatives, and the anchor phrase query itself can be refined to cull false positives by adding more specificity or even adding an additional list of required terms to build a more specialized anchor phrase query. As a key researcher goal when using anchor phrase is precise identification, there is a general preference to tolerate false negatives moreso than false positives. The former mainly reduces power (and biases studies toward non-results) whereas the latter can lead to confounders as the false positives might bring different content into the measure that might matter through a different and unwanted channel.

Fortunately, NLP tools are available to researchers make running the anchor phrase queries fast and straightforward. Tools used in this research include, for example, the meta-Heuristica platform, the WRDS SEC Analytics package or custom Python programming techniques. These methods are also highly scalable, as for example Hoberg and Moon (2017) were able to run (in less than one day on a standard PC) thousands of anchor phrase queries over the entire corpus of 10-Ks, with queries run separately for every country and separately for multiple specific types of offshoring activities. The result is a dynamic network of U.S. public firm offshoring practices worldwide.

The anchor-phrase methodology has the following strengths and weaknesses.

- [**Strength**] The method can generate very precise hits that are easy to report with full transparency. Every hit can be extracted and can be reported in appendices, or they can be hand-scored for accuracy using research assistants. The reduction of false positives to near zero is a crucial strength that needs to be emphasized given the increased importance of identification in financial economics research. Perhaps the most core weakness of standard word lists is the proliferation of difficult to understand false positives that can permit economic confounders. Anchor phrase methods are specifically designed to reduce false positives to a very low level.
- [**Strength**] The method is flexible and can be extended or refined to obtain more specialized variations of the key theme being examined. In the above mentioned example of financial constraints, the anchor phrase method allowed the authors to develop novel measures of constraints that specifically indicate challenges raising equity capital or specifically regarding debt capital.
- [**Strength**] Anchor phrase queries are typically implemented as complex string searches. This method is fast given current computing power. Many gigabytes can be processed in minutes on the cloud or day (s) using custom coding.
- [**Weakness**] The biggest weakness is exposure to false negatives, which results in some degradation of signal strength. This can be mitigated by extending the anchor phrases to be increasingly complex and thorough to pick up missed hits that might use rare or less common vocabulary. A second approach to mitigate false negatives is the use of machine learning in a second pass (see Linn and Weagley (2019) who extend the anchor phrase measures from Hoberg and Maksimovic (2015)).

- [**Weakness**] Some specialized concepts are adequately complex to render the process of constructing anchor phrase queries too difficult. An intermediate approach could be to prompt a generative AI model to generate more appropriate queries. Alternatively, the concepts being measured might be "too similar" to another concept that is not desired in the measure, making the phrases susceptible to the more problematic false positives. The complexity issue is more common than the overly-similar concepts problem, but both are possible. When these issues are insurmountable, the best alternative is generative AI.

Overall, due to its high precision, high transparency and perfect replicability at moderate cost, it is likely that anchor phrase queries will continue to be used in future years. Yet as the cost of LLMs comes down and their transparency and replicability improves, we expect increased usage of generative AI to address the same class of research problems.

Many research papers in financial economics use the anchor-phrase method. Representative examples construct financial constraints (Hoberg and Maksimovic (2015) and Linn and Weagley (2019)), offshoring activities networks to study hedging and asset pricing (Hoberg and Moon (2017) and Hoberg and Moon (2019)), measures of contracting frictions (Frésard et al. (2020)), product life cycles (Hoberg and Maksimovic (2022)), and inflation exposure (Chava et al. (2022). A related approach by Bena et al. (2022) develops precise N-grams with only one interpretation to code patents as process-based innovation.

4.6 Text Regression

Text regression methods address a fundamental challenge in financial research: how to extract predictive signals from high-dimensional textual data while maintaining statistical validity. A key advantage of regression methods is their ability to control for observable confounding variables, allowing researchers to isolate the incremental information content in text.

For example, text regression transforms our running example into a regularized prediction problem:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

where x_{ij} represents normalized frequencies of words or phrases, such as:

$$\begin{bmatrix} x_{i,\text{cloud}} & x_{i,\text{growth}} & x_{i,\text{constraints}} & \cdots \end{bmatrix} = \begin{bmatrix} 0.02 & 0.04 & 0.02 & \cdots \end{bmatrix}$$

Text regression provides a direct statistical framework for linking textual content to economic outcomes, facilitating formal hypothesis testing and inference. The primary challenges lie in managing high dimensionality, potential overfitting, and the need for large training samples to achieve reliable parameter estimates.

Manela and Moreira (2017) illustrate these challenges in their effort to predict the VIX index using news text from *Wall Street Journal* front pages. With tens of thousands of unique words appearing in news articles but only hundreds or thousands of time periods with VIX observations, ordinary least squares regression would massively overfit the data, picking up spurious relationships that fail to generalize out of sample.

Regularization provides a solution to this high-dimensional regression problem. While Manela and Moreira (2017) employ Support Vector Regression (SVR), other regularization approaches like the Lasso serve a similar function. These methods add a penalty term to the regression objective that shrinks coefficient estimates toward zero, effectively reducing the model's complexity. The penalty helps avoid overfitting by forcing the model to focus on the most important predictive signals in the text. Mamaysky et al. (2023) find that using Lasso regressions is superior to an LLM (BERT) when forecasting credit spreads.

Taddy (2013) develops a more sophisticated approach specifically designed for textual analysis. His multinomial inverse regression framework employs independent Laplace priors with unknown variance for regression coefficients, resulting in the "gamma-lasso" estimation procedure. The method provides stable results for high-dimensional problems while maintaining computational efficiency. Taddy (2015) suggests Distributed Multinomial Regression be applied with distributed computing, making it practical for large-scale text analysis. García et al. (2023) apply this framework to develop finance-specific sentiment dictionaries, showing superior performance in predicting stock returns compared to traditional approaches.

Kelly et al. (2021a) further advance this line of work by developing the Hurdle Distributed Multinomial Regression model that addresses both the high dimensionality of text and the prevalence of zeros in word counts. Their framework explicitly models the two-stage process of word selection (whether a word appears at all) and word intensity (how frequently it appears when present). They demonstrate superior predictive performance compared to Support Vector Regression and to Distributed Multinomial Regression across various economic prediction tasks.

Perhaps the most important advantage of text regression methods is their ability to control for confounding explanatory variables. For example, one of the applications considered in Kelly et al. (2021a) involves forecasting macroeconomic outcomes (e.g. employment) using newspaper text. One could run a text regression of employment growth on news text. But that is not the most interesting question because other macroeconmic variables, including lagged employment growth, are readily available and may offer superior signals for forecasting. In fact, the newspaper coverage likely reflects this available macro data. A more interesting question is therefore if the news text is useful for forecasting over and above the macro data. Kelly et al. (2021a) show how to control for the macro data and find that WSJ text offers significant forecasting ability beyond the macro data itself.[5]

Text regression methods are particularly well-suited for ROC-Holistic research objectives, as they can effectively process all content in a corpus to predict or explain economic variables. They can also support ROC-Targeted objectives by estimating the relationship between document characteristics and outcomes of interest. In contrast, text regression methods are less applicable to ROC-Comparative objectives (which focus on document-to-document relationships).

`scikit-learn` provides `Lasso` (`linear_model.Lasso`) and Support Vector Regression (`svm.SVR`). R users often rely on `glmnet` for penalized regression. `lightgbm` or `xgboost` can also handle bag-of-words or embedding-based features effectively.

4.7 Thematic Topic Models

Thematic topic models represent one of the most influential approaches for ROC-holistic objectives in financial research. Rather than targeting specific content elements, these models

[5] Note that a two stage approach, where one first reduces the dimension of the text with another method and then controls for confounders in a second stage is unlikely to partial out the relationship between the confounders and the text, resulting in biased second stage estimates.

aim to decompose large text corpora into distinct thematic components that characterize the full scope of discourse. The field gained significant momentum with Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003), which models documents as probability distributions over topics, where each topic is itself a distribution over words. Steyvers and Griffiths (2007) provides a primer on efficient estimation of LDA using Gibbs sampling.

Applied to our running example, LDA might discover three latent topics such as:

$$\begin{aligned}
\text{Topic}_1 &= 0.4 \times [\text{cloud} : 0.3, \text{computing} : 0.2, \text{AI} : 0.2, \ldots] \\
\text{Topic}_2 &= 0.3 \times [\text{margins} : 0.3, \text{costs} : 0.2, \text{constraints} : 0.2, \ldots] \\
\text{Topic}_3 &= 0.3 \times [\text{market} : 0.3, \text{growth} : 0.2, \text{expansion} : 0.2, \ldots]
\end{aligned}$$

For example, the model might allocate 0.4 probability to Topic_1, which conditional on being mentioned, generates the word "cloud" with probability 0.3. Looking at the word probabilities for each topic, researchers usually proceed to manually label the topics. Here we might label Topic_1 as "technology" or "computing technology".

Topic models provide an unsupervised framework for discovering latent thematic structure while reducing dimensionality. However, the resulting topics can be unstable and often require significant parameter tuning and post-processing to achieve coherent, interpretable results.

In a recent example, Bybee et al. (2024) leverage the power of topic models for ROC-Holistic analysis in their study of business news. Without taking strong prior stances on which themes matter, they let the data guide topic discovery across Wall Street Journal articles. This approach reveals how various business themes evolve over time, with some topics—particularly those related to recessions—emerging as especially important for predicting economic activity and stock returns. Their methodology, which includes systematic procedures for selecting the optimal number of topics and online estimation to avoid look-ahead bias, provides a template for successful implementation.

Benson et al. (2024) build on this work and estimate that 80 percent of the variation in the excess bond premium can be explained by attention to the topics of Bybee et al. (2024), and that it can forecast macroeconomic fluctuations.

The asset pricing literature has particularly benefited from topic models' ability to capture multiple dimensions of risk and sentiment. Israelsen (2014) examines how risk disclosure topics relate to factor portfolios. Lopez-Lira (2023) uses topic models to analyze risk disclosures, showing their relevance for return prediction. Bybee et al. (2023) develop systematic risk factors from news text that help explain cross-sectional returns.

Topic models have proven especially valuable for understanding emerging risks and changes in the financial sector. Hanley and Hoberg (2019) combine topic modeling with word embeddings to identify and track emerging risks, showing how this framework can provide early warning of financial instability. Hoberg and Lewis (2017) use topic models to examine how fraudulent firms' disclosures differ from their peers, while Dyer et al. (2017) document the evolution of disclosure content in 10-K filings, revealing an increase in standardized "boilerplate" language over time.

Beyond risk analysis, topic models help illuminate various aspects of corporate behavior and communication. Ball et al. (2015) use them to study how firms discuss business changes and earnings quality. Bellstam et al. (2021) analyze analyst reports to measure corporate innovation, finding that topic-based innovation measures predict future firm performance and growth opportunities. In the regulatory sphere, Kalmenovitz et al. (2021) employ topic models to study regulatory fragmentation, while Lowry et al. (2020) examine how the SEC's regulatory process shapes IPO disclosures.

Recent methodological advances have enhanced topic models' utility. Bybee et al. (2024) adopt the online estimation approach of Hoffman et al. (2010) to address look-ahead bias by processing documents sequentially. Calomiris et al. (2024a) use rolling topic models estimated using network modularity to forecast energy market outcomes. Hybrid methods that combine topic models with embeddings or neural architectures improve performance while maintaining interpretability. These innovations particularly benefit applications in finance, where temporal concerns and semantic precision are crucial.

Topic models work best when documents have natural thematic structure. Risk factors in 10-Ks, for instance, often follow standardized patterns that facilitate topic discovery. The success of applications like Bybee et al. (2024) suggests that even more fluid discourse, such as news articles, often contains sufficient thematic structure for effective modeling.

Python libraries such as `gensim` (includes `LdaModel`) and `tomotopy` or the MALLET toolkit (Java-based but callable from Python/R) are common. For online or distributed training, `gensim` has built-in support to handle large-scale text.

Looking forward, topic models will likely remain central to financial text analysis, potentially in combination with newer methods like large language models. While LLMs offer sophisticated text understanding, topic models provide unique advantages in interpretability and their ability to extract clear factors from large document collections. Hirshleifer et al. (2025, ming) demonstrate this enduring value in their analysis of war discourse in newspapers, using topic models to extract systematic patterns in coverage that help predict asset returns. The sustained success of topic models across diverse applications in finance, from macroeconomic forecasting to corporate governance, underscores their importance as tools for holistic text analysis.

A common methodological weakness in topic modeling is that after the model identifies a given number of latent topics, researchers manually label these topics based on the most probable words or phrases in each topic. This approach can gain from the researcher's domain expertise, but it can also introduce biases into the analysis. Finding ways to automate topic labeling, perhaps with the use of LLMs, is a promising area for future research aiming to discipline these labeling choices.

Another potential weakness of LDA in particular is that there is often more than one unique set of latent topic representations of a given corpus. Ke et al. (2024) argues that LDA parameters are not identified, which suggests that the choice of prior matters.

4.8 Embeddings

Word and document embeddings represent a fundamental advancement in NLP technology by enabling dense numerical representations that capture semantic relationships in text. Unlike bag-of-words approaches or topic models that treat words as independent units, embeddings preserve context and meaning by mapping words and documents into a continuous vector space where similar concepts are positioned near each other. The seminal work of Mikolov et al. (2013) introduced the influential word2vec model, establishing core principles that subsequent embedding technologies build upon.

