# Language Modeling for the Future of Finance: A Survey into Metrics, Tasks, and Data Opportunities

**Nikita Tatarinov** ♠ ✉, **Siddhant Sukhani** ♠ ✉, **Agam Shah** ✉, **Sudheer Chava**
Georgia Institute of Technology, Atlanta, USA
♠ Indicates equal contribution
✉ Corresponding Authors: {ntatarinov3, ssukhani3, ashah482}@gatech.edu

## Abstract

Recent advances in language modeling have led to a growing number of papers related to finance in top-tier Natural Language Processing (NLP) venues. To systematically examine this trend, we review 374 NLP research papers published between 2017 and 2024 across 38 conferences and workshops, with a focused analysis of 221 papers that directly address finance-related tasks. We evaluate these papers across 11 quantitative and qualitative dimensions, with particular attention to evaluation practices, metric choices, dataset coverage, and reproducibility in a high-stakes applied LM domain. Our study identifies the following opportunities for NLP researchers: (i) expanding the scope of forecasting tasks; (ii) enriching evaluation with finance-specific metrics; (iii) leveraging multilingual and crisis-period datasets for robustness-oriented evaluation; and (iv) balancing PLMs with efficient or interpretable alternatives. We identify actionable directions supported by dataset and tool recommendations, with implications for both academic evaluation practices and industry deployment.

## 1 Introduction

Language modeling is a core method in natural language processing (NLP) for analyzing unstructured text (Peters et al., 2018; Brown et al., 2020). At the same time, finance has become one of NLP's primary application domains: as seen in Figure 5, the number of *finance-related papers in top-tier NLP venues* has been rising rapidly year over year. The tasks in these papers span general NLP problems on financial data, including sentiment analysis (Balakrishnan et al., 2022), information extraction (Huang et al., 2023), and summarization (Khanna et al., 2022), as well as finance problems addressed with NLP techniques, such as stock prediction (Jain and Agrawal, 2024), volatility forecasting (Niu et al., 2023), etc. The primary goal of our study is to **highlight the opportunities to advance NLP research when applied to finance**. This focus also makes finance a useful setting for studying how language models are evaluated beyond standard NLP benchmarks, since model usefulness in this domain depends on robustness, reliability, interpretability, and practical financial impact.

Our scope includes 374 papers published from 2017 to 2024 across 38 NLP conferences and workshops. After further filtering (Section 2), we retain 221 papers that directly address finance-related tasks and evaluate them across 11 quantitative and qualitative dimensions, including tasks, methodologies, datasets, evaluation metrics, and accessibility, allowing us to assess how evaluation and benchmarking practices have evolved in financial NLP. To the best of our knowledge, we present the first systematic study of finance-domain research in NLP venues, reducing subjective selection compared to prior surveys: existing surveys (Table 1) manually select the reviewed papers. In addition, some works are focusing on specific NLP approaches such as deep learning (Ozbayoglu et al., 2020) and Large Language Models (LLMs) (Nie et al., 2024; Li et al., 2023b), or on specific tasks such as sentiment analysis (Mishev et al., 2020).

Our analysis reveals not only valuable insights but also actionable directions for both research and practice with dataset and tool recommendations. Financial forecasting tasks, while well-established, leave room for exploration, especially in areas such as risk and macroeconomic prediction (Section 3). Financial evaluation metrics are gaining traction and could further improve the practical applicability of models (Section 4.1). The growing availability of temporally diverse (Sections 4.2, 4.3), multilingual and multimodal (Sections 4.4) datasets enables more robust, globally applicable models. Together, these findings connect finance-oriented NLP to broader questions in language model evaluation: what should be measured, which datasets



*Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 718–744
July 4, 2026 

Table 1: Comparison of previous Natural Language Processing surveys in finance based on their focus areas, number of papers reviewed, and analytical features. Due to the absence of systematic collection methods in most prior research work, entries marked with (A) in the "Review Years" column indicate the approximate range. We report the 374 papers manually screened before retaining 221 finance-focused papers for detailed analysis.

| Paper | Review Years | Papers Reviewed | Features | Area surveyed | Systematic Collection | Domain trends | Quantitative Analysis | Temporal Analysis |
|---|---|---|---|---|---|---|---|---|
| (Gao et al., 2021) | 1959 (A)-2020 | 87 | 5 | General Overview | × | × | × | × |
| (Liu, 2024) | 2022-2024 | 49 | 11 | General Overview | × | × | × | × |
| (Millo et al., 2024) | 2018-2023 | 30 | 1 | Methodologies | × | × | ✓ | × |
| (Chen et al., 2020b) | 2016-2019 | 62 | 3 | Financial Technology | × | × | × | × |
| (Chen et al., 2022a) | 2018-2022 | 38 | 2 | Financial Technology | × | ✓ | × | × |
| (Xing et al., 2017) | 1998 (A)-2016 | 127 | 4 | Financial Forecasting | × | ✓ | × | × |
| (Zhao et al., 2024a) | 2004 (A)-2024 | 146 | 1 | Large Language Models | × | × | ✓ | × |
| (Li et al., 2024) | 2020-2023 | 68 | 1 | Large Language Models | × | × | × | × |
| (Dong et al., 2024) | 2023-2024 | 206 | 2 | Large Language Models | × | × | ✓ | ✓ |
| (Nie et al., 2024) | 2019-2024 | 318 | 1 | Large Language Models | × | × | × | ✓ |
| (Lee et al., 2025) | 2018-2023 | 51 | 3 | Large Language Models | × | × | × | × |
| (Man et al., 2019) | 2004-2019 | 89 | 1 | Machine Learning | × | × | × | × |
| (Ozbayoglu et al., 2020) | 1998 (A)-2020 | 151 | 6 | Deep Learning | × | × | ✓ | ✓ |
| (Mishev et al., 2020) | 2003 (A)-2020 | 89 | 1 | Sentiment Analysis | × | ✓ | × | ✓ |
| **Language Modeling for the Future of Finance** | **2017-2024** | **374** | **11** | **Tasks, Methodologies, Data, Metrics, Code, Authorship, Funding** | ✓ | ✓ | ✓ | ✓ |

constitute meaningful benchmarks, and how academic metrics translate to real-world use.

## 2 Paper Extraction Process

To study how NLP has been applied to finance, we selected papers from 38 NLP venues, including ACL, NAACL, EMNLP, LREC, CoLM, COLING, and workshops like FinNLP and the Workshop on Economics and Natural Language Processing. As illustrated in Figure 1, we began by filtering papers that mentioned “finance” or “financial” in their abstracts (Mackenzie et al., 2018; Nogueira and Lin, 2020), enabling us to cast a wide net, allowing for a high volume of papers that could be manually postprocessed. Out of all such papers published from 1975 to the present, more than 94% appeared from 2017 onward, coinciding with the emergence of transformer-based models (Vaswani et al., 2017), which greatly expanded the scope of NLP applications (Devlin et al., 2019; Dai et al., 2019; Lewis et al., 2020), and we therefore use 2017 as a threshold year for our analysis.

This initial filtering returned 374 papers. Among them, 88 were tied to shared tasks (e.g., SemEval-2017 (Kar et al., 2017), FinCausal-2022 (Mondal et al., 2022)), and 65 missed an actual financial application (e.g. referenced financial resources (Sekeres et al., 2024; Ding and Riloff, 2018)). After removing these, we retained 221 papers to analyze.

Unlike previous surveys as shown in Table 1, which often emphasize qualitative observations (Gao et al., 2021; Liu, 2024; Millo et al., 2024), our study combines quantitative and qualitative methods for a broader view of the field. In comparison to other works, we categorized the selected papers by their primary focus into four task-based categories (Figure 2). This classification helps us examine trends in how NLP techniques are used to tackle finance-relevant problems and supports more detailed comparisons across types of tasks.

## 3 Task Distribution in NLP for Financial Applications

To understand NLP use in finance, we categorize papers into four task-based groups (Table 2), with their distribution shown in Figure 2.

**Financial Forecast (Category I)** papers cover predictive tasks such as stock price and volatility forecasting. While these tasks are well-studied, areas like economic forecasting (Arno et al., 2023), risk assessment (Zhou et al., 2020), and cryptocurrency prediction (Seroyizhko et al., 2022) receive less focus. These areas offer opportunities to expand the reach of predictive models, improving robustness and comparability (Section 4).

**Financial Resources (Category II)** papers often focus on dataset construction (Figure 7). Spanning annotated earnings calls, news, and speeches, these datasets support tasks like financial event extraction (Huang et al., 2024; Ju et al., 2023), fraud detection (Erben and Waldis, 2024; Wang et al., 2019), and annotation (Aguda et al., 2024; Khatuya et al., 2024). These tasks enhance financial data processing, while others remain underrepresented.

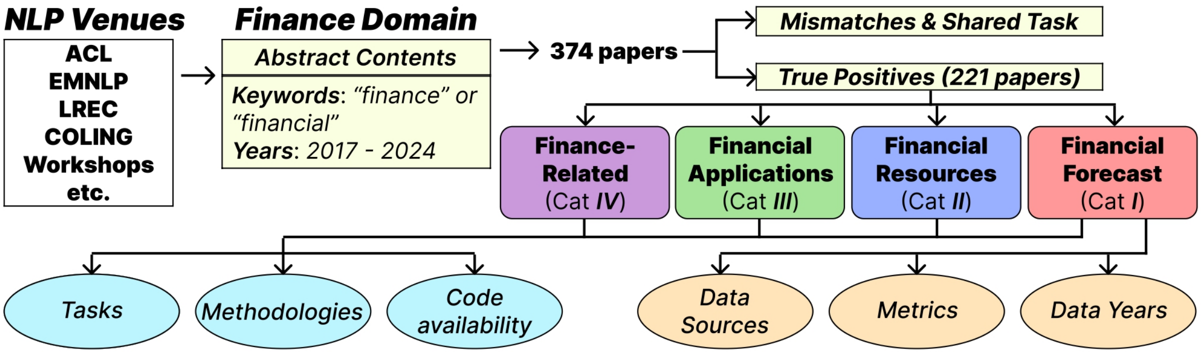


Figure 1: Overview of our paper selection process and analysis dimensions. We collected papers from a broad range of NLP venues using abstract-level keyword filtering, yielding 374 candidates. After removing mismatches and shared task papers, we retained 221 papers, categorized into four groups by their connection to financial tasks.