Modern embedding approaches span multiple generations of increasing sophistication.

Early models like word2vec and doc2vec focused on learning static word representations through neural network training on large text corpora. In these models, each word receives a fixed vector representation regardless of context - for example, “bank” would have the same embedding whether referring to a financial institution or a river bank. More advanced transformer-based models like BERT (Devlin et al. (2018)) and RoBERTa (Liu et al. (2019)) subsequently enabled contextual embeddings that dynamically represent words based on their usage in a sentence. These models can disambiguate multiple word senses and capture subtle variations in meaning - the same word “bank” would receive different vector representations depending on surrounding context. Most recently, embedding capabilities within large language models like GPT have further pushed the frontier of semantic representation. While large language models like GPT can also produce embeddings as a byproduct of their architecture, their primary purpose and training objectives are fundamentally different from traditional embedding models. We discuss these generative AI models and their applications in the next section.

Returning to our running example, if we denote the embedding function as $\phi(\cdot)$, we obtain:

$$\phi(\text{“cloud computing”}) \approx \phi(\text{“digital technology”}) \tag{3}$$

$$\phi(\text{“market position”}) \approx \phi(\text{“competitive standing”}) \tag{4}$$

The document itself is typically represented as an aggregation of its constituent embeddings:

$$\vec{d} = f(\phi(w_1), \phi(w_2), ..., \phi(w_n)) \in \mathbb{R}^k$$

where k is typically between 100 and 2000, and $f(\cdot)$ might be a simple average or a more sophisticated aggregation function.

Embedding models capture rich semantic relationships and contextual dependencies while enabling sophisticated downstream tasks through dense vector representations. Their primary limitations stem from the opacity of the learned representations and potential temporal instability in the embedding space, particularly when applied to historical financial documents.

Embeddings are particularly valuable for ROC-Comparative objectives as they enable more nuanced document similarity calculations compared to traditional cosine measures. For example, Hoberg and Phillips (2021) demonstrates that doc2vec embeddings improve the measurement of product market relationships between firms by 20-25% compared to earlier text-based methods. Sentence transformers such as those developed by Reimers and Gurevych (2019) are another popular approach.[6] The enhanced performance stems from embeddings' ability to recognize semantic relationships even when different words are used to express similar concepts.

For ROC-Holistic objectives, embeddings serve as powerful building blocks for downstream analysis. Cong et al. (2019) develop an innovative approach combining embeddings with clustering to extract interpretable factors from financial text. This method addresses a key limitation of pure topic models by incorporating word context while maintaining interpretability. Similarly, Hanley and Hoberg (2019) combine word2vec embeddings with topic modeling to track emerging risks in the financial sector with greater sensitivity than traditional approaches. Lv (2024) classifies sentence embeddings into a set of researcher-defined topics, and then aggregates them to the document level to ensure that its final output is interpretable. Sarkar (2024a) uses embeddings of financial news to measure how the market's perception of firms evolves over time and examines how these changes relate to asset prices and analyst forecasts.

The use of embeddings for ROC-Targeted tasks has also grown, often through hybrid approaches. Several studies demonstrate success using expertly-curated seed words combined with embedding-based expansion to identify specific concepts. For instance, Chen et al. (2022) employs this technique to measure institutional participation in information production, while Wu (2024) applies it to assess supply chain risk exposure. More sophisticated applications leverage contextual embeddings to measure specific constructs: Jha et al. (2022) utilize BERT to embed mentions of "finance" from millions of books, enabling precise measurement of societal attitudes toward finance across centuries and countries. In asset pricing, Jiang et al. (2022) harness contextual embeddings to identify return-predictive content in financial news with greater precision than traditional approaches. These applications demonstrate how embeddings can enhance measurement of specific constructs while maintaining the interpretability advantages of targeted approaches.

[6] Available at https://sbert.net and on Hugging Face.

Recent research increasingly utilizes embeddings from large language models, though care must be taken regarding temporal biases when analyzing historical text (Sarkar and Vafa (2024)). For many applications, simpler embedding models like word2vec may be preferable as they can be trained separately for different time periods, avoiding look-ahead bias while still capturing much of the semantic benefit that embeddings provide.

Key implementation considerations include choosing between static and contextual embeddings, selecting appropriate training corpora, and handling out-of-vocabulary terms. While more sophisticated embedding models generally offer superior performance, they also require greater computational resources and technical expertise to implement effectively. The choice of embedding approach should thus be guided by both the specific research objectives and practical constraints.

For static embeddings (Word2Vec, Doc2Vec), `gensim` is the most popular Python library. For contextual embeddings (BERT, RoBERTa, GPT-based), `Hugging Face Transformers` is standard, with pretrained or fine-tuned checkpoints. The Hugging Face community tends to incorporate innovations in this space quickly.

Looking forward, embedding technologies continue to evolve rapidly, with newer models offering increasingly nuanced semantic representations. However, the core benefit of capturing meaning through geometric relationships in a continuous vector space remains central to their utility across all ROC categories. This fundamental capability explains why embeddings have become essential components in many modern NLP applications in finance, often used in combination with other techniques to achieve optimal results.

4.9 Generative AI

Generative AI models, also termed "prompt-based large language models" and perhaps most accurately described as "instruction-tuned language models," represent a significant methodological advancement in natural language processing. These models are characterized by their ability to perform diverse analytical tasks through natural language instructions rather than task-specific architectural modifications. Although the terminology varies across academic and industrial contexts, the core innovation lies in the models' capability to interpret and execute complex analytical instructions expressed in natural language.

In the context of our running example text, we might prompt an LLM with:

```
Identify key business risks and growth opportunities in this text, providing specific evidence for each.
```

The model may generate structured analysis such as:

```
RISKS:

Supply chain constraints affecting margins
Rising labor costs in semiconductor manufacturing
Macroeconomic uncertainties (interest rates, geopolitics)

OPPORTUNITIES:

Cloud computing and AI-enabled solutions
Cybersecurity market expansion via SecureNet acquisition
Southeast Asian market penetration
```

Generative AI models offer unprecedented flexibility in analyzing complex narratives and can perform sophisticated reasoning tasks. However, their deployment in financial research faces challenges related to reproducibility, computational costs, and potential hallucination, particularly when processing documents that predate or postdate their training data. Prompt engineering remains somewhat of an art but should be more straightforward when the research question is narrow and clear.

Researchers with ROC-targeted can either directly prompt a generative AI model for an answer. Doing so is becoming easier with models which support structured output (e.g. JSON responses). ROC-comparative projects can use the embedding ability of generative AI models to compare texts. They can also prompt the model to point out or score differences and similarities between two documents.

The theoretical foundations of instruction tuning derive from the transformer architecture (Vaswani et al. (2017)), augmented by substantial innovations in pre-training objectives and instruction optimization. Notable early implementations such as GPT-3 (Brown et al. (2020)) demonstrate emergent capabilities through scale, achieving strong performance on numerous tasks without task-specific architectural modifications. While domain-specific models such as BloombergGPT (Wu et al. (2023)) have been developed for financial applications, empirical evidence suggests that general-purpose models like GPT-4 achieve comparable performance on financial tasks, challenging assumptions about the necessity of domain-specific training.

It is worth noting that the label "large" in "large language model" is inherently relative and temporary. When BERT was introduced in 2018 with 340M parameters, it was considered remarkably large compared to previous models. By 2019, GPT-2's 1.5B parameters dwarfed BERT, only to be overshadowed by GPT-3's 175B parameters in 2020. As of this writing in 2025, models with trillions of parameters exist, and we expect that what we currently consider "large" will seem modest by the time this review is published. This rapid scaling of model size reflects both technological advances in computing infrastructure and empirical evidence that larger models often exhibit qualitatively different capabilities. However, the fundamental distinction between traditional embedding models and generative AI models lies not in their size but in their training objectives, rich training data, and intended use cases.

Empirical evidence suggests significant analytical advantages in financial applications:

First, these models demonstrate sophisticated contextual understanding in financial analysis. Jha et al. (2024) document the extraction of forward-looking signals regarding capital expenditure from earnings call transcripts, with predictive power extending to nine quarters. Kim et al. (2024) establish performance parity with professional analysts in financial statement interpretation, notably achieving this without domain-specific architectural modifications.

Second, generative AI models facilitate novel methodological approaches to traditional financial research questions. Lopez-Lira and Tang (2023) demonstrate systematic extraction of market-relevant information from news content through carefully constructed instruction sets, while Fedyk et al. (2024) establish these models' capacity to capture demographic heterogeneity in investment preferences.

The application of these models in financial research necessitates careful attention to several critical methodological challenges:

- **Look-ahead Bias:** The pre-training process may inadvertently incorporate future information into historical analyses, thereby skewing results. To address this, Glasserman and Lin (2023) and Sarkar and Vafa (2024) propose rigorous frameworks for isolating temporal information, including systematic masking protocols. Engelberg et al. (2025) propose entity neutering, using the LLM itself to remove all identifying information from text, before using the text for prediction. Levy (2024) and Lopez-Lira et al. (2025) argue that generative AI models are so good at memorizing economic and financial data, and in particular numeric data, that earnings forecasts that include memorized tokens can be misleadingly accurate. It is important to note that simply validating results on data beyond the model's "knowledge cutoff date" is insufficient. This approach is problematic because many closed models like ChatGPT and Claude undergo subsequent fine-tuning based on human feedback collected after the initial training period, particularly through reinforcement learning from human feedback (RLHF) (Ludwig et al., 2025).

A nascent literature has also begun to offer more holistic solutions. A natural approach to avoid look-ahead bias is to pretrain these model from scratch only on historical data that predates the predicted period. While this idea is straightforward, implementing it requires efficient use and curation of historical timestamped data, so that the resulting models exhibit sufficient intelligence. Sarkar (2024b) pretrains timestamped BERT models with knowledge cutoffs between 1900 and 1963 (StoriesLM). He et al. (2025) pretrain what they call "chronologically consistent" LLMs (ChronoBERT and ChronoGPT) timestamped with the availability date of their training data (1999–2024), yet accurate enough that their performance is comparable to much larger open-weight models like Llama 3.1. By comparing these unbiased models to potentially biased Llama, they find that look-ahead bias is modest is in their setting of predicting stock returns with news. They note, however, that look-ahead bias is application-specific. Fortunately, these chronologically consistent models are publicly available so researchers can easily assess look-ahead bias in other settings.

- **Methodological Replicability:** Variability in outputs across different model versions and instruction sets presents a significant challenge for research replication. This underscores the need for standardized protocols governing both the construction of prompts and the deployment of models. At a minimum, researchers should eliminate randomness in text generation by setting parameters such as "temperature" (for OpenAI models) to zero, ensuring consistent and reproducible outputs. The use of open-source models such as Meta's Llama is another option.

- **Resource Allocation:** The implementation of these models requires substantial computational infrastructure. However, they offer the potential to streamline research workflows by reducing reliance on traditional data collection methods, thereby improving efficiency in financial analysis.

- **Hallucination:** Early models like GPT-3.5 tend to generate plausible-sounding answers, which are simply not true. For example, when asked for a source for the claim that "Hallucination is rampant in LLM use in finance," the model may generate a citation of "Hoberg, Gerard, and Asaf Manela, 2022, Hallucination and LLMs, Journal of Rigorous Finance, Vol 5, Issue 1, pp. 1-24," even though such a paper does not exist. Hallucination is much less of a problem with newer models, especially ones which use search tools to augment the model with knowledge it lacks. But being mindful of this possibility is important for a researcher who uses generative AI in their work.

An important methodological consideration when employing these models is their training objective: optimization for human approval rather than objective accuracy. The models are fundamentally trained to generate responses that human evaluators find satisfactory and convincing. This training paradigm introduces potential methodological concerns, particularly in analytical contexts. When prompted to explain their reasoning or justify their outputs, these models may generate plausible-sounding explanations that align with evaluator expectations rather than reflect their actual computational process. This characteristic necessitates careful validation protocols and suggests that model-generated explanations should be treated as hypotheses rather than definitive analytical insights.

The methodological implications of generative AI models extend beyond immediate analytical applications. Kazinnik (2023) employs a large language model to simulate depositor behavior during a bank run, using generative AI to create synthetic data that aligns with empirical studies. This approach highlights the potential of generative models to provide insights into human behavior in economic contexts, offering a cost-effective method for analyzing hypothetical scenarios and informing policy communication strategies. Novy-Marx and Velikov (2024) raise important considerations regarding research methodology, particularly concerning the automation of hypothesis generation and testing procedures. This suggests a need for careful integration of these models into existing research frameworks rather than wholesale methodological replacement.

Researchers can access large language models like GPT-4 or Claude via APIs (`openai` or `anthropic` Python clients). For open-source models (Llama, Bloom, etc.), `Hugging Face Transformers` offers inference pipelines (`pipeline(‘‘text-generation’’)`).

Looking forward, generative AI presents significant opportunities for methodological innovation in financial research, particularly in tasks requiring sophisticated contextual reasoning. In fact, Eisfeldt and Schubert (2024) already find that the development and adoption of generative AI is driving a significant technological shift for financial research (and firms). However, optimal implementation will require careful consideration of methodological rigor and systematic validation protocols.

4.10 Named Entity Recognition

Although the potential of named entity recognition (NER) for solving problems in finance is compelling, the use of this technology in finance remains somewhat nascent. This is likely due in part to the fact that NER technologies have only recently increased in quality enough to materially improve researcher efficiency.