In the **Financial Applications (Category III)** group, sentiment analysis (Rodriguez Inserte et al., 2023), information extraction (Lior et al., 2024), and question answering (Kosireddy et al., 2024) are the dominant tasks. These methods are widely used to analyze earnings calls, reports, and market commentary, where investor sentiment and factual extraction are key inputs for decision-making(Chen et al., 2021; Zhu et al., 2021; Mukherjee et al., 2022; Qin and Yang, 2019). Recent efforts have aimed to improve QA models for handling financial text (Theuma and Shareghi, 2024; Liu et al., 2024; Mavi et al., 2023).

Table 2: Summary of task categories based on the specifics of their financial focus, ranging from direct prediction of financial outcomes to general NLP tasks with potential financial value.

| Category | Description |
|---|---|
| **Financial Forecast (Category I)** | Targets predicting financial events, including stock movements, volatility, bankruptcy, and currency exchange rates. |
| **Financial Resources (Category II)** | Covers tasks addressing finance-specific issues beyond prediction, such as constructing financial datasets, detecting fraud in finance-related documents, and extracting financial events. |
| **Financial Applications (Category III)** | Focuses on general ML/NLP tasks like sentiment analysis and information/relation extraction for financial datasets. |
| **Finance Related (Category IV)** | Covers tasks not applied to financial data or targeting financial problems, but potentially useful in finance, such privacy-preserving and explainable AI. |

Finally, among **Finance-Related (Category IV)** papers, the most studied areas are dataset construction and numerical reasoning. While these datasets are not finance-specific, they have potential applications in finance. For instance, fake news detection datasets (Vargas et al., 2021) can help reduce misinformation in markets. Numerical reasoning (Akhtar et al., 2023), important for understanding financial statements and assessing risk, is receiving more attention. However, other areas – such as explainable AI (XAI) (Klein and Walther, 2024) and privacy-preserving methods (Abbe et al., 2012) remain rarely explored, despite their relevance to secure and interpretable decision-making (Basu et al., 2021).

## 4 Potential in Financial Forecasting

**Category I** papers reveal several areas where forecasting models could be improved. Most studies use general ML metrics, leaving out finance-specific measures that better capture practical utility (Section 4.1). Crisis periods are rarely considered, limiting insights into how models behave under stress (Section 4.2). Finally, there is a strong U.S. and English-language bias in data, with limited adoption of global financial datasets (Section 4.3, 4.4). Expanding these areas of exploration could strengthen model robustness and support broader real-world applicability.

### 4.1 Advancing Evaluation with Finance Specific Metrics

As shown in Figure 3, **Category I** papers mostly rely on ML metrics such as accuracy, F1, or MSE (Sawhney et al., 2020b; Wu, 2020; Yangjia et al., 2022). While useful, these do not fully reflect financial performance. Finance-specific metrics offer

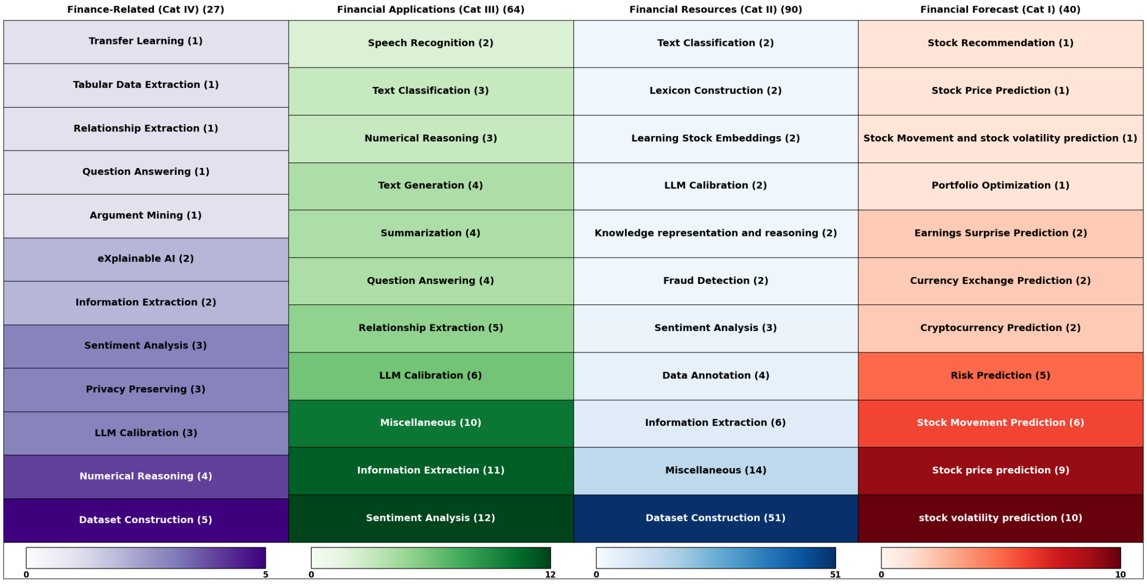


Figure 2: Distribution of primary tasks across categories. Each cell shows the task name and paper count (e.g., "Sentiment Analysis (3)"), with color gradients indicating frequency – darker shades represent more papers. "Miscellaneous" groups tasks that appear only once within Categories II and III.

risk-adjusted insights that ML measures overlook. Incorporating such metrics would improve *comparability across models and align research outcomes with practical needs* (Zhang et al., 2024; Zou et al., 2022; Sawhney et al., 2021c).

As highlighted by Tamar et al. (2012); Liu et al. (2022), the financial machine learning literature stresses that conventional statistical metrics such as mean squared error are insufficient for assessing trading strategies, since they overlook the dimensions of real-world profitability and risk management (Bailey et al., 2015; de Prado, 2018). Instead, financial metrics such as Sharpe Ratio, Maximum Drawdown, and Cumulative Return are essential because they reflect the model's risk-adjusted performance and robustness across market regimes (Lo, 2002; Krauss et al., 2017; Takahashi et al., 2009) as well as the applicability of the model to practical real world situations.

**Key Financial Metrics and Python Libraries** In financial machine learning, specialized metrics are essential for evaluating a model's real-world applicability. The *Sharpe Ratio* quantifies risk-adjusted return by comparing excess returns to volatility (Investopedia, 2025a), while *Maximum Drawdown (MDD)* captures the largest peak-to-trough loss, reflecting downside risk and robustness (Hayes, 2024). Similarly, *Cumulative Return* provides a direct benchmark by measuring total profit over a period (Investopedia, 2007). Several Python libraries facilitate the computation of these metrics within model-driven investment workflows: `PyPortfolioOpt` (Martin, 2021) offers portfolio optimization tools including Sharpe and risk-return analysis; `QuantLib` (QuantLib Team, 2000) supports pricing, drawdown, and advanced risk modeling; `Backtrader` (Backtrader Team, 2015) enables backtesting and performance evaluation of trading strategies, including NLP-driven pipelines; and `bt` (Morissette, 2025) streamlines prototyping and comparison of forecasting models with built-in financial metrics.

### 4.2 Strengthening Robustness by Incorporating Crisis Periods

Figure 4 shows that researchers use financial data dating back to 1993, when electronic filings became publicly available, with another rise in 2005 after the HTML filing mandate (U.S. SEC, 2025). However, most publications use post-2009 (the global financial crisis) and pre-2020 (the COVID-19 pandemic) data, thus missing the crisis periods in evaluation (Kim et al., 2022b; Fan et al., 2023; Yao et al., 2022). *Crisis periods are essential for evaluating model robustness, allowing stress testing* (Investopedia, 2025b; Hafiz et al., 2023; Aldasoro et al., 2025) and helping build forecasting models that remain reliable under instability.

For example within industrial applications of such models, LTCM's overdependence on stable market data and highly leveraged VaR models left it blind to tail risks, so when the Asian and Russian crises hit, spreads exploded, liquidity vanished, and the fund lost over 90% of its equity, sparking systemic risk concerns (Department of Land Economy, 2003; Federal Reserve History, 2025; Vanity Fair Staff, 1998). While there were data availability challenges prior to 2015 (Bloomberg L.P., 2015), the crisis years remain underused. In Section 4.3, we highlight valuable data sources, including those with crisis years data.

### 4.3 Opportunities in Data Coverage

Most forecasting models use common financial sources: stock prices, SEC filings, financial news, earnings calls. But many valuable datasets are either rarely used (e.g., Federal Reserve reports (Shah et al., 2023; Menzio et al., 2024a)) or completely absent from NLP research (e.g., shareholder letters). Public resources like FRED (Federal Reserve Bank of St. Louis, 2024) and Fama-French (Fama and French, 2024) offer rich but underexplored macro indicators. Incorporating these datasets into ablation studies and further analysis could help align the various findings with macroeconomic trends and market dynamics, improving model reliability (Chakraborty et al., 2016; Xu and Cohen, 2018a; Sawhney et al., 2020b). We highlight sources of valuable financial data, including those already used by the community and the other beneficial ones in Appendix G. For evaluation-oriented future work, especially relevant sources include FOMC communications, macroeconomic indicators, analyst reports, and non-U.S. regulatory filings, as they can support robustness checks, broader market coverage, and more realistic benchmark construction.

### 4.4 Data Concerns and Biases

**Data Accessibility Concerns** Despite the presence of a large variety of data sources, there are significant limitations in data accessibility. Legacy datasets described in Section 4.3 often become outdated, and APIs for accessing data may be discontinued, as the Yahoo Finance API (EODHD, 2024), or restricted by paywalls. Other essential databases, such as CRSP, require paid access and are unavailable to most individual researchers. These barriers severely limit reproducibility and access to reliable data, underscoring the need for open, regularly updated financial datasets for forecasting.

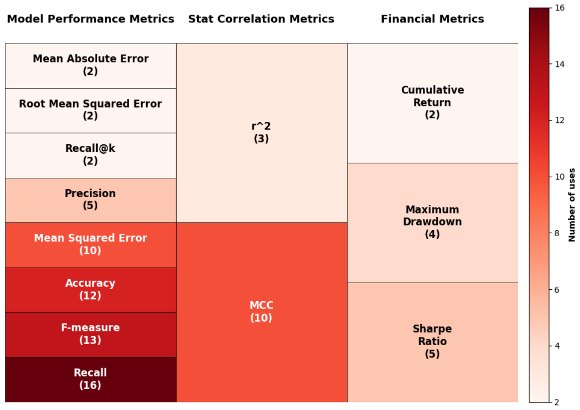


Figure 3: Distribution of evaluation metrics used in **Category I** papers. Most rely on ML-based metrics, while only a few financial metrics appear repeatedly.

**Data Language and Country Bias** As shown in the previous sections, financial NLP research is heavily skewed toward English-language data, particularly from U.S. markets (Chen et al., 2021; Reddy et al., 2024; Mukherjee et al., 2022). Most forecasting datasets rely on U.S. regulatory filings, earnings calls, and financial news, which is often centered on large-cap firms like those in the S&P 500. This narrow focus limits the applicability of models to non-English and non-U.S. contexts. Figure 7 highlights that while some multilingual datasets exist, they are mostly designed for language modeling rather than financial forecasting (Ding et al., 2014; Cortis et al., 2017). Very few datasets enable forecasting tasks in other languages, with rare exceptions like the bilingual CMIN corpus (Luo et al., 2023). *Expanding multilingual, market-diverse datasets is essential for building globally robust forecasting models* and mitigating systemic biases in NLP for finance, as also noted by (Jørgensen et al., 2023).

**Multimodal Data Scarcity** Despite significant progress in multimodal learning, the domain of NLP in Finance still lacks comprehensive datasets that integrate text, audio, and other critical financial modalities (Sawhney et al., 2020b, 2021a; Kaikaus et al., 2022). MAEC, a dataset aligning transcripts with audio from over 3,400 S&P 500 earnings calls (920+ hours of speech) for financial risk forecasting, and FinAudio, the first standardized benchmark for audio-first tasks with 400+ hours of financial audio for evaluating Audio-LLM capabilities, represent two of the key recent resources in financial audio research (Li et al., 2020a; Cao et al., 2025). Our survey indicates that while many ex-

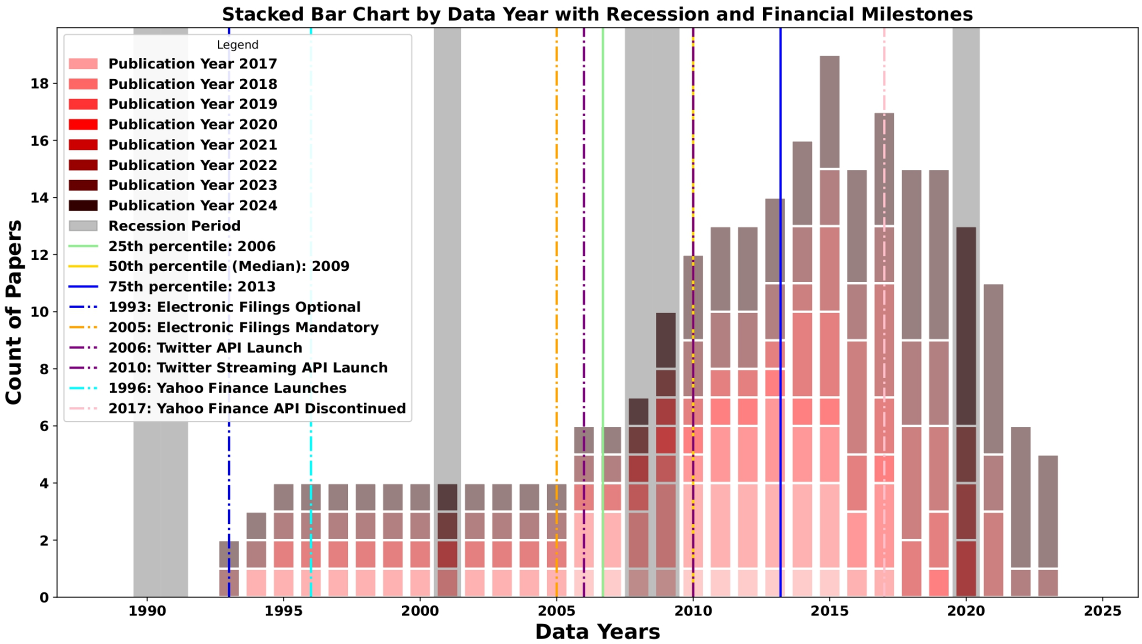


Figure 4: Data year distribution in financial forecasting papers, annotated with major financial events and infrastructure milestones. Highlights underuse of crisis periods despite their importance for model robustness.

isting datasets support exploration of acoustic and textual cues in financial forecasting, truly multimodal resources (those integrating visual data such as stock charts, presentation slides alongside structured financial metrics) remain very limited. Notable exceptions include Mathur et al. (2022b) and Fons et al. (2024), which begin to address this gap. This modality gap highlights the pressing need for richer, tri- or quad-modal datasets to better capture the multifaceted signals that influence market dynamics (Liang and et al., 2022; Li and et al., 2024).

# 5 Broader Insights for NLP in Finance

## 5.1 Complementing Foundation Models with Practical Alternatives

Based on our observations from our vast corpus of research, the methodological landscape in NLP for Finance has shifted decisively toward pretrained language models (PLMs) and large language models (LLMs). We group existing approaches into several broad categories, which allows us to track adoption trends and the rise of PLMs, beginning with the use of BERT-based models in finance around 2019 (Araci, 2019a; Yang et al., 2020c), and accelerating rapidly with the emergence of general-purpose LLMs like ChatGPT (Wu et al., 2023; Gao et al., 2023). However, NLP tasks in Finance often pose domain-specific constraints such as interpretability, regulatory transparency, data scarcity, and the need for low-latency systems (Luo et al., 2018; Maia et al., 2018). These factors make non-PLM methods (such as rule-based classifiers, feature-engineered models, and domain-specific embeddings) not only viable but often preferable in production environments (Shah et al., 2023; Lopez et al., 2021). The rest of this section revisits key methodological alternatives to PLMs, each addressing practical and conceptual needs that remain critical in industrial financial tasks.

**Latency and Scalability** Classical NLP pipelines are far more resource-efficient than LLMs (Strubell et al., 2019). In time-time contexts like news analytics or alert systems, lightweight models (e.g., logistic regression or small CNNs) deliver insights instantly and can scale to massive corpora without requiring GPU infrastructure (Zhai et al., 2019). Systems like RavenPack (Cornez, 2018) and Refinitiv News Analytics (TRNA) (Waeosri, 2020) use rule-based NLP and horizontally scalable architectures to deliver structured news data with sub-second latency, supporting real-time trading and high-throughput analysis.

**Interpretability and Compliance** Lexicon- and rule-based methods remain essential in regulated settings (Moreno-Ortiz et al., 2020; Du et al., 2023). Financial sentiment lexicons (e.g., Loughran-McDonald) (Loughran and McDonald, 2011) and keyword rules are transparent, auditable, and often outperform neural models in compliance or risk-flagging tasks (Hosseini et al., 2018). Lexicon-based surveillance tools, such as those used in communications monitoring (NICE Actimize, 2023) and financial crime compliance (Risk & Compliance, 2021), are widely adopted for their transparency and ease of audit, allowing firms to justify alerts with specific keywords.

**Information Extraction and Knowledge Graphs** Named-entity and relation extraction systems power knowledge graphs and schema-driven applications (Elhammadi et al., 2020; Hamad et al., 2024). These systems offer precision, structure, and explainability, which are ideal for fraud detection, Know Your Customer (KYC), and regulatory alignment, where PLMs often lack controllability (Szarvas et al., 2007; Tarnopolski et al., 2019). Financial institutions use rule-based NLP for entity and relation extraction to build and update knowledge graphs in KYC and Anti-Money Laundering (AML) workflows, as seen in solutions from KYCHub (2025) and TigerGraph (2024), enabling automated risk monitoring and explainable link discovery.

**Data Efficiency and Domain Adaptability** Feature-based pipelines using TF-IDF, static embeddings, and classical classifiers (SVMs, XGBoost) work well with limited data(Wang and Manning, 2012). They enable explicit domain knowledge integration (such as tagging financial terms or numerical cues) for credit scoring, bankruptcy prediction, and event detection (Li et al., 2018; Khandani et al., 2021; Alanis et al., 2022). Financial firms often use TF-IDF (Simha, 2021) and XGBoost (Kamau, 2024) to build effective models with limited labeled data without requiring costly infrastructure.

**Structure and Summarization** Unsupervised techniques like LDA and extractive summarization remain valuable for discovering latent topics or summarizing dense financial documents (Filippova et al., 2009; Agrawal et al., 2021). These methods are fast, interpretable, and competitive on factual corpora like earnings calls or filings (Blei et al., 2003; Nenkova and McKeown, 2011). Companies like Bank of America (Phoenix Strategy Group, 2025) and American Express (Arora and Radhakrishnan, 2020) apply topic modeling and extractive summarization to earnings calls and filings, using unsupervised or minimally supervised methods for reliable and traceable outputs.

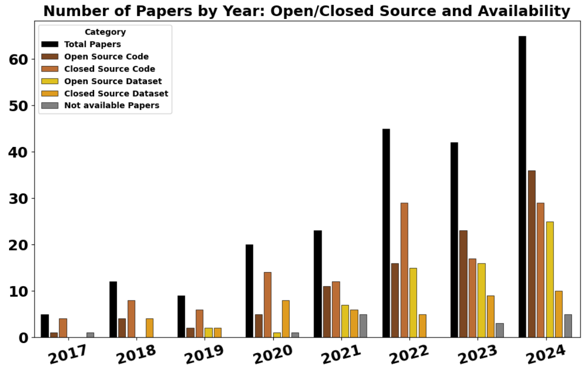


Figure 5: Trends in code and dataset availability, highlighting the shift toward open-source practices and the growing accessibility of NLP resources for finance.

## 5.2 Implications of Shift Toward General-Purpose Models

General-purpose language models have rapidly become the default choice in NLP applied to Finance, with models like RoBERTa (Liu et al., 2019), GPT-4 (Achiam et al., 2023), and LLaMA-2 (Touvron et al., 2023) adopted far more widely and quickly than domain-specific alternatives. As Figure 7 shows, even earlier versions of FinBERT (Araci, 2019b; Yang et al., 2020b; Liu et al., 2020; Huang et al., 2022) are used more than newer ones, and recent finance-tuned models like FinGPT (Yang et al., 2023) remain rare. For practitioners, this means *many models are optimized for benchmarks, not for the realities of finance*, where interpretability, latency, or regulation often matter more than raw accuracy.

This shift is also reflected in the decline of custom architectures. Before 2022, task-specific models were more common: TagOp for TAT-QA (Zhu et al., 2021), HyBrider for HybridQA (Chen et al., 2020c), and MDRM (Qin and Yang, 2019), HTML (Yang et al., 2020a), and HAN (Hu et al., 2021) for volatility prediction (Niu et al., 2023; Mathur et al., 2022a). Today, such efforts are rare. However, as Section 5.1 discusses, real-world financial applications often require design trade-offs, such as transparency for compliance or fast inference for deployment, which off-the-shelf LLMs do not handle well. *This move away from custom models narrows the solution space* when financial tasks still require methodological flexibility.