An NER system may decompose the text in our running example into structured entity annotations:

$$\begin{aligned}
\text{MONETARY} &= \text{\$850 million}, \text{\$275 million}, \text{\$2.1 billion} \\
\text{ORGANIZATION} &= \text{SecureNet Technologies} \\
\text{LOCATION} &= \text{Southeast Asia} \\
\text{DATE} &= \text{fiscal } 2024, 2023
\end{aligned}$$

Named Entity Recognition provides precise extraction of structured information from unstructured text, facilitating quantitative analysis of specific entity types. However, its utility is constrained by predefined entity categories and potential difficulties in capturing complex relationships between identified entities. The method is most effective when combined with other approaches that can leverage the extracted structured information within a broader analytical framework.

One step in improving efficiency is to develop methods and share code, as is done by Shah et al. (2023) and Hilt and Schwenkler (2024) for example. The former establishes a new financial database that can be used to evaluate NER algorithms and identifies Snorkel platforms as the best performing. The latter uses the `spaCy` NER algorithm available in the `cleanNLP` library in R. Hilt and Schwenkler (2024) uses NER in newspaper articles to develop a dynamic network of news-related firms. This is in line with the standard goal in financial economics to identify company names as entities. The study illustrates the ability of the resulting network to identify periods of elevated risk. Using code from established researchers is part of "good practice" as it comes with documented performance metrics and established uses.

Another excellent use case is the identification of company name mentions in trade association websites. Hoberg and Neretina (2023) uses the Flair SequenceTagger model implemented in Python to achieve this task (see Akbik et al. (2019)). Interestingly given the use of newspapers above, Flair is trained using the CoNLL-2003 named entity recognition dataset (Tjong Kim Sang and De Meulder (2003)) which contains many Reuters newspaper articles. The NER procedures, supplemented by human review in this case, identifies which companies are likely members of which trade associations using websites. The authors then study the impact of trade association membership on a wide array of corporate outcomes. Ferrell et al. (2021) is another example of an implementation using similar NER technology, this time to detect lawyer and employee name mentions on law firm websites.

Many earlier studies use fuzzy matching to identify names. The methods noted here are more advanced versions of this approach, and perhaps most notable, the more advanced tools have the ability to identify named entities in free form text such as newspapers or corporate websites. Notwithstanding the technological advances, NER technology remains a technique where we expect that future research will continue to provide improvements. As of present day, for example, even these more advanced techniques fall short of a 90% success rate. As a result, using more advanced techniques combined with human input and quality control is a potential best practice.

Finally we note that NER techniques have also been used to identify entities other than companies. For example, Cohen and Nguyen (2024) identify product, money and percent entities in conference call transcripts. This allows the authors to identify and assess the consequences of firms changing their performance targets, where they then find negative abnormal stock returns.

4.11 Visual and Voice

Recent advances in machine learning have enabled sophisticated analysis of non-textual data sources including voice recordings and visual content. These modalities offer unique insights into financial markets and corporate behavior that complement traditional textual analysis.

Voice analysis has emerged as a powerful tool for extracting information from earnings calls and other corporate communications. Mayew and Venkatachalam (2012) pioneered this area by analyzing vocal markers of managerial emotional states during earnings calls, finding that positive emotional markers predict future firm performance. Building on this work, Hobson et al. (2012) demonstrate that vocal cues can help detect financial misreporting, suggesting that voice analysis may capture deceptive behavior that is not apparent in the transcript alone. More recently, Gorodnichenko et al. (2023) extend voice analysis to monetary policy by examining Federal Reserve communications, showing how vocal features of central banker speeches provide additional signals about policy intentions beyond their written content.

Visual analysis represents another frontier in financial machine learning. Hu and Ma (2021) analyze video recordings of management presentations to understand how visual cues influence investor decision-making. Their work reveals that non-verbal communication channels significantly impact investment choices, even when controlling for the information content of the presentations. On a larger scale, Mukherjee et al. (2021) demonstrate how satellite imagery can be used to generate novel macroeconomic indicators, highlighting the potential for visual data to provide real-time insights into economic activity. Curti and Kazinnik (2023) apply facial recognition analysis to FOMC press conference videos, quantifying the impact of the Fed Chair's facial expressions on investor expectations. They find that negative expressions can adversely affect market reactions, even when accounting for verbal communication, underscoring the importance of nonverbal cues in financial markets.

Guenzel et al. (2025) employs AI to deduce personality traits from LinkedIn photos of MBA graduates, finding that these traits predict various career outcomes, including school rank, compensation, and job seniority. While the AI-derived traits show only modest correlations with cognitive measures like GPA and standardized test scores, they offer comparable predictive power for labor outcomes.

The combination of multiple modalities - text, voice, and visual - can be particularly powerful. Ludwig and Mullainathan (2024) show how integrating different data types can improve hypothesis generation in financial research. Their approach demonstrates that cross-modal analysis can reveal patterns and relationships that might not be apparent when examining each modality in isolation.

Key implementation challenges in voice and visual analysis include:

- Data availability and quality, particularly for historical recordings
- Computational requirements for processing large-scale video and audio data
- Need for specialized expertise in audio and visual processing techniques
- Potential biases in visual and vocal features across different demographic groups

The libraries `pyAudioAnalysis` or `librosa` in Python can assist in voice/emotional speech analysis; `OpenCV` or `Pillow` handle image processing. Modern solutions often combine these with deep-learning frameworks (`PyTorch`, `TensorFlow`) for multimodal NLP tasks.

Looking forward, the integration of voice and visual analysis with traditional text-based approaches promises to provide increasingly nuanced understanding of financial markets and corporate behavior. As computational capabilities continue to advance, these modalities will likely become standard components of the financial researcher's toolkit.

4.12 Other

There exist many specialized NLP-based methods that are often used to measure specific document features not discussed above, or to improve accuracy when processing corpora. Although this list is not comprehensive, we highlight a number of select examples. We note that many of these methods are under-utilized in the study of financial economics and further exploration might be particularly fruitful. They are also well-suited to ROC-Targeted research objectives.

- [**Part of Speech Tagging**] Many important concepts in financial economics are more precisely identified when conditioning on parts of speech. Verbs indicate actions, which might be relevant for separating speculative content from definitive acts. Nouns indicate specificity regarding things or entities, which can be important when measuring items needed to construct a precise list. For example, Cabezon and Hoberg (2022) identifies a list of hundreds of new technologies that emerged since 1997 by identifying noun-phrases using the metaHeuristica platform appearing in paragraphs where the word root for technology is present. This technology facilitated efficient identification of emerging technologies even though some technologies are best expressed as bigrams, trigrams or unigrams. Accuracy was further ensured with help from research assistants as the technology made manual review of each term manageable given the high initial dimensionality of the corpus. This facilitated analysis of how technologies transmit in network settings such as through board connections to improve identification of a research conclusion.

- [**Dependency Parsing**] Dependency parsing goes beyond parts of speech and attempts to codify the relationships between words appearing in proximity. For example, from the perspective of the verb "runs" in "She runs fast", "She" is the subject and fast is an adverb modifier. Cao et al. (2023b) use techniques developed by Chen and Manning (2014) as part of a larger framework to identify the extent to which portfolio managers discuss forward looking risk assessments in narrative discussions. The added sophistication of this approach is helpful given the nuanced way managers discuss forward looking risks.

- [**Boilerplate Content**] Major corpora used in financial economics often contain significant boilerplate content, i.e., content that serves a more general purpose for the document writer (such as style requirements or required legal statements) and that lacks an important economic signal. Some research projects can benefit from purging boilerplate content to improve signal strength, whereas other projects aim to measure and study boilerplate content itself. Work in this area is growing but remains quite nascent, and more work remains fruitful. An example of a study that purges boilerplate content from the complex corpus of company websites is Hoberg et al. (2024), and a study that measures the extent of boilerplate content is Lang and Stice-Lawrence (2015). The former study uses ground truth regressions to purge dimensions in embeddings that are uninformative about an industry signal. The second uses 4-grams to detect boilerplate content as highly repeated phrases relative to a benchmark.

- [**Plagiarism Detection**] The technology used to identify plagiarism has advanced significantly, and has experienced growing use in evaluating student assignments. A strength is the ability to identify content that is very similar (even nearly identical) in free form text even though the content might not be exactly the same and copied material might have heterogeneous length. Yet higher-end technologies might not be scalable. Cagé et al. (2020) run large-scale analysis by tracking small portions of text that appear verbatim in news articles released within certain time periods to identify original versus copied news. The paper uses 5-grams to implement the research and finds that a quarter of news stories in their sample are reproduced online within 4 minutes.

  An interesting dichotomy is that 4-grams were used to identify boilerplate content in Lang and Stice-Lawrence (2015) and 5-grams to measure copied content in Cagé et al. (2020). These uses might initially appear to be in conflict. Yet we view both articles to be well-implemented, and the success of both is explained by context. Content that is pervasively repeated in regulatory settings is a strong indication of boilerplate content, and repeated content in articles released in short time-windows in the media is an indication of copied content as boilerplate is less prevalent in the media.

- [**Readability**] Readability and related metrics assess the extent to which a reader can understand the contents of a given document. Li (2008) is an early study that explored annual report readability and its link to current earnings and earnings persistence. This study explored early measures from computational linguistics including the Gunning Fog Index and the Flesch-Kincaid formula. Loughran and McDonald (2014) explore readability along dimensions relevant to financial economists and find that the early measures do not work well for financial applications and that document length is superior. Bonsall et al. (2017) take a different approach and consider a plain English measure of financial reporting readability, the "Bog Index". This existing work lays an important foundation for assessing readability, an issue relevant to providing transparent information that is equitable across heterogeneous economic agents. This is also an area where new technologies might offer significant potential for additional gains in measurement quality and research in the area thus remains fertile.

- [**Numeracy**] Numeracy in computer science is the study of using numbers in the context of computational problems. This area is particularly nascent regarding uses in financial economics. Yet high quality numeracy likely has outstanding potential to improve research in this area. Important corpora used by financial economists contain both text and numbers, and too little research has been able to integrate both into impactful research agendas. Developing accurate methodologies with interpretation in this space is a potentially fruitful area for future research (Thawani et al., 2021).

- [**Unusualness**] Text is unusual if it is unlikely to appear in a random sample of text. Thus measuring unusualness requires a probabilistic language model. Entropy measures of unusualness have been studied by Glasserman and Mamaysky (2019), who find that unusually positive news forecasts lower volatility. Glasserman et al. (2023) extends their work using an LSTM architecture.

5 Corporate Finance

We briefly summarize key contributions to the corporate finance literature that were facilitated by novel technologies and NLP tools. This summary is not meant to be a comprehensive review of all NLP-based contributions in corporate finance, as that is not possible with available page counts. Rather we select highly targeted discussions that indicate how major areas arise with these new technologies, and to provide examples of our above-mentioned framework of research objectives. We also note key visions for how technologies can more dramatically accelerate research progress and address new questions not possible before.

5.1 Disclosure, Regulation and Issuance

Many of the earliest contributions in corporate finance focused on disclosure and information environments. This was necessary as researchers must first understand the information content of large textual corpora before they can apply them to address the wide range of theoretical questions addressed thereafter. Early researchers documented both content and how it changes dynamically over time. Hanley and Hoberg (2010) examines the content of

IPO prospectuses using cosine similarities and illustrates their strong link to IPO pricing, and Ball et al. (2015) provides a related analysis for 10-Ks using more advanced topic model technology (LDA). Brown and Tucker (2011) explored the rate of change in the content of MD&A sections of the 10-K over time and document a strong link to changes in fundamentals driving changes in disclosure, although Dyer et al. (2017) later show using topic models that boilerplate content has also been increasing. These studies are prime examples of ROC-comparative and ROC-holistic objectives as some are based on large numbers of document comparisons and others on assessing information content of a corpus holistically.

Loughran and McDonald (2011) establish the use of expertly curated word lists to identify sentiment from a financial interpretation context. Sentiment is perhaps the most heavily researched feature of content, and the authors document sentiment's relevance to economic issues such as fraud, material weaknesses, and risk. Li (2008) and Loughran and McDonald (2014) analogously assess document readability. The former illustrates that the fog index correlates with earnings and its persistence, and the latter questions the fog index and notes that document length might be more relevant. Bonsall et al. (2017) then propose a new approach based on grammatical attributes of the content. These are foundational examples of ROC-Targeted objectives as sentiment, readability, length, and grammar are important document features.

It is noteworthy that literature in this area often spans both corporate finance and accounting, as the information environment is crucial to issues of interest to both disciplines such as understanding the quality of earnings, the prevalence of fraud, and the ability of firms to raise capital and avoid litigation. Hanley and Hoberg (2012) illustrate the role of disclosure in reducing litigation risk in IPOs, and along with Brown and Tucker (2011) and Cohen et al. (2020b) (who examine slow market reactions to changes in disclosure), are great examples of how vector representations can be used to model the extent to which documents are revised. We note that document revision is a core ROC-comparative objective. These are great examples illustrating how document revisions can inform important economic outcomes using ROC-comparative methods.

As the quality of NLP tools improved, much research sought to examine overall informativeness of textual disclosures. Ball et al. (2015) use LDA on MD&A sections of the 10-K and document a strong ability to explain valuations. Dyer et al. (2017) use topic models to revisit how overall disclosure quality has been changing over time and find evidence of increased boilerplate content. Finally, Hoberg and Lewis (2017) use topic models to identify a set of topics that most predict which firms commit fraud using SEC filings.