Table 3: Summary of trends and research directions in NLP research applied to finance, highlighting areas of focus, methodological shifts, and key opportunities for future exploration.

| Criteria | Trends and Observations | Potential Opportunities and Recommendations |
|---|---|---|
| **NLP Tasks in Finance** | Sentiment analysis, information extraction, and question answering are the most frequently addressed tasks. Forecasting is mostly centered around stock and volatility prediction. | Underexplored areas such as explainability, privacy-preserving methods, and tasks like bankruptcy or cryptocurrency prediction present valuable directions for future work (Section 3). |
| **Evaluation Metrics** | Most studies rely on ML metrics such as accuracy and MSE, which do not reflect financial performance. | Incorporating financial metrics like Sharpe Ratio or Maximum Drawdown would improve the practical relevance and comparability of predictive models (Section 4.1). |
| **Crisis Periods** | Few studies include data from financial crises such as 2008–2009 or 2020–2021, focusing instead on stable periods. | Including data from volatile periods can support stress-testing and help build models more resilient to real-world fluctuations (Section 4.2). |
| **Data Selection and Bias** | Most studies use English-language datasets and U.S. financial sources. While there is growing diversity in the types of data used (e.g., news, filings, social media), much of it is static, with limited updates or adaptation to changing financial contexts. | Developing multilingual and globally representative datasets (Section 4.4), along with maintaining and updating existing ones (Section 4.3), would support better generalization and long-term applicability of models across financial domains. |
| **PLM/LLM Adoption** | General-purpose PLMs and LLMs have largely replaced finance-specific and custom architectures. At the same time, the use of statistical NLP and conventional ML continues to decline. | Revisiting finance-specific models and exploring alternative methods such as graph-based learning or hybrid statistical models could offer improvements for financial tasks (Sections 5.1, 5.2). |
| **Open Accessibility** | Code and dataset sharing has become more common, especially after 2021, enhancing transparency and reproducibility. | Open access should be a standard in NLP-for-finance research when legally feasible. Maintaining functional and up-to-date repositories is key for long-term reproducibility (Section 5.3). |

### 5.3 Enabling Reproducibility through Open Resources

As we witness an accelerated adoption of language models, the issue of open accessibility in NLP research, especially in finance, becomes more relevant. As shown in Figure 5, open-source practices have become more common in recent years. Before 2021, closed-source code dominated, but the trend has shifted toward sharing code and datasets.

This move enhances transparency, reproducibility, and collaborative progress in NLP research applied to finance (Whited, 2023). Despite this, some papers still include inactive links or empty repositories, limiting reproducibility. At the same time, dataset availability has steadily increased since 2017, pointing to growing awareness around open data. *Maintaining functional, up-to-date repositories remains essential to support meaningful benchmarking and model development*.

## 6 Conclusion & Discussion

Drawing on Table 3, our survey shows that sentiment analysis, information extraction, and forecasting have driven NLP in finance, yet critical gaps remain: finance-specific evaluation metrics, stress-testing on crisis data, and truly global, multilingual datasets. These findings underscore both the field's progress and the work ahead to build resilient, transparent, and inclusive financial NLP solutions. More broadly, these gaps position finance as a concrete case study for evaluation as measurement beyond raw model capability, where success depends not only on predictive accuracy but also on robustness, reproducibility, interpretability, and deployment relevance.

**Implications for NLP and Language Modeling Researchers** The application of LMs to financial domains presents challenges that differ from traditional NLP benchmarks. Our findings suggest that domain-specific requirements, such as interpretability, latency, and compliance, make classical or hybrid approaches highly relevant. The lack of financial evaluation metrics, underuse of crisis-period data, and limited attention to multilingual and global financial corpora signal areas where NLP researchers can contribute meaningfully. Moreover, the scarcity of domain-adapted models points to missed opportunities for more effective domain transfer. As a result, our observations indicate that future work should prioritize robustness, transparency, and real-world relevance over benchmark performance alone, , aligning finance-oriented NLP with broader discussions on evaluation validity, benchmark documentation, and reproducible assessment.

**Considerations for Financial Practitioners** For finance professionals and institutions considering NLP solutions, this survey provides guidance on selecting appropriate tools. Off-the-shelf LLMs may appear attractive, but their performance in finance-specific contexts is often limited by lack of customization, interpretability, or latency guarantees. Classical and lexicon-based methods, when tailored to regulatory or operational constraints, can outperform black-box models in compliance or auditing settings. Practitioners should also evaluate model performance using finance-aligned metrics and test under historical stress scenarios. Finally, leveraging multilingual datasets and broadening market coverage can help mitigate geographic and systemic biases in automated decision-making. These considerations reflect the broader challenge of navigating the gap between academic metrics and real-world impact in language model evaluation.

## Limitations

For readability, we standardize task names and merge closely related variants when counting. When a paper spans multiple areas, we assign a single primary tag to keep statistics interpretable; a multi-tag alternative would yield slightly different totals. These are intentional choices to aid comparability. As for venue coverage, our corpus intentionally focuses on NLP conferences and workshops, so we therefore exclude adjacent venues where relevant work may also appear. Adding these venues, as well as adding more keywords, would broaden the survey, but would also add noise to an already large corpus that required manual filtering. Thus, our findings should be read as a systematic survey of finance-related work in NLP venues, not an exhaustive survey of all financial NLP research. Finally, our survey does not aim to exhaustively cover the most recent FinLLM or agent-based finance work, which evolves quickly and often appears outside the NLP venues we study. Nevertheless, the gaps we identify around finance-specific evaluation metrics and stress-testing on crisis-period data remain relevant considerations for these newer directions.

## Ethical Considerations

This study does not assess the merit or quality of individual papers. We do not suggest that any category or method is inherently better than others. Our goal is to map the volume and distribution of research applying NLP methods to finance, and to offer a foundation for further study, without making judgments about the value of specific approaches.

## References


Emmanuel A. Abbe, Amir E. Khandani, and Andrew W. Lo. 2012. Privacy-preserving methods for sharing financial risk exposures. *American Economic Review*, 102(3):65–70.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Angus Addlesee. 2024. Grounding LLMs to in-prompt instructions: Reducing hallucinations caused by static pre-training knowledge. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, pages 1–7, Torino, Italia. ELRA and ICCL.

Yash Agrawal, Vivek Anand, Manish Gupta, S. Arunachalam, and Vasudeva Varma. 2021. Goal-directed extractive summarization of financial reports. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*.

Toyin D. Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, and Charese Smiley. 2024. Large language models as financial data annotators: A study on effectiveness and efficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10124–10145, Torino, Italia. ELRA and ICCL.

Noujoud Ahbali, Xinyuan Liu, Albert Nanda, Jamie Stark, Ashit Talukder, and Rupinder Paul Khandpur. 2022. Identifying Corporate Credit Risk Sentiments from Financial News. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 362–370, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Maximilian Ahrens and Michael McMahon. 2021. Extracting economic signals from central bank speeches. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 93–114, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association* for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Emmanuel Alanis, Sudheer Chava, and Agam Shah. 2022. Benchmarking machine learning models to predict corporate bankruptcy. *arXiv preprint arXiv:2212.12051*.

Imad Aldasoro and 1 others. 2025. Predicting financial market stress with machine learning. Technical report, Bank for International Settlements (BIS).

Dogu Araci. 2019a. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Dogu Araci. 2019b. Finbeunknownrt: Financial sentiment analysis with pre-trained language models. *Preprint*, arXiv:1908.10063.

Henri Arno, Klaas Mulier, Joke Baeck, and Thomas Demeester. 2023. From numbers to words: Multimodal bankruptcy prediction using the ECL dataset. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 11–21, Bali, Indonesia. Association for Computational Linguistics.

Piyush Arora and Priya Radhakrishnan. 2020. AMEX AI-labs: An investigative study on extractive summarization of financial documents. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 137–142, Barcelona, Spain (Online). COLING.

Backtrader Team. 2015. backtrader: Python backtesting & trading library. https://www.backtrader.com/. Open-source project (GPL-3.0), accessed 2025-07-21.

David H. Bailey, Jonathan Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2015. The probability of backtest overfitting. *Journal of Computational Finance*. Forthcoming.

Sudhandar Balakrishnan, Yihao Fang, and Xiaodan Zhu. 2022. Exploring robustness of prefix tuning in noisy data: A case study in financial sentiment analysis. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 78–88, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, and Zumrut Muftuoglu. 2021. Privacy enabled financial text classification using differential privacy and federated learning. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 50–55, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bloomberg L.P. 2015. 2015 data for good exchange: Data science and its impact on society. Accessed: 2025-03-05.

Bloomberg L.P. 2025a. Bloomberg news api / data subscriptions. https://www.bloomberg.com/subscriptions/uspaid. Accessed: 2025-07-21.

Bloomberg L.P. 2025b. Bloomberg terminal: Real-time data, research, and analytics. https://www.bloomberg.com/professional/products/bloomberg-terminal/. Accessed: 2025-07-21.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

Christopher Burger, Lingwei Chen, and Thai Le. 2023. “are your explanations reliable?” investigating the stability of LIME in explaining text classifiers by marrying XAI and adversarial attack. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12831–12844, Singapore. Association for Computational Linguistics.

Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao yang Liu, Koduvayur Subbalakshmi, Meikang Qiu, Sophia Ananiadou, and Jian-Yun Nie. 2025. Finaudio: A benchmark for audio large language models in financial applications. *Preprint*, arXiv:2503.20990.

Tanmoy Chakraborty and 1 others. 2016. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Yash Chaudhary. 2020. Stock-market sentiment dataset.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020a. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110, Marseille, France. European Language Resources Association.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020b. Nlp in fintech applications: Past, present and future. *Preprint*, arXiv:2005.01320.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022a. An overview of financial technology innovation. In *Companion Proceedings of the Web Conference 2022*, WWW ’22, page 572–575, New York, NY, USA. Association for Computing Machinery.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaeyoung Choe, Keonwoong Noh, Nayeon Kim, Seyun Ahn, and Woohwan Jung. 2023. Exploring the impact of corpus diversity on financial pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2101–2112, Singapore. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2022. Incorporating stock market signals for Twitter stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4074–4091, Dublin, Ireland. Association for Computational Linguistics.