As much disclosure is regulated, additional studies explored the efficacy of regulation in ensuring high quality disclosures. Lang and Stice-Lawrence (2015) document how the regulatory process achieves improved disclosure quality internationally, and Lowry et al. (2020) examines the mechanism of SEC comment letters in detail. Kalmenovitz and Chen (2020) and Kalmenovitz (2023) measure regulatory impact with the former focusing on regulatory intensity and the latter on regulatory fragmentation. Calomiris et al. (2024b) show that higher regulation leads to slower firm growth and lower profits, with especially large impact for smaller firms. Bourveau et al. (2020) suggest that increased regulatory disclosure is not always helpful as it can also facilitate bad behavior relevant to regulators such as collusion. Finally, Cao et al. (2023a) notes that firms might alter disclosures given the knowledge that it will be processed not only by humans but also by machines.

5.2 Industrial Organization

This section explores the impact of NLP technologies on the study of industrial organization, with a focus on industry competition, complementary product market relationships, vertical relationships, and economies of scope. This area is noteworthy due to its early adoption of NLP methods and the high impact of work in this area. A key driver of impact is the widespread need to model the concept of industry in business research both as a core research topic or simply to control for industry effects, which are known to be first order.

A second foundation for its success is the quality and availability of textual data. There is a large amount of text focused on product markets such as item 1 business descriptions reported annually to the SEC by publicly traded firms. These filings are mandatory (hence no selection biases) and fully cover U.S. public firms going back to the late 1980s (high power through long time series). Prior to the use of NLP, the main source of industry data was government classifications such as SIC and NAICS, which (A) are not updated frequently even though firms and industries are highly dynamic, (B) they lack flexibility regarding partial relatedness, and (C) they are sourced through a manual process and have limited signal strength. Beyond SEC filings, additional research extracts industry information using websites to cover private firms, short business descriptions from Refinitiv to cover foreign firms, and various international regulatory databases to extend classifications.

Hoberg and Phillips (2016) and Hoberg and Phillips (2010) first introduced NLP methods to this area by constructing the dynamic text-based industry classification (TNIC) database using cosine similarities between 10-K item 1 filings. This ROC-comparative framework was improved by Hoberg and Phillips (2021) using doc2vec embeddings, resulting in a 20-25% improvement in signal strength. Despite the existence of larger language models, doc2vec (unlike other embedding algorithms) has the advantage of being easily trained separately in each year. This ensures that the resulting industry classifications are not exposed to look-ahead biases that might arise using models trained on recent data such as GPT. Given that industry classifications are so widely used in both corporate finance and asset pricing settings, care to avoid look-ahead issues is particularly important in this area as it avoids unintended negative externalities on other researchers using generated industry data.

The growing array of textual industry data has been expanded beyond public U.S. firms in important ways. Hoberg et al. (2024) extend the database to include up to one million private firms using corporate websites from 2000 to present, Frésard et al. (2020) extends the framework to include vertical relationships, and Breitung and Müller (2023) extends the framework to include international competitors using short Refinitiv business descriptions and a GPT based embedding that performs well even on these shorter descriptions. These authors note that masking techniques should significantly mitigate the look ahead issues associated with GPT. Hoberg et al. (2014) provide measures of dynamic product market fluidity from the TNIC spatial database. All of these referenced studies are squarely ROC-comparative research projects, and they indicate a vast new resource of spatial modeling of business relationships that can address many important theoretical questions in a dynamic setting.

Additional studies have used text-based industry classifications in structural models to study corporate profits and markups over time (Pellegrino, Bruno (2022)). Others have used related data resources such as 8-K filings to study supply chain risk (Wu (2024)), others study supply chain inflation using text (Chava et al. (2022)), and others use string-search methods to obtain direct mentions of competitive influences (Li et al. (2013)). Beyond creating measures, text-based industry data has enriched studies of how U.S. firms deal with competition from China (Hombert and Matray (2018)), invest following peer stock price changes (Foucault and Fresard (2014)), and the study of trade credits and supplier competition (Chod et al. (2019)).

5.3 Investments and Financial Constraints

The literature using new technologies to examine investment and financial constraints is large and growing. This is due in part to a growing realization that NLP tools can measure many theoretically important concepts that were not possible to measure previously. Other studies focus on improving measurement of existing concepts or studying the role of artificial intelligence itself in corporate outcomes. The research objectives in this area weight heavily on ROC-Targeted and ROC-Comparative objectives with a key goal of measuring concepts not possible using previously available researcher tools.

5.3.1 Corporate Investments and Acquisitions

Hoberg and Phillips (2010) is one of the earliest NLP studies in this area, illustrating the compelling nature of measuring concepts not previously measurable. A key gap in the prior literature is the absence of direct evidence of product market synergies as a key merger motive. This study uses basic cosine similarities from the above-mentioned TNIC industry representations to show that acquisitions are significantly more common among firms positioned as complements (high similarity but not too high) rather than substitutes (very high similarity) in the product market. The study also documents that the size of textual business descriptions grows as product synergies are realized. Extending this line of work, Frésard et al. (2020) document vertical integration synergies relating to commercialization of mature innovation but not earlier stage innovation. This study brought data on vertically related vocabularies from the Bureau of Economic Analysis to the TNIC platform. Bernile and Lyandres (2019) extend this literature in a different way and use text searches from newspaper articles to extract forecasted operational synergies and cost cutting. Unlike the above ROC-comparative studies, this study is in the ROC-targeted category as it seeks highly specialized content from its corpus and is not based on document comparisons.

Beyond acquisitions, studies have explored investments more broadly along with its determinants. As an illustration of the potential of cutting-edge measurement, Jha et al. (2024) use ChatGPT to extract information from conference calls that predicts future capital expenditures for up to 9 quarters. Baker et al. (2016) measure economic policy uncertainty from newspapers using anchor phrase methodologies that identify passages containing words from each of three groups: uncertainty, economy, and policy vocabularies. Hassan et al. (2019) identify related measures that are firm specific in conference calls using bigrams that uniquely associate with discussions of politics that appear in proximity to risk words. This is another example of anchor-phrase methodology applied on bigram lists, in this case curated using pattern-based sequence identification. The authors find that firms more exposed to economic policy uncertainty experience increased risk and reduced investment. These studies are firmly in the ROC-targeted category.

5.3.2 Financial Constraints

The study of financial constraints is seminal in corporate finance, and dates back at least to Kaplan and Zingales (1997). This early study pre-dates the advent of NLP tools in finance, and the authors manually reviewed a small number of 10-Ks to construct an index based on a regression model. Though pathbreaking at the time, contrasting this approach with modern techniques underscores the profound progress that technologies bring to research. Other notable studies in this area that are precursors to the advent of NLP-based studies include Hadlock and Pierce (2010) who expand on Kaplan and Zingales (1997) and emphasize the role of firm size and age, and Whited and Wu (2006) who use a structural model to measure constraints and show that constrained firms earn higher returns. The core question many studies ask in this area is simply: is a given firm financially constrained or not? This line of inquiry aligns with using ROC-Targeted NLP methods.

The first study to bring NLP to this area, Hoberg and Maksimovic (2015), uses precise anchor-phrase queries to identify direct statements indicating constraints. Specifically, the authors search for terms relating to delay and its synonyms that appear within close proximity to mentions of a type of investment. Moreover, any hits on this query must exist in the firm's discussion of liquidity that typically appear in the MD&A section of the 10-K. This approach relies heavily on highly accurate document parsing that can detect subsections in major sections of a 10-K, a challenging task that the metaHeuristica platform excels at (this tool facilitated many other research projects requiring fast queries, accurate document parsing, and early access to emerging technologies). A discussion of needing to delay an investment when discussing inadequate capital liquidity is the very essence of financial constraints, indicating the clarity of and power of ROC-Targeted methods in this context. The authors are further able to extend the dialog about financial constraints by adding further richness to the anchor-phrase queries to measure equity and debt financial constraints separately.

Other studies measuring financial constraints using advanced technologies include Buehlmaier and Whited (2018) (also uses anchor-phrase queries) who study the asset pricing implications of financial constraints, Bodnaruk et al. (2015) who use a word-list approach to measure financial constraints, and Linn and Weagley (2019) who extend the measures in Hoberg and Maksimovic (2015) using machine learning. This latter study brings a novel technological perspective that text might not be necessary to solve some problems, as they use machine learning techniques to project existing financial constraint measures onto an array of accounting variables, allowing the authors to extend the sample well backward in time before textual data became available. This also allows the authors to fill out the cross section to include observations that were not in the original sample due to technical limitations.

These studies overall illustrate that richer combinations of technology-enhanced methods and work by multiple research teams can bring a powerful set of empirical tools to bear on an important question. These studies also illustrate many important economic conclusions including a deeper understanding of the financial crisis, the perspective that the most severe financial constraints are faced by innovative firms attempting to raise costly equity capital, that informational frictions likely play a role, and that constraints are adequately impactful to also impact asset pricing returns.

5.4 Innovation

The use of NLP technologies in research is an innovation in its own right. It is thus not surprising that these tools have become widely used to research the topic of innovation itself.

Research in this area often focuses on the patent corpus and scores patents on important characteristics to examine ROC-targeted objectives. Early studies score patents on innovativeness. For example, Packalen and Bhattacharya (2015) use the first appearance of the words in the patent corpus overall. Bena et al. (2022) scores patents on whether they focus on process or non-process innovation. They do so by examining whether claims mention key clauses such as "A method for" or "A process for". Bowen et al. (2023) examine rapidly emerging technologies by scoring individual words based on how rapidly they are growing year-over-year in the patent corpus overall. They find that startups having patents with high growth vocabulary exit via IPO and those with low growth vocabulary exit via acquisitions.

Other studies use more advanced technologies to study innovation. Kelly et al. (2021b) score patents as to whether they are breakthroughs by examining their textual similarity to future patents relative to their similarity to past patents. This approach is ROC-comparative and requires a large number of comparisons between patent text and dynamic vocabularies. Acikalin et al. (2022) use transformer-based embedding models and train a machine learning model to predict which patents are exposed to the Alice vs CLS Bank Supreme Court case that eliminated patent protection for many methods patents. The authors find that losses in IP protection significantly harm the competitive position of smaller firms likely due to their limited barriers to entry, whereas larger firms experience no losses and even experience modest gains likely at the expense of their smaller rivals.

Chen et al. (2019) take an industry-specific approach and score patents on the extent to which they are in the fintech area. They manually classify 1000 patents based on the type of fintech technology the patent addresses, followed by a second-stage machine learning algorithm that uses the trained sample to score a much larger sample of 67,948 patents. This is a novel approach given the existence of known categories, and the authors go on to document that fintech innovations bring positive value to inventors, but often impact industry valuations negatively.

Other articles study innovation using corpora other than the patent corpus. Bellstam et al. (2021) use Latent Dirichlet Allocation on analyst reports to measure innovation by extracting a core topic from these documents that zeroes in on innovation. Their innovation measure predicts future performance and growth opportunities. Hoberg and Maksimovic (2022) uses 10-Ks to score firms based on the extent to which their products are in each of four stages of the product life cycle with product innovation being the first and process being the second. The authors find that the relationship between observed investment and Tobins' Q is significantly stronger (and informative about the type of investment) when the Q-model conditions its predictive relationship on the ex-ante life cycle stages. Finally, Liu et al. (2021) study cryptocurrency white papers and use machine learning to measure the technological sophistication of cryptocurrencies. The authors find that technologically more advanced currencies earn significantly higher initial returns until listing, but underperform subsequently, consistent with overreaction.

5.4.1 Labor and Artificial Intelligence

Some studies examine the impact of AI itself on corporate finance outcomes, and do so by using AI technology itself to facilitate the research. Babina et al. (2024) is an early contribution in this area and the authors first develop a measure of AI exposure at the firm level as a crucial first step. They do so by first identifying terms indicating core AI skills: artificial intelligence, machine learning, natural language processing and computer vision. They then examine all job skills listed in existing labor databases by Cogism and Burning Glass and examine the extent to which each skill co-occurs with the core terms. The result indicates AI exposure at the skill level, which can then be aggregated to the occupation and firm level, facilitating targeted measures of how much each firm is investing in artificial intelligence (Cogism and Burning Glass data provide foundations via resume and job postings data). This approach indicates a strategy that taps expert work in AI already conducted by Cogism and Burning Glass to categorize 535 million employment histories of individuals worldwide based on resumes and 180 million job vacancies. The result is significant power and the authors find that AI firms grow faster and have higher market valuations, and notably, that growth is focused especially on larger firms and that AI-induced growth is correlated with higher industry concentration. Additional work by the same authors (Babina et al. (2023)) illustrates that workforce composition itself experiences shifts toward STEM and more highly educated workers, and at the same time, towards a flattening of hierarchical structures.

Another study in this area, Jiang et al. (2021), focuses on the role of AI specifically in the area of FinTech. These authors draw from earlier work classifying patents based on Fintech contributions by Chen et al. (2019) and first extract keywords from these patent titles and abstracts to develop keywords. The authors then use similarity analysis (ROC-comparative) to score job descriptions and thus occupations based on fintech exposures, thus facilitating the ability to measure fintech hiring. These authors find declines in hiring by fintech firms but also find that hiring is uneven. For example, workers with finance and technology experience increased openings indicating complement and substitution effects.