Jason Cornez. 2018. Market-moving event detection: Latest innovations. `https://www.ravenpack.com/blog/event-detection-innovations/`. Accessed: 2025-07-21.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Marcos López de Prado. 2018. *Advances in Financial Machine Learning*. John Wiley & Sons.

University of Cambridge Department of Land Economy. 2003. Long-term capital management: Lessons from a collapse. Technical report, University of Cambridge. Accessed: 2025-07-21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haibo Ding and Ellen Riloff. 2018. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, New Orleans, Louisiana. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425, Doha, Qatar. Association for Computational Linguistics.

Mengming Michael Dong, Theophanis C. Stratopoulos, and Victor Xiaoqi Wang. 2024. A scoping review of chatgpt research in accounting and finance. *International Journal of Accounting Information Systems*, 55:100715.

Kangning Du, Erik Cambria, and 1 others. 2023. Finsenticnet: A concept-level lexicon for financial sentiment analysis. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, Mexico City, Mexico. IEEE.

Hanyu Duan, Yi Yang, and Kar Yan Tam. 2021. Learning numeracy: A simple yet effective number embedding approach using knowledge graph. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2597–2602, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarah Elhammadi, Laks V.S. Lakshmanan, Raymond Ng, Michael Simpson, Baoxing Huai, Zhefeng Wang, and Lanjun Wang. 2020. A high precision pipeline for financial knowledge graph construction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 967–977, Barcelona, Spain (Online). International Committee on Computational Linguistics.

EODHD. 2024. Exploring alternatives to the yahoo finance api. `https://medium.com/@eodhd/exploring-alternatives-to-the-yahoo-finance`

-api-aa0415085cc5. API discontinued and paywalled CSV access reported via Reddit and verified; Accessed: 2025-07-21.

Stefan Erben and Andreas Waldis. 2024. ScamSpot: Fighting financial fraud in Instagram comments. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–81, St. Julians, Malta. Association for Computational Linguistics.

Eurostat. 2025. Eurostat database. https://ec.europa.eu/eurostat/data/database. [Accessed: September 30, 2025].

FactSet Insight. 2025. Company events transcript datafeed. https://insight.factset.com/resources/at-a-glance-document-distributor-xml-company-events-transcript-datafeed. Accessed: 2025-07-21.

Eugene F. Fama and Kenneth R. French. 2024. Fama-french data library. Accessed: 2024-03-05.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. 2023. Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7522–7529.

Federal Reserve Bank of St. Louis. 2024. FRED, Federal Reserve Economic Data. Accessed: 2024-03-05.

Federal Reserve History. 2025. Near failure of long-term capital management. https://www.federalreservehistory.org/essays/ltcm-near-failure. Accessed: 2025-07-21.

Federal Reserve System. 2025. Fomc historical materials: Policy statements and meeting transcripts. https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm. Accessed: 2025-07-21.

Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2009. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 246–254, Athens, Greece. Association for Computational Linguistics.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. 2024. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21598–21634, Miami, Florida, USA. Association for Computational Linguistics.

Ruizhuo Gao, Zeqi Zhang, Zhenning Shi, Dan Xu, Weijuan Zhang, and Dewei Zhu. 2021. A review of natural language processing for financial technology. In *International Symposium on Artificial Intelligence and Robotics 2021*, volume 11884, page 118840U. International Society for Optics and Photonics, SPIE.

Tianyu Gao and 1 others. 2023. Large language models in financial nlp: A survey. *arXiv preprint arXiv:2306.01511*.

Google News. 2025. Google news aggregator. https://news.google.com/home. Accessed: 2025-07-21.

Yue Guo, Chenxi Hu, and Yi Yang. 2023. Predict the future from the past? on the temporal data distribution shift in financial sentiment classifications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1029–1038, Singapore. Association for Computational Linguistics.

Faizal Hafiz, Jan Broekaert, and Akshya Swain. 2023. Evolution of neural architectures for financial forecasting: A note on data incompatibility during crisis periods. *arXiv preprint arXiv:2311.14604*.

Hassan Hamad, Abhinav Kumar Thakur, Nijil Kolleri, Sujith Pulikodan, and Keith Chugg. 2024. FIRE: A dataset for financial relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3628–3642, Mexico City, Mexico. Association for Computational Linguistics.

Adam Hayes. 2024. Maximum drawdown (mdd): Definition and formula. https://www.investopedia.com/terms/m/maximum-drawdown-mdd.asp. Accessed: 2025-07-21.

Seyed Hosseini, Mohammad Reza Mousavi, Pooya Amirian, and ... 2018. Auditnlp: Towards auditable and interpretable nlp systems in finance. In *Proceedings of the 2nd Workshop on Interpretability and Robustness in Natural Language Processing*, page 28–35.

Yongli Hu, Puman Chen, Tengfei Liu, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2021. Hierarchical attention transformer networks for long document classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Allen Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text†. *Contemporary Accounting Research*, 40.

Ta-wei Huang, Jia-huei Ju, Yu-shiang Huang, Cheng-wei Lin, Yi-shyuan Chiang, and Chuan-ju Wang. 2023. FISH: A financial interactive system for signal highlighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 50–56, Dubrovnik, Croatia. Association for Computational Linguistics.

Yusheng Huang, Ning Hu, Kunping Li, Nan Wang, and Zhouhan Lin. 2024. Extracting financial events from raw texts via matrix chunking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7035–7044, Torino, Italia. ELRA and ICCL.

International Monetary Fund. 2025. Imf data. https://data.imf.org/. [Accessed: September 30, 2025].

Investopedia. 2007. Cumulative return: Definition, calculation, and example. https://www.investopedia.com/terms/c/cumulativereturn.asp. Accessed: 2025-07-21.

Investopedia. 2025a. Sharpe Ratio Definition. Accessed: 2025-03-05.

Investopedia. 2025b. Stress Testing Definition. Accessed: 2025-03-05.

Jainendra Kumar Jain and Ruchit Agrawal. 2024. FB-GAN: A novel neural sentiment-enhanced model for stock price prediction. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 85–93, Torino, Italia. Association for Computational Linguistics.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Jia-Huei Ju, Yu-Shiang Huang, Cheng-Wei Lin, Che Lin, and Chuan-Ju Wang. 2023. A compare-and-contrast multistage pipeline for uncovering financial signals in financial reports. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14307–14321, Toronto, Canada. Association for Computational Linguistics.

Kaggle Contributor. 2021. Reddit cryptocurrency data for august 2021. https://www.kaggle.com/datasets/pavellexyr/Reddit-cryptocurrency-data-for-august-2021. Accessed: 2025-07-21.

Kaggle Contributor. 2025. Sentiment140 dataset with 1.6 million tweets. https://www.kaggle.com/datasets/kazanova/sentiment140. Accessed: 2025-07-21.

Jamshed Kaikaus, Jessen L. Hobson, and Robert Brunner. 2022. Truth or fiction: Multimodal learning applied to earnings calls. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. Multimodal (text+audio) modeling for restatement and market movement prediction.

Isaak Kamau. 2024. When to use xgboost over llm: A case study in stock market analysis. https://medium.com/@isaakmwangi2018/when-to-use-xgboost-over-llm-a-case-study-in-stock-market-analysis-2f46571ad536. Accessed: 2025-07-21.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2017. RiTUAL-UH at SemEval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 877–882, Vancouver, Canada. Association for Computational Linguistics.

Amir Khandani, Adlai Kim, and Andrew Lo. 2021. A practical machine learning approach to credit scoring. *Journal of Financial Data Science*, 3(2):55–72.

Urvashi Khanna, Samira Ghodratnama, Diego Molla, and Amin Beheshti. 2022. Transformer-based models for long document summarisation in financial domain. In *Financial Narrative Processing Workshop (4th: 2022)*, pages 73–78. European Language Resources Association (ELRA).

Subhendu Khatuya, Rajdeep Mukherjee, Akash Ghosh, Manjunath Hegde, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2024. Parameter-efficient instruction tuning of large language models for extreme financial numeral labelling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7391–7403, Mexico City, Mexico. Association for Computational Linguistics.

Donggyu Kim, Garam Lee, and Sungwoo Oh. 2022a. Toward privacy-preserving text embedding similarity with homomorphic encryption. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 25–36, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022b. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*.

Tony Klein and Thomas Walther. 2024. Advances in explainable artificial intelligence (xai) in finance. *Finance Research Letters*, 70:106358.

Tagore Rao Kosireddy, Jeffrey David Wall, and Evan Lucas. 2024. Exploring the readiness of prominent small language models for the democratization of financial literacy. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 124–149, Miami, Florida, USA. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Jonas Sebastian Krauss, Xiaotong A. Do, and Nicolas Huck. 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702.

KYCHub. 2025. What is adverse media screening? `https://www.kychub.com/blog/adverse-media-screening/`. Accessed: 2025-07-21.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020a. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020b. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3063–3070, New York, NY, USA. Association for Computing Machinery.

Qi Li, Charles Lim, and Jianhua Huang. 2018. Credit scoring with historical bank data: Classical vs. ml-based approaches. *Journal of Banking and Finance*, 90:128–145.

X. Li and et al. 2024. Time-MMD: A multi-domain multimodal dataset for time series analysis. In *NeurIPS 2024 Datasets and Benchmarks Track*.

Xurui Li, Yue Qin, Rui Zhu, Tianqianjin Lin, Yongming Fan, Yangyang Kang, Kaisong Song, Fubang Zhao, Changlong Sun, Haixu Tang, and Xiaozhong Liu. 2023a. STINMatch: Semi-supervised semantic-topological iteration network for financial risk detection via news label diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9304–9315, Singapore. Association for Computational Linguistics.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA. Association for Computing Machinery.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2024. Large language models in finance: A survey. *Preprint*, arXiv:2311.10723.

Weixin Liang and et al. 2022. Understanding the modality gap in multi-modal contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gili Lior, Yoav Goldberg, and Gabriel Stanovsky. 2024. Leveraging collection-wide similarities for unsupervised document structure extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9538–9550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yi-Ting Liou, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Dynamic graph transformer for implicit tag recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1426–1431, Online. Association for Computational Linguistics.

Chuang Liu, Junzhuo Li, and Deyi Xiong. 2023. Tab-CQA: A tabular conversational question answering dataset on financial reports. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 196–207, Toronto, Canada. Association for Computational Linguistics.

Jiaxin Liu, Yi Yang, and Kar Yan Tam. 2024. Beyond surface similarity: Detecting subtle semantic shifts in financial narratives. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2641–2652, Mexico City, Mexico. Association for Computational Linguistics.