All studies noted above benefit from more complex pipelines than do many other studies in corporate finance. In particular, these methodological strategies benefit from AI and NLP techniques optimized by professional data providers such as Cogism and Burning Glass, and also from prior work by other scholars (such as patent classifications by Chen et al. (2019)), in addition to their own organic work on NLP. This style of work illustrates a key prediction by the authors of this study: over time, some of the most impactful research in finance will likely be facilitated by larger and more cross-disciplinary teams that potentially might also include industry partners.

5.5 Emerging Areas in Corporate Finance

As noted above, it is not possible for our study to summarize the entire literature in corporate finance as the full impact of advanced technologies has become vast. We end this section with an overview of sample areas that are likely to grow in coming years.

One key example is the study of risk management. Here, a diverse set of research objectives helped to motivate important findings that will likely stimulate more research. Examples are the measurement of specific operational hedging strategies (Hoberg and Moon (2017)), economic policy uncertainty (Bloom et al. (2020) and Hassan et al. (2019)), product market competitive risks (Hoberg et al. (2014)), and financial instability risk (Hanley and Hoberg (2019)). As was the case for the study of financial constraints, precise anchor-phrase technologies facilitated by the metaHeuristica platform were instrumental in the first study's ROC-targeted objectives. Customized anchor phrase methods were used in the second's ROC-targeted framework, ROC-comparative methods using the TNIC platform were used in the third, and ROC-holistic methods based on topic models and word2vec embeddings in the last.

Other exciting areas include corporate culture (Li et al. (2021)) and regulatory impact

(Kalmenovitz et al. (2021) and Kalmenovitz and Chen (2020)). These topics are particularly challenging to measure due to existence of potential confounders and the complexity of the themes associated with their respective foundations. The first study uses embedding technologies in tandem with machine learning techniques to zero in on culture-identifying content in earnings conference calls. The second measures the complex topic of regulatory fragmentation using an ROC-holistic approach using Latent Dirichlet Allocation to quantify the content of the vast Federal Register corpus, and then map exposures to the firm-year level using LDA on firm annual reports. The third takes an ROC-comparative approach and identifies the extent to which two firms face the same regulations at the pair level, documenting unique components of firm comovement not explained by other forms of similarity.

A final example is the study of topics relevant to network connections among firms and political economy. In addition to the concept of EPU noted above (a foundational contribution to the study of political economy), Goldman et al. (2024) use a large number of newspaper articles and a sentiment-based framework to study the role of political polarization in the financial news and its impact on stock prices, and Hoberg and Neretina (2023) use website data and entity recognition techniques to explore the role of trade associations in shaping corporate finance strategies and outcomes. These studies thus adopt ROC-targeted methods.

We believe these areas have strong potential for continued research. More broadly, the authors of this study believe the profession is still in its early stages regarding the growth and potential that technology-enhanced research will bring to the study of corporate finance.

6 Asset Pricing

NLP is revolutionizing empirical asset pricing by enabling researchers to systematically analyze vast amounts of qualitative information that can affect asset values. This includes news articles, corporate disclosures, social media posts, and other textual data that can provide insights into investor sentiment, information diffusion, and market efficiency. The literature has evolved from simple dictionary-based approaches to sophisticated machine learning techniques, paralleling broader advances in natural language processing.

A common null hypothesis in financial economics is that markets are informationally efficient Fama (1970). In such markets, prices reflect all the information available to investors. A large body of work, starting with Fama et al. (1969) studies how asset prices respond to new information and how this information becomes incorporated into prices. Tests of the efficient markets hypothesis require the econometrician to observe changes in the information set of the marginal investor. It is therefore quite natural that some of the first applications of natural language tools in finance involve better measurement of the information released to the public in the form of news articles and corporate announcements.

Measuring the information set available to investors is, in fact, a cornerstone of modern asset pricing. The stochastic discount factor / generalized method of moments view of asset pricing emphasized in Cochrane (2005) postulates that the time t equilibrium price p_t of an asset generating a future payoff x_{t+1} can be represented as a conditional expectation,

$$p_t = E[m_{t+1}x_{t+1}|\mathcal{F}_t], \tag{5}$$

where m_{t+1} is a stochastic discount factor. Importantly, the expectation is taken conditional on the information $\mathcal{F}_t$ available to investors at time t. A central insight of Hansen and Richard (1987) is that the unobservability of agents' conditioning information sets is a fundamental challenge to testing conditional asset pricing models. Consequently, even if we could observe the stochastic discount factor, the inherent inability to capture agents' complete information sets renders empirical tests of equilibrium asset pricing models problematic.

Some of the early work applying natural language to macro-asset pricing has seen the fruitful use of textual data like news coverage to measure risk and risk premia. We survey this literature below, organized by themes of research.

6.1 Short-term Return Predictability

Return predictability arising from textual sources can reflect both rational responses to information and behavioral biases in information processing. Predicting returns with text is a clear example of ROC-targeted. Niederhoffer (1971) documented that major news events reported in newspaper headlines predict stock price movements. This early contribution used one of the simplest possible measures of salience, headline size, to identify important news. But this early work inspired a literature examining how financial markets process textual information.

Antweiler and Frank (2004) analyze stock message boards, showing that online discussion volume predicts volatility and trading activity. Das and Chen (2007) study message board posts about stocks, developing methods to extract sentiment from internet discussions and showing that aggregate message board sentiment relates to stock index levels, volumes, and volatility.

Tetlock (2007) examines daily content from the Wall Street Journal's "Abreast of the Market" column, finding that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals. High or low pessimism also forecasts increased trading volume. These patterns are consistent with theories of noise trading rather than fundamental information.

García (2013) studies financial news from the New York Times over 1905-2005, documenting that news content more strongly predicts stock returns during recessions, particularly at the onset of economic downturns. This state-dependence suggests sentiment effects are amplified during periods of uncertainty and market stress.

Jiang et al. (2019) construct a sentiment index from corporate financial disclosures, finding negative correlation with future returns. The predictive power exceeds traditional sentiment measures and is stronger for firms that are difficult to value.

Chen et al. (2022) analyze Seeking Alpha posts, showing that user-generated content predicts future returns and earnings surprises. García et al. (2023) develop a machine learning approach to generate finance-specific sentiment dictionaries, showing that color words in financial text predict stock price movements.

Sentiment can capture the positivity of a change in economic fundamentals and market attitudes about a firm, but it can also capture the biases of the authors of the text or the slant of the publication. For example, in a foundational example of ROC-Targeted, Dougal et al. (2012) use exogenous scheduling of Wall Street Journal columnists to identify a causal relation between financial reporting and stock market performance. They find that financial journalists amplify or attenuate prevailing sentiment.

Dim et al. (2023) argues that media bias-driven return volatility is inversely related to price informativeness. It uses a topic model to measure stock return exposure to media narratives, as a proxy for media bias, and finds support for this mechanism. Mamaysky (2024) shows that media and markets generate a negative feedback loop during the onset of the COVID-19 pandemic, with especially pronounced effects for high narrativity news content.

6.2 Slow Adjustment to Information

Some textual information appears to be incorporated gradually into prices, suggesting limits to investor attention and processing capacity. Peress (2014) documents a causal impact of the media on trading volume and price volatility by exploiting national newspaper strikes. He uses precise anchor phrase queries to identify strikes based on media coverage text, and finds that the media contribute to the efficiency of the stock market by improving the dissemination of information among investors and its incorporation into stock prices. Manela (2014) models how information diffusion affects asset prices, using media coverage of drug approvals to measure information transmission rates. This example of an ROC-targeted analysis measures media exposure by searching for unique pharmaceutical drug names in the text. The analysis reveals that the value of information is hump-shaped in its future transmission rate, and that the estimated amount of noise trading is small. These findings help explain the relationship between information diffusion speed and price discovery.

Boudoukh et al. (2018) uses a text regression approach called VIP to identify news that is relevant to individual stocks. The paper leverages the richness of textual content to resolve a long standing puzzle that market model R^2s are similar on news and no news days (Roll, 1988). Using the set of relevant news, it finds that when relevant news can be properly identified, stock price movements are closely linked to the arrival of this information.

Hoberg and Phillips (2018) show that text-based industry momentum captures gradual information diffusion about similar firms. Cohen et al. (2020b) study changes in the language of 10-K and 10-Q reports, finding that firms with substantial year-over-year changes in their disclosure language experience negative future returns. A portfolio that shorts firms with large textual changes earns abnormal returns of over 22% annually, with no immediate price reaction. Breitung and Müller (2022) develop NID, a language model that extracts semantically novel information from financial texts, and apply it to annual reports. They find that NID outperforms word-based methods in terms of alpha and show that the lazy prices anomaly is confined to strictly regulated markets and small stocks. Fedyk (2024) examines news positioning using Bloomberg terminal data, documenting higher trading volume and larger excess returns for front-page news compared to similar non-front-page content.

The last four are prominent examples of ROC-comparative methods, which become well-motivated when financial economists are interested in treatment effects relative to counterfactual non-treated observations. Fedyk (2024), for example, measures news importance using machine learning, including a comparison of the distributions of topics covered by the news using topic modeling and a deep-learning classifier of news importance based on the text of the news headlines. The similarity of the news is central to the paper's ability to attribute return differences to the appearance on the front-page.

6.3 Risk and Uncertainty Measurement

Textual analysis helps measure various forms of risk and uncertainty. Baker et al. (2016) construct a news-based index of economic policy uncertainty that predicts investment, employment, and output. Manela and Moreira (2017) develop a news-based measure of uncertainty and rare disaster concerns by analyzing Wall Street Journal front pages from 1890-2009. Their news-based VIX predicts future volatility and captures disaster concerns not reflected in standard measures. Both papers are targeted at measuring risk using news text, but they opt for different methods. Baker et al. (2016) use anchor phrase queries, searching for news than mentions a word from three *economic*, *policy*, and *uncertainty* word lists, relying on researcher expertise in selecting the best phrases. In contrast, Manela and Moreira (2017) use supervised machine learning, and run a regularized text regression of VIX on word counts mentioned in contemporaneous front-page news coverage, letting the data choose the relevant words.

Hassan et al. (2019) create text-based measures of firm-specific political risk from earnings call transcripts, showing how political uncertainty affects corporate investment and hiring decisions. The paper offers a novel approach to creating political and non-political word lists relevant to their purpose using subsets of textbook material. Jiang et al. (2022) leverage large language models to predict returns, showing that contextual embeddings can capture risk-related information beyond traditional approaches.

Theoretically, efficient market prices should change when new information is released, but not when the same news is republished, although that is not always the case in practice (Huberman and Regev, 2001). Several papers use ROC-comparative approaches to measure news novelty based on textual analysis. Tetlock (2010) defines the staleness of a news story as its textual similarity to the previous ten stories about the same firm. It finds that stock returns respond less to stale news. Glasserman et al. (2023) measures news novelty through an entropy measure calculated using a recurrent neural network. It finds that cross-sectional entropy exposure carries a negative risk premium.

Recent examples of ROC-holistic include Bybee et al. (2023) and Lopez-Lira (2023) which use thematic topic model techniques to extract systematic risk factors from text, linking news narratives and risk disclosures to the cross-section of returns.

6.4 Institutional investors

The relationship between news and institutional trading provides insights into market efficiency and information advantages. Huang et al. (2020) examine how institutions trade around news releases using comprehensive newswire data matched with institutional trading records. They find evidence that institutions anticipate certain types of news content, suggesting sophisticated information processing capabilities. Huang et al. (2023) study bond fund trading around corporate news events, showing how institutions provide liquidity in response to news-driven price pressure.

6.5 Social media

Networks of investors generate text data that reveals how information and opinions propagate through markets. The literature on social networks in finance is large and growing. We therefore survey only a small subset of this literature and refer the interested reader to a recent survey by Cookson et al. (2024b).

In seminal work, Antweiler and Frank (2004) examines the impact of internet stock message boards on market behavior by analyzing over 1.5 million posts from Yahoo! Finance and Raging Bull regarding traded companies. This early contribution is an example of ROC-

Targeted objectives, which uses a Naïve Bayes text regression to classify a training set of messages as buy, hold, or sell. Based on this measure of bullish sentiment it finds that message board discussions can help predict market volatility, though their influence on stock returns is statistically significant but economically small. The authors also observe that disagreement among message posters correlates with increased trading volume.

Cookson and Niessner (2020) investigates the sources of investor disagreement by analyzing sentiment data from StockTwits, a social media investing platform, alongside users' investment approaches, such as technical or fundamental analysis. It relies on self-reported sentiment by users, as well as a maximum entropy sentiment classifier. The paper distinguishes between disagreement caused by differing information sets and disagreement stemming from the interpretation of the same information. The authors find that overall disagreement is equally driven by both factors, but within-group disagreement correlates more strongly with trading volume. Their results suggest that differences in information are more significant for trading behavior than differences in investment styles.

Cookson et al. (2023a) analyze 250 million posts on a social investment platform, finding that users develop increasingly similar language and views over time, with correlated trading patterns becoming less profitable as echo chambers strengthen. Cookson et al. (2024a) combine text from social media posts with trade data, documenting that investors with linguistically similar posts engage in correlated trading, especially during market stress. Hirshleifer et al. (2024) study social transmission of earnings news using county-level networks and StockTwits messages. They find that firms in more connected counties see stronger immediate price reactions and faster post-announcement adjustment but persistent high volume, consistent with their social churning hypothesis.

Cookson et al. (ming) shows that social media sentiment predicts merger withdrawals. To better understand what generates this predictability, they use a topic model to decompose the social media signal into tweets about different topics that are informative of fundamental information (company discussions, discussion of deal terms, etc.) versus discussions that are not (meme tweets, and technical tweets). They find that this predictability is driven by longer acquisition-related tweets by fundamental investors, rather than memes and price trend tweets.

Cassella et al. (2023) study posts by social media forecasters. It documents that optimistic forecasters react to negative news by shifting their expectations to a longer forecast horizon. To measure the degree to which these posts are about the long-term, an ROC-targeted goal, they use word embeddings to construct a lexicon made of the top terms closest to "long-term" and to "long-haul". They then use this lexicon to obtain long-term language scores for each post.

Li (2025) studies the role of ideological narratives as a source of asset pricing risk. It uses social media and large language models to measure ideology dynamics. This ROC-targeted study starts with a set of ideologies that are important in the context of cryptocurrencies, anarchism and decentralization, and then uses both precise anchor phrases and large language models to measure their salience in social media text. It finds that fluctuations in these ideology dynamics are priced in the cross-section of cryptocurrency returns.

6.6 High-Frequency (Intraday) Return Predictability

High-frequency intraday analysis offers the distinct advantage of large numbers of observations, which is crucial for reliable statistical inference from high-dimensional sets of explanatory variables. The ultra high dimensionality of text data, therefore, allows teasing out dependencies with text without resorting to excessive dimensionality reduction.

The processing of textual information at high frequencies provides a unique laboratory for studying market efficiency. Chinco et al. (2019) use the LASSO method to predict one-minute ahead returns from lagged returns across the entire market, showing the value of machine learning for high-frequency prediction. Aleti and Bollerslev (2024) conduct a comprehensive analysis of intraday news effects on asset prices, finding that monetary policy and financial sector news generate the strongest price responses. Both of these ROC-targeted studies use text regression methods, leveraging the high frequency nature of their data.

7 Other Topics in Finance

7.1 Sustainability

Our focus on sustainability includes studies that focus on traditional areas such as climate finance and the environment, and also more emerging but related areas such as biodiversity. We view the study of sustainability to have excellent potential, especially regarding potential applications using NLP tools to pioneer the creation of new databases that can empower future research previously unexplored. The study of all aspects of ESG including the social and governance aspects are fertile for future work, as are the political and economic consequences of ESG movements.

News coverage, corporate disclosures, and patent documents help measure emerging climate risks and market responses where traditional numerical data fall short. Engle et al. (2020) develop a dynamic hedging procedure for climate risk, extracting innovations from newspaper-based climate news series. Using third-party ESG scores to model firm exposures, they construct industry-balanced mimicking portfolios that hedge climate news innovations both in and out of sample.

Cohen et al. (2020a) analyze patent texts to measure green innovation, finding that firms with more climate-mitigation patents receive lower environmental ratings than industry peers. Bingler et al. (2022) develop ClimateBert, a language model trained on climate-related disclosures, documenting widespread use of vague language and selective disclosure in corporate climate risk reporting.

Sautner et al. (2023) develop a machine learning keyword discovery algorithm to measure firms' climate change exposures in earnings calls. Applied to over 10,000 firms across 34 countries from 2002-2020, their measures predict job creation in green technologies and green patenting, with evidence of pricing in options and equity markets. Leippold (2023) explore GPT-3's knowledge of climate finance through a structured interview approach, assessing the model's understanding of climate risks and sustainable investing concepts.

In recent years, biodiversity finance has gained increased interest as stakeholders increasingly acknowledge the profound links between biodiversity loss, climate change, and other urgent environmental concerns. This emerging domain underscores the need for novel financial mechanisms, and also novel measures of biodiversity risk and attention which have often relied on text as data.

Giglio et al. (2023) develop a news-based measure of aggregate biodiversity risk and utilize textual analysis of firms' 10-K statements to construct firm-level measures of exposure. Their findings show that these exposures, which differ from climate-related risk profiles, already shape equity prices, even though many market participants believe current pricing do not fully account for biodiversity risk.

Chen et al. (2023) examine how financial markets price the costs associated with biodiversity conservation by focusing on China's "Green Shield Action". The authors draw on text data from local government documents and official bond disclosures to measure market sentiment and policy-related information. They find that the initiative improved biodiversity, but also led to significant increases in bond yields for municipalities with these reserves, effectively increasing the cost of public capital.

Chen et al. (2024) investigate the socioeconomic impact of a strict seasonal fishing ban along the Yangtze River, highlighting how preserving local fish resources spurred higher-quality scientific research. By analyzing the textual content of research funding applications, publications, and media coverage, they demonstrate how biodiversity policies can indirectly stimulate innovation and academic output.

Overall we believe that measurement of sustainability concepts is a big theme in many works, and given the complexity of all three parts of ESG and related movements, continuing to build accurate measures of the components and their impact on business remains a fruitful area for future researchers.

7.2 Central Bank Communication

Central banks rely extensively on language to signal policy and shape expectations. Jefferson (2025) provides an in depth review of this literature. Hansen et al. (2018) analyze FOMC transcripts to study how transparency affects monetary policy deliberations. Cieslak and Vissing-Jorgensen (2021) examine FOMC documents from the mid-1990s, showing that negative stock returns predict policy accommodations primarily through policymakers' concerns about consumption wealth effects. They find no evidence that the Fed overreacts to the market beyond the market's effect on growth expectations. Cieslak et al. (2021) analyze private FOMC deliberations to quantify different types of uncertainty, finding that inflation uncertainty leads to tighter policy stances, particularly when inflation approaches or exceeds targets. This response reflects the FOMC's focus on upper-tail inflation risks.

7.3 Banking and regulation

Bank supervision generates extensive textual records through examination reports and regulatory communications that reveal information not captured in numerical filings. Goldsmith-Pinkham et al. (2016) analyze supervisory issues ("matters requiring attention") raised by Federal Reserve examiners. They find that supervisory issues increase with bank size and complexity, and decrease with profitability and loan quality. Their analysis shows that stressed banks resolve issues faster, but banks facing multiple issues take longer to resolve each one, suggesting capacity constraints. Cookson et al. (2023b) study social media's role in bank runs, showing how digital networks can accelerate deposit withdrawals during banking crises.

Engelberg et al. (2023) study the partisanship of financial regulators: the Federal Reserve and the Securities and Exchange Commission. They suggest a text-based partisanship measure that captures the extent to which one can guess a speaker's affiliated political party based only on the language they use. For example, in the 2010s, "red tape" is much more likely to be used by a republican, and "climate change" by a democrat. They find that Fed Governors have remained fairly nonpartisan from 1930 to 2019, but SEC Commissioner partisanship rose to an all-time high during the 2010–2019 period.

The use of natural language data in banking and regulation is still in its infancy. Much more can be done to build this literature to leverage unstructured data produced by bank employees and bank supervisors.

7.4 Financial Crises and Systemic Risk

Financial crises are fortunately rare. For that reason, studying them requires long and wide panel data. Historical texts and media coverage provide long-term perspectives on how societies perceive and respond to financial crises. Jha et al. (2022) analyze millions of books published across eight countries, measuring sentiment toward finance over hundreds of years. They find persistent cross-country differences in finance sentiment, with more positive sentiment in more capitalist countries. They find that finance sentiment correlates with financial market participation and income inequality, and sentiment declines one year before financial crises.

This example of an ROC-targeted approach uses BERT to embed book sentences mentioning “finance” into a vector space. Its finance sentiment index measures if the average finance-mentioning sentence is semantically closer to “finance benefits society” than to “finance damages society.” As part of the paper’s validation tests, it shows that similar sentiment indices can be constructed for other important industries (paper, coal, tobacco, etc.) and other entities (banks, stocks, mutual funds, etc.). Exploring these other sentiment indices and their relation to the economy is an interesting avenue for future work.

Jha et al. (2021) rely on that measure of finance sentiment to examine how natural disasters affect public sentiment toward finance. They find that finance sentiment declines after epidemics and earthquakes but rises following severe droughts, floods, and landslides. These heterogeneous effects suggest finance sentiment responds differently to the realization of insured versus uninsured risks. Understanding the effects of an on-going crisis can be critical for policymakers. Timely newspaper coverage allows Jha et al. (2021) to measure finance sentiment at the onset of the COVID-19 pandemic. They find that finance sentiment declines at the start of the pandemic, but recovers in countries that experienced high stock markets returns and that responded with large fiscal spending. Mamaysky (2024) also examines the COVID-19 pandemic and shows that news coverage became hypersensitive during the onset of the crisis, and transitioned back to a more normal pattern by mid-March of 2020.

Also relevant to this topic, Hanley and Hoberg (2019) measure systematically important emerging risks for financial institutions in a dynamic way by linking textual content financial institutions report in their 10-K risk factor sections to dynamic evolutions in the pairwise-firm return covariance matrix. When the relationship between reported risks and covariance changes in a major systematic way in the data, the results suggest that the risks of systemic financial instability is rising. The approach used by the authors can alert market participants to potential risks years before they fully manifest. The approach is based on LDA and word2vec embeddings, and this also facilitates an ability to provide interpretations of the emerging risks so that they might be actionable to regulators.

8 Major Challenges and Directions for Future Work

In earlier sections, this article discussed many avenues for future research. In this section, we summarize some big picture issues future researchers might consider and we note larger and more challenging gaps in the literature. Our first section discusses methodology challenges and our second section discusses field-specific research areas.

8.1 Addressing Methodological Limitations

The methods noted throughout this article indicate a tremendous amount of progress in developing aspirational methods that can answer many questions not possible to answer with earlier methods. Even if no further progress is made, adequate tools exist to improve research in financial economics for many decades to come. Yet, the progress to date also reveals some gaps that indicate fruitful areas for future methodology-focused researchers.

- [**Improved thematic topic models**] While topic models such as Latent Dirichlet Allocation and follow-on models that link to embeddings are a clear breakthrough technology with some compelling works already in the finance literature, many researchers in financial economics believe that improvements in interpretability and the clarity of topics remains fruitful. Existing topic models appear to work better when documents have a strong verbal factor structure as is the case with risk factor sections of the 10-K and media content, but do not work as well with more heterogeneous text.
- [**Interpretability**] This comment reflects overwhelming demand by financial economists for improved tools regarding interpretability for essentially all NLP tools. Clarity of mechanisms and identification are core to success in financial economics as the black box critique has real consequences. The goal of improved interpretability is to produce results that are rigorous, well-identified, and that come with appropriate risk management practices. This requires clear understanding of economic mechanisms driving any particular result.

- [**Automated hypothesis formation**] While some pioneering work has been done, this area remains fruitful for many reasons. First, it is unlikely that even the best economists can reliably postulate all possible hypotheses that might explain a given result. The consequences of this human limitation include (A) exposure to systemic risk that could impose a high cost on society and (B) investment strategies that work well most of the time but crash at any time with low probability, or (C) policy choices that are easily unraveled through regulatory arbitrage or (D) policy changes that are simply ineffective because they target the wrong economic forces.

- [**GPT and bias documentation and minimization**] There is little doubt that generative AI will play a major and growing role in research and practical applications. Yet these tools are susceptible to significant biases because they were trained using recent data and financial economists need to back-test econometric models using older data. This can create bias because content from these LLMs might incorporate information that was not available at the time of these earlier observations. More work to rigorously document when this bias is significant and when it is not significant can improve the soundness of inferences made using these models. Additional work also needs to develop rigorous methods to help researchers debias estimates.

- [**Best practices for LLM prompting**] It is well known among practitioners that prompts to LLMs such as ChatGPT and Claude can be well designed or of poor quality. Developing best practices for prompting in financial economics is a largely un-studied topic that is likely to be important. Such standards and best practices can significantly improve the rigor of work in this area.

- [**Numeracy: extracting numbers from text**] Many large corpora that financial economists develop contain both text and numbers, where the numbers often refer to quantities or dollar amounts that are indicated by text. Further work developing methods to better link these textual and numerical quantities should prove useful in many future research projects.

8.2 Field-Specific Topics for Future Researchers

NLP methods have brought tremendous progress to advancing research on core topics in financial economics. Yet an even larger number of research topics remain given the depth of these tools and available corpora. We note select examples to illustrate fruitful areas for future research that have high potential for impact.

- [**Empower more quasi-natural experiments**] The literature in financial economics remains focused on inferences with rigorous and unambiguous interpretations. The standard is to tap economic variables with plausibly exogenous variation. As we noted in this review, NLP methods can be used to create plausibly exogenous instruments for important economic events where numerical measures of intensity are not available in some cases.

- [**Mechanism revelation**] Another approach to improving identification is the crafting of mechanism-revealing tests, a topic discussed throughout this review. Proving mechanisms has become pervasive to how financial economists evaluate research, and hence this approach can improve many research projects that are more tentative on supportive evidence.

- [**International textual corpora**] Existing research has made significant progress on understanding major U.S. corpora, but exploring international corpora remains under-researched, especially in developing economies. Work in this area will require overcoming challenges of multi-lingual content and heterogeneous data formats. Yet novel databases (especially data created by practitioners or joint teams of academics and practitioners) might exist or reduce the cost of creating frameworks in some settings.

- [**ESG measurement**] Much pioneering work has already been done in this area. Yet we believe that measurement of how various economic agents achieve non-pecuniary goals remains fertile as a research area. This statement reflects the fact that the way firms contribute to society and ESG goals is likely higher dimensional than is currently represented in the literature.