Junhua Liu. 2024. A survey of financial ai: Architectures, advances and open challenges. *Preprint*, arXiv:2411.12747.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhaoran Wang, and Jian Guo. 2022. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *NeurIPS 2022 Datasets and Benchmarks Track*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Andrew W. Lo. 2002. The statistics of sharpe ratios. *Financial Analysts Journal*, 58(4):36–52.

Gustavo Lopez and 1 others. 2021. Rule-based and feature-engineered approaches for nlp in finance: A practical guide. *Journal of Financial Data Science*, 3(2):21–38.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. 2023. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12164–12176, Toronto, Canada. Association for Computational Linguistics.

Liang Luo, Dong Li, Shoushan Liu, Guodong Wu, and Bin Li. 2018. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4244–4250.

Joel M. Mackenzie, J. Shane Culpepper, Roi Blanco, Matt Crane, Charles L. A. Clarke, and Jimmy Lin. 2018. Query driven algorithm selection in early stage retrieval. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 396–404.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018 (WWW Companion)*, pages 1941–1942. ACM.

Xiliu Man, Tong Luo, and Jianwu Lin. 2019. Financial sentiment analysis(fsa): A survey. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, pages 617–622.

Robert Andrew Martin. 2021. Pyportfolioopt: portfolio optimization in python. *Journal of Open Source Software*, 6(61):3066.

Puneet Mathur, Mihir Goyal, Ramit Sawhney, Ritik Mathur, Jochen Leidner, Franck Dernoncourt, and Dinesh Manocha. 2022a. DocFin: Multimodal financial prediction and bias mitigation using semi-structured documents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1933–1940, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022b. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2276–2285, New York, NY, USA. Association for Computing Machinery.

Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. 2023. Retrieval-augmented chain-of-thought in semi-structured domains. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 178–191, Singapore. Association for Computational Linguistics.

Martina Menzio, Davide Paris, and Elisabetta Fersini. 2024a. Unveiling currency market dynamics: Leveraging federal reserve communications for strategic investment insights. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 94–102, Torino, Italia. Association for Computational Linguistics.

Martina Menzio, Davide Paris, and Elisabetta Fersini. 2024b. Unveiling currency market dynamics: Leveraging federal reserve communications for strategic investment insights. In *Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP*, pages 94–102. ELRA.

Denisa Millo, Blerina Vika, and Nevila Baci. 2024. Integrating natural language processing techniques of text mining into financial system: Applications and limitations. *Preprint*, arXiv:2412.20438.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. 2020. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131662–131682.

Joydeb Mondal, Nagaraj Bhat, Pramir Sarkar, and Shahid Reza. 2022. ExpertNeurons at FinCausal 2022 task 2: Causality extraction for financial documents. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 128–130, Marseille, France. European Language Resources Association.

Antonio Moreno-Ortiz, Javier Fernandez-Cruz, and Chantal Pérez Chantal Hernández. 2020. Design and evaluation of SentiEcon: a fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5065–5072, Marseille, France. European Language Resources Association.

Philippe Morissette. 2025. bt: Flexible backtesting framework for python. `https://pmorissette.`

github.io/bt/. Accessed: 2025-07-21; MIT License; GitHub repo starred 2.6k times :contentReference[oaicite:0]index=0.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. A survey of text summarization techniques. *Foundations and Trends in Information Retrieval*, 5(2):103–233.

NICE Actimize. 2023. Smarter communications surveillance with ai and nlp. https://info.niceactimize.com/rs/338-EJP-431/images/Smarter_Communications_Surveillance_with_AI_White_Paper.pdf?version=0. Accessed: 2025-07-21.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *Preprint*, arXiv:2406.11903.

Hao Niu, Yun Xiong, Xiaosu Wang, Wenjing Yu, Yao Zhang, and Weizu Yang. 2023. KeFVP: Knowledge-enhanced financial volatility prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11499–11513, Singapore. Association for Computational Linguistics.

Rodrigo Nogueira and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Organisation for Economic Co-operation and Development. 2025. Oecd data. https://data.oecd.org/. [Accessed: September 30, 2025].

Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. 2020. Deep learning for financial applications : A survey. *Applied Soft Computing*, 93:106384.

Huzaifa Pardawala, Siddhant Sukhani, Agam Shah, Veer Kejriwal, Abhishek Pillai, Rohan Bhasin, Andrew DiBiasio, Tarun Mandapati, Dhruv Adha, and Sudheer Chava. 2025. Subjective-qa: Measuring subjectivity in earnings call transcripts' qa through six-dimensional feature analysis. *Preprint*, arXiv:2410.20651.

Denis Peskoff, Adam Visokay, Sander Schulhoff, Benjamin Wachspress, Alan Blinder, and Brandon M. Stewart. 2023. Gpt deciphering fedspeak: Quantifying dissent among hawks and doves. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6529–6539.

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

Phoenix Strategy Group. 2025. Ai-driven topic modeling for financial trends. https://www.phoenixstrategy.group/blog/ai-driven-topic-modeling-for-financial-trends. Accessed: 2025-07-21.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

QuantLib Team. 2000. Quantlib: Open-source software for quantitative finance. https://www.quantlib.org/docs.shtml. Accessed: 2025-07-21; initial release December 11, 2000; licensed under modified BSD.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

Risk & Compliance. 2021. Lexicon surveillance is not dead. https://riskandcompliancemagazine.com/lexicon-surveillance-is-not-dead. Accessed: 2025-07-21.

Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. 2023. Large language model adaptation for financial sentiment analysis. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 1–10, Bali, Indonesia. Association for Computational Linguistics.

Subhro Roy and Dan Roth. 2018. Mapping to declarative knowledge for word problem solving. *Transactions of the Association for Computational Linguistics*, 6:159–172.

Yunxin Sang and Yang Bao. 2022. DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1623–1633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021a. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762, Online. Association for Computational Linguistics.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020b. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, Online. Association for Computational Linguistics.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021b. FAST: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2164–2175, Online. Association for Computational Linguistics.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021c. Quantitative day trading from natural language using reinforcement learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4030, Online. Association for Computational Linguistics.

Seeking Alpha. 2025. Earnings call transcripts. `https://seekingalpha.com/earnings/earnings-call-transcripts`. Accessed: 2025-07-21.

Hedwig G. Sekeres, Wilbert Heeringa, Wietse de Vries, Oscar Yde Zwagers, Martijn Wieling, and Goffe Th. Jensma. 2024. Developing infrastructure for low-resource language corpus building. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 72–78, Torino, Italia. ELRA and ICCL.

Pavlo Seroyizhko, Zhanel Zhexenova, Muhammad Zohaib Shafiq, Fabio Merizzi, Andrea Galassi, and Federico Ruggeri. 2022. A sentiment and emotion annotated dataset for bitcoin price forecasting based on Reddit posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 203–210, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anirudha Simha. 2021. Understanding tf-idf for machine learning. `https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/`. Accessed: 2025-07-21.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic. Association for Computational Linguistics.

S&P Global. 2025. S&p capital iq platform. `https://www.capitaliq.com/CIQDotNet/Login-okta.aspx`. Accessed: 2025-07-21.

S&P Global Market Intelligence. 2025. Machine-readable earnings call transcripts dataset. `https://www.marketplace.spglobal.com/en/datasets/machine-readable-transcripts-(45)`. Accessed: 2025-07-21.

StockTwits. 2025. Stocktwits: Social media for investors and traders. `https://stocktwits.com/`. Accessed: 2025-07-21.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

György Szarvas, Richárd Farkas, Róbert Busa-Fekete, Erzsébet Süveges, and Giulio Békés. 2007. Unsupervised relation extraction through pattern clustering. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, page 419–427.

Hideki Takahashi, Yosuke Omori, and Takeo Watanabe. 2009. Designing trading strategies using financial and economic data. *Japanese Economic Review*, 60(2):183–200.

Aviv Tamar, Shie Mannor, and Huan Xu. 2012. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Alex Tarnopolski, Alexander Panchenko, and Preslav Nakov. 2019. Knowledge graph embeddings for financial entity resolution and link prediction. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, page 150–161.

The Motley Fool. 2025. Earnings call transcripts. `https://www.fool.com/earnings-call-transcripts/`. Accessed: 2025-07-21.

The World Bank. 2025. World bank open data. `https://data.worldbank.org/`. Accessed: 2025-07-21.

Adrian Theuma and Ehsan Shareghi. 2024. Equipping language models with tool use capability for tabular data analysis in finance. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–103, St. Julian's, Malta. Association for Computational Linguistics.

Thomson Reuters. 2025. Streetevents transcripts portal. `https://www.streetevents.com/login.aspx`. Accessed: 2025-07-21.

TigerGraph. 2024. Knowledge graph and nlp revolutionize the way aml looks at peps. `https://www.tigergraph.com/uncategorized/knowledge-graph-and-nlp-revolutionize-the-way-aml-looks-at-peps-3/`. Accessed: 2025-07-21.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

UC Irvine. 2023. Reuters-21578 text categorization collection. `https://huggingface.co/datasets/ucirvine/reuters21578`. Accessed: 2025-07-21.

U.S. BEA. 2025. Official bea website. `https://www.bea.gov/`. Accessed: 2025-07-21.

U.S. SEC. 2024. Edgar application programming interfaces (apis). `https://www.sec.gov/search-filings/edgar-application-programming-interfaces`. Accessed: 2025-07-21.

U.S. SEC. 2025. Sec edgar: Overview of the system. Accessed: 2025-03-05.

Vanity Fair Staff. 1998. Wall street lays another egg. *Vanity Fair*. Accessed: 2025-07-21.

Francielle Vargas, Fabrício Benevenuto, and Thiago Pardo. 2021. Toward discourse-aware models for multilingual fake news detection. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 210–218, Online. INCOMA Ltd.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Wasin Waeosri. 2020. How to get mrn news analytics data via websocket api. `https://developers.lseg.com/en/article-catalog/article/how-to-get-mrn-news-analytics-data-via-elektron-websocket-api`. Accessed: 2025-07-21.

Shengkun Wang, Taoran Ji, Jianfeng He, Mariam AL-Mutairi, Dan Wang, Linhan Wang, Min Zhang, and Chang-Tien Lu. 2024a. AMA-LSTM: Pioneering robust and fair financial audio analysis for stock volatility prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 379–386, Mexico City, Mexico. Association for Computational Linguistics.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Weikang Wang, Jiajun Zhang, Qian Li, Chengqing Zong, and Zhifei Li. 2019. Are you for real? detecting identity fraud via dialogue interactions. In *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 1762–1771, Hong Kong, China. Association for Computational Linguistics.

Zezhong Wang, Siyang Hao, Inez Maria Zwetsloot, and Simon Trimborn. 2024b. Social network datasets on reddit financial discussion. *Preprint*, arXiv:2410.05002.

Toni Whited. 2023. Costs and Benefits of Reproducibility in Finance and Economics. *Harvard Data Science Review*, 5(3). Https://hdsr.mitpress.mit.edu/pub/mnhzk8gq.

Xianchao Wu. 2020. Event-driven learning of systematic behaviours in stock markets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2434–2444, Online. Association for Computational Linguistics.

Ziang Wu and 1 others. 2023. How powerful is chatgpt? *arXiv preprint arXiv:2301.07597.*

X (formerly Twitter). 2025. X api documentation. `https://developer.x.com/en/docs/x-api`. Accessed: 2025-07-21.

Yan Xiao, Yaochu Jin, and Kuangrong Hao. 2024. Federated document-level biomedical relation extraction with localized context contrast. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7163–7173, Torino, Italia. ELRA and ICCL.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2017. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Yumo Xu and Shay B. Cohen. 2018a. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1980.

Yumo Xu and Shay B. Cohen. 2018b. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Yahoo Finance. 2025. Yahoo finance news feed. `https://finance.yahoo.com/news/`. Accessed: 2025-07-21.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020a. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 441–451, New York, NY, USA. Association for Computing Machinery.

Yi Yang, Mark UY, and Allen Huang. 2020b. Finbert: A pretrained language model for financial communications.

Yi Yang and 1 others. 2020c. Finbert: A pretrained language model for financial communications. *SSRN Electronic Journal*.

Zheng Yangjia, Li Xia, Ma Junteng, and Chen Yuan. 2022. Fundamental analysis based neural network for stock movement prediction. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 973–984, Nanchang, China. Chinese Information Processing Society of China.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. 2022. Wild-time: A benchmark of in-the-wild distribution shift over time. In *NeurIPS 2022 Datasets and Benchmarks Track*.

Chengxi Zhai, Yang Zhang, Dexin Zhao, and Eric H. Chi. 2019. Charting latency-sensitive natural language processing models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 1099–1110.

Zhilu Zhang, Procheta Sen, Zimu Wang, Ruoyu Sun, Zhengyong Jiang, and Jionglong Su. 2024. FinBPM: A framework for portfolio management-based financial investor behavior perception model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 246–257, St. Julian's, Malta. Association for Computational Linguistics.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Hanqi Jiang, Yi Pan, Junhao Chen, Yifan Zhou, Gengchen Mai, Ninghao Liu, and Tianming Liu. 2024a. Revolutionizing finance with llms: An overview of applications and insights. *Preprint*, arXiv:2401.11641.

Yiyun Zhao, Prateek Singh, Hanoz Bhathena, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram, and Saket Sharma. 2024b. Optimizing LLM based retrieval augmented generation pipelines in the financial domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 279–294, Mexico City, Mexico. Association for Computational Linguistics.

Zishuo Zhao, Ziyang Ma, Zhenzhou Lin, Jingyou Xie, Yinghui Li, and Ying Shen. 2024c. Source-free domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15076–15086, Torino, Italia. ELRA and ICCL.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *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 337–346, Hong Kong, China. Association for Computational Linguistics.

Fan Zhou, Shengming Zhang, and Yi Yang. 2020. Interpretable operational risk classification with semi-supervised variational autoencoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 846–852, Online. Association for Computational Linguistics.

Yifan Zhou, Yiming Liu, Qitian Wu, Shuya Li, Zhiguang Wang, Ruoxuan Xiong, Bin Liu, Panpan Zheng, and Fangjian Guo. 2024. Fnspid: A large-scale financial news-stock price impact dataset. `https://arxiv.org/abs/2402.06698v1`. Accessed: 2025-07-21.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual*

*Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Nadhem Zmandar, Tobias Daudert, Sina Ahmadi, Mahmoud El-Haj, and Paul Rayson. 2022. CoFiF plus: A French financial narrative summarisation corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1622–1639, Marseille, France. European Language Resources Association.

Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022. Astock: A new dataset and automated stock trading based on stock-specific news analyzing model. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 178–186, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

# A Timeline Figure

This appendix includes the timeline Figure 7.

# B Authorship and Funding

Based on the analysis we conducted, we also made a binary classification of the following categories:

- Academic authors
- Industrial authors
- Funding (industrial)
- Funding (governmental)
- Funding (academic)

As seen in the Figure 6, we see that there is no major correlation between funding and authorship. Thus it does not imply that if an industrial/academic author is present in the list of authors, there will necessarily be funding from the industry/academia/government. This finding is surprising actually as one would expect that a paper written by only academics would have academic/governmental funding and the same for industrial authors and industrial funding.

Table 4: Summary of Authors and Funding Categories

| Category | Sum |
|---|---|
| Academic and Industrial | 78 |
| Only Academic authors | 118 |
| Only Industrial authors | 25 |
| Funding (industrial) | 49 |
| Funding (governmental) | 61 |
| Funding (academic) | 35 |

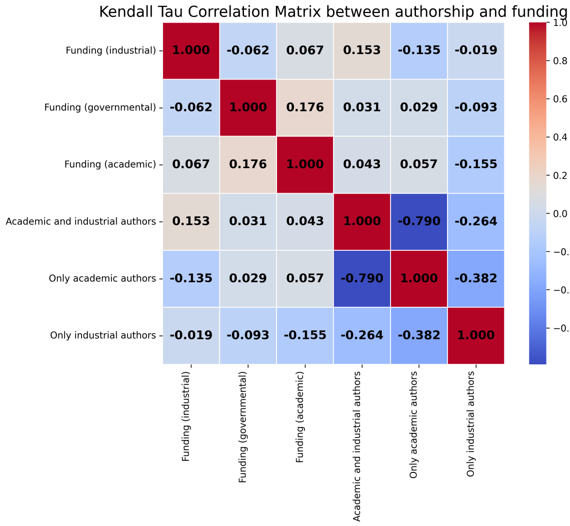


Figure 6: This figure displays the correlation between authors from industrial and academic backgrounds within various types of funding such as academic, industrial and governmental.

# C Survey Dimensions, Methodologies, and Task Categories in Financial NLP Research

This appendix consolidates the key reference tables from our study, providing an overview of the dimensions examined (Table 5), the methodologies employed (Table 6), and the task categories identified in NLP research applied to the financial domain (Tables 7 and 8).

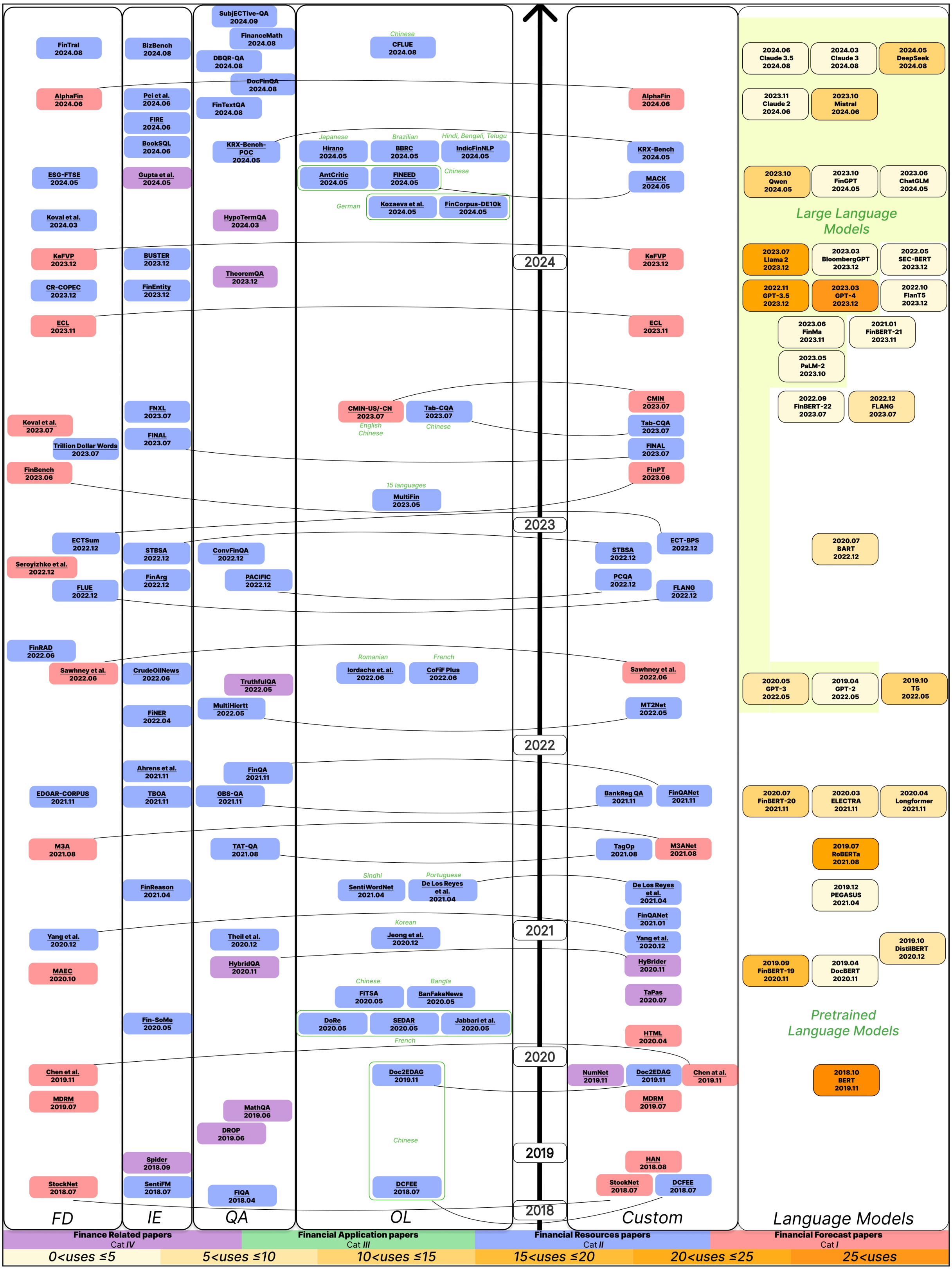

Figure 7: Timeline of PLM/LLM adoption in NLP research applied to finance, alongside key datasets by task type: C (custom models), QA (question answering), IE (information extraction), OL (other language), and FD (financial documents). For each model, the top date indicates release; the bottom, first usage in the surveyed papers.