- [**Better links between subdisciplines such as corporate finance, asset pricing, accounting, and economics**] Many disciplines in finance are relevant to one another, but research linking these disciplines remains fruitful and NLP might offer novel ways to link these topics.

- [**Corporate culture & strategy**] Pioneering work in these areas has illustrated the importance of these topics. Yet both culture and strategy are likely higher dimensional and more inter-related than is currently depicted in the literature. As both are highly represented in various corpora, the potential to further inform both remains fruitful.

- [**Policy and regulation**] A large fraction of the total workforce of major economies works for government and regulatory agencies. While pioneering work has established strong results indicating importance and novel findings, the dimensionality of this space is likely far larger than what has been documented thus far, and large corpora exist to further explore this topic.

- [**Managerial responses to investor use of AI**] While some work begins to explore this issue (Cao et al. (2023a)), it is likely that the process of managers adjusting conference calls, press releases and SEC disclosures to accommodate (or obfuscate) investor attempts to use AI is still in progress and evolving.

- [**Consumer welfare**] Many studies that bring concepts from industrial organization into the study of finance have made significant progress especially in studying how important theories dynamically impact firm strategies and performance. Yet a significant gap is consumer welfare, as existing numerical data on prices such as the Nielsen database, while outstanding, only covers a small fraction of the public firm (and private firm) samples studied in most research projects. Novel textual corpora might exist to more systematically evaluate consumer welfare and price levels, thus empowering many research projects as many theories have strong predictions in this area.

- [**Household finance and advisory services**] Although some work on robo-advising already exists (see D'Acunto et al. (2019) for earlier work and Lo and Ross (2024) regarding ChatGPT), we believe much remains to be learned regarding how AI-based financial services can best serve customers in dynamic environments and without bias.

- [**Impact of artificial intelligence on finance**] Some pioneering work has opened this as a new field, and it is clear that new AI technologies will have a large impact in business. For example, as adoption of natural language tools by practitioners grows, it is possible that some well documented relationships between news and market prices could change (Glasserman et al., 2025). Yet much uncertainty remains and too little is currently known across the many sub-disciplines in financial economics.

- [**Network databases and pairwise interactions**] Research using textual corpora to generate dynamic networks is maturing. Yet new corpora offering identification of new dynamic network evolutions remains fruitful and could leverage impact.

- [**Risk management**] The study of risk management has been challenging in earlier years partly due measurement challenges as second moment concepts are less commonly reported numerically in public databases. This area is relevant to both corporate finance and asset pricing, as both firms and investors can face significant risks as well as potential systemic exposures.

- [**Social media**] Although work using text from social media has been growing, the high dimensionality and highly dynamic nature of corpora in this area indicates many continued opportunities. This is particularly true when topics can link individual behaviors or decisions to business or investing outcomes.

- [**AI and allocation efficiency**] Many studies find results consistent with return predictability relating to slow price adjustment. A bigger question would be to assess overall market quality and the allocation of capital (see Bai et al. (2016) for earlier work), as basic arguments would suggest improvements are likely to materialize but it is at present difficult to assess the magnitude.

9 Conclusion

This article is designed to help readers build a confident understanding of the history regarding how tools in natural language processing (NLP) have evolved in the study of financial economics and related disciplines. With this foundation in place, a core objective is to then inform researchers first regarding how to assess if these tools can offer benefits to their specific research or applications, and how to select specific NLP tools that are a best fit for their applications. The content focuses on academic research as examples of the framework and context, but the framework for project evaluation is relevant to academic researchers, practitioners, and regulators alike who are building financial applications.

We first illustrate the channels through which many research projects can realize major benefits from NLP tools such as: measuring economic quantities not possible before, improving the ability to identify causal channels, documenting economic mechanisms in action, achieving improved signal strength by reducing measurement error, and improving interpretability. Yet researchers should also be aware that some projects might not be ideally suited to these methods, especially when available corpora do not have well-matched content, corpora are too small, numerical constructs provide best measurement, or if verbal content is too susceptible to voluntary or strategic reporting. Researchers also should be aware of the black box critique, although this concern can be addressed through improved methodological choice focused on methods with high interpretability.

Once a researcher has determined that NLP tools are a good fit for their research objectives, this article next guides researchers on how to choose best-practice methods from the vast array of available NLP tools. This is done by recognizing that research objectives fall into three research objective categories (ROC): ROC-targeted, ROC-comparative, and ROC holistic. This classification, alongside considerations regarding resource availability and the priority of interpretability, usually indicates a best-practice methodology for a specific research question.

This article ends with a focused survey of developments in corporate finance, asset pricing and other areas in financial economics regarding how these major areas have evolved using specific methods and why. We annotate this discussion noting areas that should be of interest to future researchers. We also note how future methodological selection is likely to shift toward large language models over time, although this transition will take time and will be uneven. We believe advances in financial economics are still in the early stages, and the potential for future research to benefit from these tools far exceeds the work that has been done to date.

References

Acikalin, U. U., Caskurlu, T., Hoberg, G., and Phillips, G. M. (2022). Intellectual property protection lost and competition: An examination using machine learning.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In Ammar, W., Louis, A., and Mostafazadeh, N., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleti, S. and Bollerslev, T. (2024). News and Asset Pricing: A High-Frequency Anatomy of the SDF. *The Review of Financial Studies*, page hhae019.

Antweiler, W. and 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.

Ash, E. and Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, 15(1):null.

Athey, S. and Imbens, G. (2015). Machine Learning Methods for Estimating Heterogeneous Causal Effects. *GitHub*.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151:103745.

Babina, T., Fedyk, A., He, A. X., and Hodson, J. (2023). Firm investments in artificial intelligence technologies and changes in workforce composition. 31325.

Bai, J., Philippon, T., and Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654.

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

Ball, C., Hoberg, G., and Maksimovic, V. (2015). Disclosure, business change and earnings quality. *Available at SSRN 2260371.*

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2024). Inference for Regression with Variables Generated by AI or Machine Learning.

Bellstam, G., Bhagat, S., and Cookson, J. A. (2021). A Text-Based Analysis of Corporate Innovation. *Management Science*, 67(7):4004–4031.

Bena, J., Ortiz-Molina, H., and Simintzi, E. (2022). Shielding firm value: Employment protection and process innovation. *Journal of Financial Economics*, 146(2):637–664.

Benson, K., Cheng, I.-H., Hull, J. C., Martineau, C., Nozawa, Y., Strela, V., Wu, Y., and Yuan, J. (2024). Understanding the Excess Bond Premium.

Bernile, G. and Lyandres, E. (2019). The Effects of Horizontal Merger Operating Efficiencies on Rivals, Customers, and Suppliers. *Review of Finance*, 23(1):117–160.

Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bloom, N., Davis, S. J., Foster, L., Lucking, B., Ohlmacher, S., and Saporta-Eksten, I. (2020). Business-level expectations and uncertainty.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Bonsall, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2):329–357.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2018). Information, Trading, and Volatility: Evidence from Firm-Specific News. *Review of Financial Studies*.

Bourveau, T., She, G., and Žaldokas, A. (2020). Corporate Disclosure as a Tacit Coordination Mechanism: Evidence from Cartel Enforcement Regulations. *Journal of Accounting Research*, 58(2):295–332.

Bowen, D. E., Frésard, L., and Hoberg, G. (2023). Rapidly Evolving Technologies and Startup Exits. *Management Science*, 69(2):940–967.

Bowen, D. E., Frésard, L., and Taillard, J. P. (2017). What's Your Identification Strategy? Innovation in Corporate Finance Research. *Management Science*, 63(8):2529–2548.

Breitung, C. and Müller, S. (2022). New Information Detection using Language Models.

Breitung, C. and Müller, S. (2023). Global Business Networks.

Brown, S. V. and Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*, 49(2):309–346.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*.

Buehlmaier, M. M. and Whited, T. M. (2018). Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, 31(7):2693–2728.

Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business News and Business Cycles. *The Journal of Finance*, page jofi.13377.

Bybee, L., Kelly, B., and Su, Y. (2023). Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Cabezon, F. and Hoberg, G. (2022). Leaky Director Networks and Innovation Herding.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The Production of Information in an Online World. *The Review of Economic Studies*, 87(5):2126–2164.

Calomiris, C. W., Çakır Melek, N., and Mamaysky, H. (2024a). Big Data Meets the Turbulent Oil Market.

Calomiris, C. W., Mamaysky, H., and Yang, R. (2024b). Measuring the Cost of Regulation: A Text-Based Approach.

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

Cao, S., Yang, B., and Zhang, A. L. (2023b). Managerial risk assessment and fund performance: Evidence from textual disclosure. *Available at SSRN 4060307*.

Cassella, S., Dim, C., and Karimli, T. (2023). Optimism Shifting.

Chava, S., Du, W., Shah, A., and Zeng, L. (2022). Measuring firm-level inflation exposure: A deep learning approach.

Chen, A. J., Hoberg, G., and Zhang, M. B. (2022). Institutional Participation in Information Production and Anomaly Returns. *Wisdom of the Institutional Crowd: Implications for Anomaly Returns: Chen, AJ| uHoberg, Gerard| uZhang, Miao Ben*.

Chen, D. and Manning, C. (2014). A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750.

Chen, F., Chen, M., Cong, L. W., Gao, H., and Ponticelli, J. (2023). Pricing the Priceless: The Financing Cost of Biodiversity Conservation.

Chen, M., Cong, L. W., Gao, H., Karolyi, G. A., and Wang, H. (2024). Regulating Fishing in the Yangtze for Biodiversity Unintentionally Boosted Science Research.

Chen, M. A., Wu, Q., and Yang, B. (2019). How Valuable Is FinTech Innovation? *The Review of Financial Studies*, 32(5):2062–2106.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse Signals in the Cross-Section of Returns. *The Journal of Finance*, 74(1):449–492.

Chod, J., Lyandres, E., and Yang, S. A. (2019). Trade credit and supplier competition. *JOURNAL OF FINANCIAL ECONOMICS*, 131(2):484–505.

Cieslak, A., Hansen, S., McMahon, M., and Xiao, S. (2021). Policymakers' Uncertainty.

Cieslak, A. and Vissing-Jorgensen, A. (2021). The Economics of the Fed Put. *Review of Financial Studies*, 34(9):4045–4089.

Cochrane, J. H. (2005). *Asset Pricing*. Princeton University Press, revised edition.

Cohen, L., Gurun, U. G., and Nguyen, Q. H. (2020a). The ESG-innovation disconnect: Evidence from green patenting.

Cohen, L., Malloy, C., and Nguyen, Q. (2020b). Lazy Prices. *The Journal of Finance*, 75(3):1371–1415.

Cohen, L. and Nguyen, Q. (2024). Moving Targets.

Cong, L. W., Liang, T., Yang, B., and Zhang, X. (2020). Analyzing Textual Information at Scale. *Information for Efficient Decision Making*, pages 239–271.

Cong, L. W., Liang, T., Zhang, X., and Zhu, W. (2019). Textual Factors: A Scalable, Interpretable, and Data-Driven Approach to Analyzing Unstructured Information*.

Cong, L. W. and Yang, S. (2024). Understanding Patenting Disparities via Causal Human+Machine Learning.

Cookson, J. A., Engelberg, J. E., and Mullins, W. (2023a). Echo chambers. *The Review of Financial Studies*, 36(2):450–500.

Cookson, J. A., Fox, C., Gil-Bazo, J., Imbet, J. F., and Schiller, C. (2023b). Social media as a bank run catalyst. *Available at SSRN*, 4422754.

Cookson, J. A., Lu, R., Mullins, W., and Niessner, M. (2024a). The social signal. *Journal of Financial Economics*, 158:103870.

Cookson, J. A., Mullins, W., and Niessner, M. (2024b). Social Media and Finance. In *Oxford Research Encyclopedia of Economics and Finance*.

Cookson, J. A. and Niessner, M. (2020). Why Don't We Agree? Evidence from a Social Network of Investors. *The Journal of Finance*, 75(1):173–228.

Cookson, J. A., Niessner, M., and Schiller, C. (forthcoming). Can Social Media Inform Corporate Decisions? Evidence from Merger Withdrawals. *Journal of Finance*.

Curti, F. and Kazinnik, S. (2023). Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences. *Journal of Monetary Economics*, 139:110–126.

D'Acunto, F., Prabhala, N., and Rossi, A. G. (2019). The Promises and Pitfalls of Robo-Advising. *The Review of Financial Studies*, 32(5):1983–2020.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9):1375–1388.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Dim, C., Sangiorgi, F., and Vilkov, G. (2023). Media Narratives and Price Informativeness.

Dougal, C., Engelberg, J., García, D., and Parsons, C. A. (2012). Journalists and the Stock Market. *The Review of Financial Studies*, 25(3):639–679.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2-3):221–245.

Eisfeldt, A. L. and Schubert, G. (2024). AI and Finance.

Engelberg, J., Henriksson, M., Manela, A., and Williams, J. (2023). The Partisanship of Financial Regulators. *The Review of Financial Studies*, 36(11):4373–4416.

Engelberg, J., Manela, A., Mullins, W., and Vulicevic, L. (2025). Entity Neutering.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebel, J. (2020). Hedging Climate Change News. *The Review of Financial Studies*, 33(3):1184–1216.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2):383–417.

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. (1969). The Adjustment of Stock Prices to New Information. *International Economic Review*, 10(1):1–21.

Fedyk, A. (2024). Front-Page News: The Effect of News Positioning on Financial Markets. *The Journal of Finance*, 79(1):5–33.

Fedyk, A., Kakhbod, A., Li, P., and Malmendier, U. (2024). ChatGPT and Perception Biases in Investments: An Experimental Study. *Available at SSRN 4787249*.

Ferrell, A., Manconi, A., Neretina, E., Powley, W., and Renneboog, L. (2021). Corporate Litigation, Governance, and the Role of Law Firms.

Floackis, C., Louca, C., Michaely, R., and Weber, M. (2023). Cybersecurity risk. *The Review of Financial Studies*, 36(1):351–407.

Foucault, T. and Fresard, L. (2014). Learning from peers' stock prices and corporate investment. *Journal of Financial Economics*, 111(3):554–577.

Frankel, R., Jennings, J., and Lee, J. (2022). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science*, 68(7):5514–5532.

Frésard, L., Hoberg, G., and Phillips, G. M. (2020). Innovation activities and integration through vertical acquisitions. *The Review of Financial Studies*, 33(7):2937–2976.

García, D. (2013). Sentiment during recessions. *Journal of Finance*, 68(3):1267–1300.

García, D., Hu, X., and Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, 147(3):525–549.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3):535–574.

Giglio, S., Kuchler, T., Stroebel, J., and Zeng, X. (2023). Biodiversity Risk.

Glasserman, P., Li, F., and Mamaysky, H. (2025). Time Variation in the News–Returns Relationship. *Journal of Financial and Quantitative Analysis*, 60(1):258–294.

Glasserman, P. and Lin, C. (2023). Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis.

Glasserman, P. and Mamaysky, H. (2019). Does Unusual News Forecast Market Stress? *Journal of Financial and Quantitative Analysis*, 54(5):1937–1974.

Glasserman, P., Mamaysky, H., and Qin, J. (2023). New News is Bad News.

Goetzmann, W. N., Kim, D., and Shiller, R. J. (2022). Crash Narratives.

Goldman, E., Gupta, N., and Israelsen, R. (2024). Political polarization in financial news. *Journal of Financial Economics*, 155:103816.

Goldsmith-Pinkham, P., Hirtle, B., and Lucca, D. O. (2016). Parsing the Content of Bank Supervision. SSRN Scholarly Paper ID 2750315, Social Science Research Network, Rochester, NY.

Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, 34(7):3213–3225.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The Voice of Monetary Policy. *American Economic Review*, 113(2):548–584.

Guenzel, M., Kogan, S., Niessner, M., and Shue, K. (2025). AI Personality Extraction from Faces: Labor Market Implications.

Hadlock, C. J. and Pierce, J. R. (2010). New Evidence on Measuring Financial Constraints: Moving Beyond the KZ Index. *The Review of Financial Studies*, 23(5):1909–1940.

Hanley, K. W. and Hoberg, G. (2010). The information content of IPO prospectuses. *The Review of Financial Studies*, 23(7):2821–2864.

Hanley, K. W. and Hoberg, G. (2012). Litigation risk, strategic disclosure and the underpricing of initial public offerings. *Journal of Financial Economics*, 103(2):235–254.

Hanley, K. W. and Hoberg, G. (2019). Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies*, 32(12):4543–4603.

Hansen, L. P. and Richard, S. F. (1987). The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. *Econometrica*, 55(3):587–613.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *Quarterly Journal of Economics*, 133(2):801–870.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically Consistent Large Language Models.

Hilt, V. and Schwenkler, G. (2024). The Different Networks of Firms Implied by the News.

Hirshleifer, D., Mai, D., and Pukthuanthong, K. (forthcoming). War Discourse and the Cross Section of Expected Stock Returns. *Journal of Finance*.

Hirshleifer, D., Mai, D. Y., and Pukthuanthong, K. (2025). War Discourse and Disaster Premia: 160 Years of Evidence from Stock and Bond Markets. *Review of Financial Studies*, 38(2):457–506.

Hirshleifer, D., Peng, L., and Qiguang, W. (2024). News Diffusion in Social Networks and Stock Market Reactions | The Review of Financial Studies | Oxford Academic. *Review of Financial Studies*.

Hoberg, G., Knoblock, C. A., Phillips, G., Pujara, J., Qiu, Z., and Raschid, L. (2024). Using Representation Learning and Web Text to Identify Competitor Networks. *Working Paper*.

Hoberg, G. and Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, 43:58–85.

Hoberg, G. and Maksimovic, V. (2015). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies*, 28(5):1312–1352.

Hoberg, G. and Maksimovic, V. (2022). Product life cycles in corporate finance. *The Review of Financial Studies*, 35(9):4249–4299.

Hoberg, G. and Moon, S. K. (2017). Offshore activities and financial vs operational hedging. *Journal of Financial Economics*, 125(2):217–244.

Hoberg, G. and Moon, S. K. (2019). The offshoring return premium. *Management Science*, 65(6):2876–2899.

Hoberg, G. and Neretina, E. (2023). Do Trade Associations Matter to Corporate Strategies? *Available at SSRN 4575314*.

Hoberg, G. and Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Hoberg, G. and Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hoberg, G., Phillips, G., and Prabhala, N. (2014). Product Market Threats, Payouts, and Financial Flexibility. *The Journal of Finance*, 69(1):293–324.

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

Hoberg, G. and Phillips, G. M. (2021). Scope, scale and competition: The 21st century firm.

Hobson, J. L., Mayew, W. J., and Venkatachalam, M. (2012). Analyzing Speech to Detect Financial Misreporting. *Journal of Accounting Research*, 50(2):349–392.

Hochberg, Y. V., Kakhbod, A., Li, P., and Sachdeva, K. (2023). Are Patents with Female Inventors Under-Cited? Evidence from Text Estimation.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864.

Hombert, J. and Matray, A. (2018). Can Innovation Help US Manufacturing Firms Escape Import Competition from China? *JOURNAL OF FINANCE*, 73(5):2003–2039.

Hu, A. and Ma, S. (2021). Persuading investors: A video-based study.

Huang, A. G., Tan, H., and Wermers, R. (2020). Institutional Trading around Corporate News: Evidence from Textual Analysis. *Review of Financial Studies*.

Huang, A. G., Wermers, R., and Xue, J. (2023). "Buy the Rumor, Sell the News": Liquidity Provision by Bond Funds Following Corporate News Events. *SMU Cox School of Business Research Paper*, (22-06).

Huberman, G. and Regev, T. (2001). Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar. *The Journal of Finance*, 56(1):387–396.

Israelsen, R. D. (2014). Tell it like it is: Disclosed risks and factor portfolios. *Available at SSRN 2504522*.

Jefferson, P. N. (2025). Reading between the Lines? Textual Analysis of Central Bank Communications. *Speech*, (99608).

Jha, M., Liu, H., and Manela, A. (2021). Natural disaster effects on popular sentiment toward finance. *Journal of Financial and Quantitative Analysis*, 56(7):2584–2604.

Jha, M., Liu, H., and Manela, A. (2022). Does Finance Benefit Society? A Language Embedding Approach. *Review of Financial Studies*.

Jha, M., Qian, J., Weber, M., and Yang, B. (2024). ChatGPT and corporate policies.

Jiang, F., Lee, J., Martin, X., and Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149.

Jiang, H., Khanna, N., Yang, Q., and Zhou, J. (2024). The Cyber Risk Premium. *Management Science*, page mnsc.2022.02056.

Jiang, J., Kelly, B. T., and Xiu, D. (2022). Expected Returns and Large Language Models.

Jiang, W., Tang, Y., Xiao, R. J., and Yao, V. (2021). Surviving the FinTech disruption.

Kalmenovitz, J. (2023). Regulatory intensity and firm-specific exposure. *The review of financial studies*, 36(8):3311–3347.

Kalmenovitz, J. and Chen, J. (2020). Regulatory Similarity. *Journal of Law and Economics, forthcoming (2023)*.

Kalmenovitz, J., Lowry, M., and Volkova, E. (2021). Regulatory fragmentation.

Kaplan, S. N. and Zingales, L. (1997). Do Investment-Cash Flow Sensitivities Provide Useful Measures of Financing Constraints?*. *The Quarterly Journal of Economics*, 112(1):169–215.

Kazinnik, S. (2023). Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative AI.

Ke, S., Montiel Olea, J. L., and Nesbit, J. (2024). Robust machine learning algorithms for text analysis. *Quantitative Economics*, 15(4):939–970.

Kelly, B., Manela, A., and Moreira, A. (2021a). Text Selection. *Journal of Business & Economic Statistics*, 39(4):859–879.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021b). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320.

Kelly, B. and Xiu, D. (2023). Financial Machine Learning. *Foundations and Trends® in Finance*, 13(3-4):205–363.

Kim, A., Muhn, M., and Nikolaev, V. V. (2024). Financial Statement Analysis with Large Language Models.

Lang, M. and Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3):110–135.

Leippold, M. (2023). Thus spoke GPT-3: Interviewing a large-language model on climate finance. *Finance Research Letters*, 53:103617.

Levy, B. (2024). Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2):221–247.

Li, F., Lundholm, R., and Minnis, M. (2013). A Measure of Competition Based on 10-K Filings. *Journal of Accounting Research*, 51(2):399–436.

Li, J. (2025). Ideology and Asset Pricing.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7):3265–3315.

Linn, M. and Weagley, D. (2019). Uncovering Financial Constraints.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Liu, Y., Sheng, J., and Wang, W. (2021). How Do Investors Value Technology in Cryptocurrency? Evidence from Textual Analysis. *Available at SSRN*, 3577208.

Lo, A. W. and Ross, J. (2024). Can ChatGPT Plan Your Retirement?: Generative AI and Financial Advice. *Harvard Data Science Review*, (Special Issue 5).

Lopez-Lira, A. (2023). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The Memorization Problem: Can We Trust LLMs' Economic Forecasts?

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, 69(4):1643–1671.

Loughran, T. and McDonald, B. (2020). Textual Analysis in Finance. *Annual Review of Financial Economics*, 12(1):357–375.

Lowry, M., Michaely, R., and Volkova, E. (2020). Information revealed through the regulatory process: Interactions between the SEC and companies ahead of their IPO. *The Review of Financial Studies*, 33(12):5510–5554.

Ludwig, J. and Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation*. *The Quarterly Journal of Economics*, 139(2):751–827.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large Language Models: An Applied Econometric Framework.

Lv, L. (2024). The Value of Information from Sell-side Analysts.

Mamaysky, H. (2024). News and Markets in the Time of COVID-19. *Journal of Financial and Quantitative Analysis*, 59(8):3564–3600.

Mamaysky, H., Shen, Y., and Wu, H. (2023). Credit Information in Earnings Calls.

Manela, A. (2014). The value of diffusing information. *Journal of Financial Economics*, 111(1):181–199.

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

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated Social Science: Language Models as Scientist and Subjects.

Mayew, W. J. and Venkatachalam, M. (2012). The Power of Voice: Managerial Affective States and Future Firm Performance. *The Journal of Finance*, 67(1):1–43.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mukherjee, A., Panayotov, G., and Shon, J. (2021). Eye in the sky: Private satellites and government macro data. *Journal of Financial Economics*, 141(1):234–254.

Murphy, K. M. and Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 20(1):88–97.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, USA. Association for Computational Linguistics.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Niederhoffer, V. (1971). The Analysis of World Events and Stock Prices. *Journal of Business*, 44(2):193–219.

Novy-Marx, R. and Velikov, M. (2024). AI-Powered (Finance) Scholarship.

Packalen, M. and Bhattacharya, J. (2015). New Ideas in Invention.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International economic review*, pages 221–247.

Pellegrino, Bruno (2022). Product Differentiation and Oligopoly: A Network Approach. *American Economic Review*.

Peress, J. (2014). The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes. *The Journal of Finance*, 69(5):2007–2043.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Roll, R. (1988). R2. *The Journal of Finance*, 43(3):541–566.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sarkar, S. K. (2024a). Economic Representations.

Sarkar, S. K. (2024b). StoriesLM: A Family of Language Models With Sequentially-Expanding Pretraining Windows.

Sarkar, S. K. and Vafa, K. (2024). Lookahead Bias in Pretrained Language Models.

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023). Firm-Level Climate Change Exposure. *The Journal of Finance*, 78(3):1449–1498.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Shah, A., Vithani, R., Gullapalli, A., and Chava, S. (2023). FiNER: Financial Named Entity Recognition Dataset and Weak-Supervision Model.

Steyvers, M. and Griffiths, T. (2007). Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*. Psychology Press.

Taddy, M. (2013). Multinomial Inverse Regression for Text Analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Taddy, M. (2015). Distributed Multinomial Regression. *Annals of Applied Statistics*, 9(3):1394–1414.

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. (2010). Does public financial news resolve asymmetric information? *The Review of Financial Studies*, 23(9):3520–3557.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Thawani, A., Pujara, J., Ilievski, F., and Szekely, P. (2021). Representing Numbers in NLP: A Survey and a Vision. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Veitch, V., Sridhar, D., and Blei, D. M. (2020). Adapting Text Embeddings for Causal Inference.

Whited, T. M. and Wu, G. (2006). Financial Constraints Risk. *The Review of Financial Studies*, 19(2):531–559.

Wu, D. A. (2024). Text-Based Measure of Supply Chain Risk Exposure. *Management Science*, 70(7):4781–4801.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). BloombergGPT: A Large Language Model for Finance.