Table 5: This table outlines the dimensions examined in our analysis of NLP research focused on the financial domain. Dimensions that led to significant findings are discussed in the main paper, while an analysis of authorship and funding is provided in Appendix B.

| **Exploration dimension** | **Description** |
|---|---|
| Primary task | The main task addressed in the paper, such as volatility prediction (Wang et al., 2024a). |
| Sub-Tasks | Additional tasks that support the primary objective, like using sentiment analysis to enhance stock price predictions (Jain and Agrawal, 2024). |
| Methodology | Techniques applied, ranging from traditional machine learning to deep learning and large language models. See Section **??**. |
| Code and data availability | Whether the paper provides open-source code or data, and the quality of this accessibility (e.g., active links). |
| Contribution | What type of contribution does the paper have (Dataset, framework, model, evaluation, etc.)? |
| Comparability of research | What other research did the researchers compare their work to? |
| Data source | The types of data used, including financial reports, social media, news articles, or other financial documents. |
| Metrics | Evaluation criteria, including standard ML metrics (accuracy, precision) and finance-specific metrics (Sharpe ratio). |
| Data years | The time periods covered by the data, such as crisis years or more stable financial periods. |
| Authorship | Indicates if the paper was authored or co-authored by industry professionals, reflecting practical applications. |
| Funding source | The origin of the research funding (academic, industrial, or governmental) and its influence on the research focus. |

Table 6: This table categorizes the key methodologies into nine groups."PLMs" stands for Pretrained Language Models, while "LLMs" stands for Large Language Models.

| **Methodology** | **Description** |
|---|---|
| Statistical NLP | Methods like TF-IDF, Bag-of-Words, and n-grams focused on extracting statistical patterns from financial text. |
| Embeddings | Techniques such as Word2Vec, GloVe, and custom embeddings designed to map financial terms into vector spaces, improving downstream tasks. |
| Conventional ML | Algorithms like Logistic Regression, Support Vector Machines (SVM), and Decision Trees, often used for classification and risk prediction. |
| Deep Learning (DL) | Custom neural network architectures tailored for specific financial tasks such as stock price forecasting. |
| PLMs | Encoder-only models (e.g., BERT) and Encoder-Decoder models (e.g., T5) are used for tasks like document summarization and information extraction. |
| LLMs | Models like GPT-3.5 and LLaMA-2 focus on text generation and understanding complex financial language. |
| Statistical Modeling | Involves correlation analysis, Granger causality, etc., to understand relationships in financial data. |
| Graphs | Uses graph structures to model interactions in financial systems. |
| Knowledge Graphs | Integrating structured knowledge within financial tasks to enhance model performance and reliability. |

| Primary Task | Examples & Insights |
|---|---|
| **Finance-Related (Categories IV)** | |
| Explainable AI (XAI) refers to methods that make machine learning outputs more interpretable and less of a "black box." | These papers show how small input changes can drastically affect model explanations (Sinha et al., 2021) and highlight the instability of tools like LIME in sensitive applications (Burger et al., 2023). In finance, where decisions must be traceable and justifiable, clear and reliable explanation methods are essential for building trust in tasks like fraud detection and risk assessment. |
| Privacy-preserving methods protect sensitive data while still allowing model training and analysis. | Examples include homomorphic encryption for text similarity (Kim et al., 2022a), federated learning for distributed training without raw data exchange (Zhao et al., 2024c), and domain adaptation that shares only model parameters (Xiao et al., 2024). These methods are especially important in finance, where institutions must work with private data while staying compliant and avoiding breaches. |
| Numerical reasoning involves working with numbers and structured data to solve problems or develop algorithms. | This includes solving math word problems using declarative knowledge (Roy and Roth, 2018), learning numeracy through number embeddings (Duan et al., 2021), and pretraining models for verifying tabular claims (Akhtar et al., 2023). In finance, good numerical reasoning supports tasks like market prediction and risk evaluation by improving how models interpret and reason about numbers. |
| **Financial Applications (Category III)** | |
| Sentiment analysis helps assess market mood and forecast financial movements using sources like news, social media, and reports. | Studies analyze opinions in reports and news (Rodriguez Inserte et al., 2023), investor sentiment on social media (Guo et al., 2023), and tone in financial texts (Choe et al., 2023). As markets become more influenced by public opinion, sentiment analysis plays a growing role in trading and analysis. |
| Information extraction (IE) and relation extraction focus on identifying entities and linking them to events or other entities. | Examples include extracting financial events from documents (Zheng et al., 2019), signals in reports (Huang et al., 2023), and patterns in social media (Conforti et al., 2022). These tools help connect companies, individuals, and financial instruments to real-world developments (Liou et al., 2021), supporting real-time analysis and prediction. |
| Question Answering (QA) enables systems to find specific information in financial documents, reports, and databases. | QA models support quick access to relevant facts (Mavi et al., 2023), and recent work focuses on calibrating large language models (LLMs) for finance-specific tasks (Zhao et al., 2024b; Theuma and Shareghi, 2024; Addlesee, 2024). This area emphasizes not only building QA tools but also improving their accuracy for use in high-stakes financial contexts. |

Table 7: Overview of key primary tasks in **Categories IV (Finance-Related)** above the mid line and **Category III (Financial Applications)** below. Each entry summarizes the task's role and relevance in NLP research applied to finance, with representative examples and practical insights.

| Primary Task | Examples & Insights |
|---|---|
| **Financial Resources (Category II)** | |
| Dataset and Resource Construction involves building structured, labeled financial datasets for model development and evaluation. | These papers focus on creating large-scale annotated resources (Chen et al., 2020a), including datasets like Tab-CQA (Liu et al., 2023) and ConvFinQA (Chen et al., 2022b) designed for reasoning over tables and multi-step numerical queries. Other works assemble corpora from financial reports (Shah et al., 2022), company filings (Zmandar et al., 2022), news (Tang et al., 2023), and government documents (Shah et al., 2023), enabling downstream tasks like entity linking, numerical reasoning, or forecasting. |
| Fraud detection leverages domain-specific data and models to identify deceptive financial behavior across platforms. | Erben and Waldis (2024) identifies financial scams on Instagram using a fine-tuned BERT model deployed via browser extension and REST API. Another approach detects identity fraud through interactive dialogue, employing knowledge graphs and reinforcement learning to expose inconsistencies in claimed personal data (Wang et al., 2019). These systems highlight how NLP architectures can protect users from financial manipulation. |
| **Financial Forecast (Category I)** | |
| Stock Price and Volatility Prediction aims to forecast stock movements or market instability using text data. | These models use historical data (Sawhney et al., 2021b), news (Ahbali et al., 2022), and event-driven signals (Sawhney et al., 2020a; Wu, 2020) to predict stock trends. Volatility prediction further explores market sensitivity by analyzing unstructured text like press releases or financial articles (Qin and Yang, 2019), helping traders anticipate fluctuations. |
| Risk Prediction involves identifying the likelihood and impact of adverse financial events, such as defaults or market disruptions. | These models analyze unstructured data like earnings call transcripts (Sang and Bao, 2022), regulatory filings, and legal documents to detect early signals of financial risk (Li et al., 2023a). Applications include credit risk estimation, fraud detection, and systemic risk monitoring (Zhang et al., 2024). |

Table 8: Descriptions of major primary tasks in **Category II (Financial Resources)** above the line and **Category I (Financial Forecast)** below. Each entry summarizes the task's focus and contribution to NLP research applied to finance, with representative examples.

## D Conference Proceedings

Table 9 categorizes key conferences and workshops according to the four primary task categories identified in our survey.

## E Papers per year

Table 10 reports the number of papers per year from 1975 to 2024.

## F Glossary

Table 11 provides definitions to the key financial metrics mentioned in this paper.

## G Data Sources

**Regulatory Filings and Earnings Calls** U.S. public companies file a range of documents such as 10-K, 10-Q, 8-K, credit agreements, proxy statements, S-1, S-3, and others are accessible via the SEC's EDGAR platform, through free API calls and RSS feeds (U.S. SEC, 2024). While similar registries exist (e.g., Canada's SEDAR, Japan's EDINET, UK's Companies House), they are rarely used, reinforcing a U.S.-centric and English-language bias (Section 4.4). Earnings calls, where firms discuss quarterly performance with investors, are another rich data source (Mukherjee et al., 2022; Koval et al., 2023; Pardawala et al., 2025). Seeking Alpha transcribes 4,500 calls quarterly (Seeking Alpha, 2025); other providers include The Motley Fool (The Motley Fool, 2025), FactSet (FactSet Insight, 2025), Thomson Reuters (Thomson Reuters, 2025), and S&P Global (S&P Global Market Intelligence, 2025), though often behind paywalls. Companies also post transcripts/slides on IR websites. Public datasets include transcriptions and audio, such as S&P 500 calls from 2017 (Qin and Yang, 2019) and 1,213 companies' data covering 2015-2018 (Li et al., 2020b).

**Financial News and Social Media** Real-time news from sources like Reuters, Bloomberg, Dow Jones, CNBC, FT, and WSJ is vital for forecasting. The Reuters-21578 dataset (UC Irvine, 2023), with 10,369 articles from 1987, remains a benchmark in text analysis. Larger archives such as FNSPID (Zhou et al., 2024) link 15.7 million articles (1999-2023) to S&P 500 stocks. News is accessible via APIs (e.g., Bloomberg (Bloomberg L.P., 2025a)) and free sources like Yahoo Finance (Yahoo Finance, 2025) and Google News (Google News, 2025). On social media, "FinTwit" and platforms like X and StockTwits are used for sentiment analysis. Key datasets include Sentiment140 (Kaggle Contributor, 2025; Chaudhary, 2020), StockNet (Xu and Cohen, 2018b), and Reddit datasets (Kaggle Contributor, 2021; Wang et al., 2024b). X's API offers limited free and full historical access under paid tiers (X (formerly Twitter), 2025). StockTwits, with Bullish/Bearish sentiment tagging, offers free and commercial API access (StockTwits, 2025) and is often used in forecasting studies.

**Macroeconomic and Monetary Data** Central bank communications (policy statements, minutes, and press releases) offer critical macroeconomic guidance (Ahrens and McMahon, 2021; Peskoff et al., 2023; Menzio et al., 2024b). The U.S. Federal Reserve's FOMC releases detailed statements and transcripts on interest rates, inflation, and economic outlook (Federal Reserve System, 2025), with archives available on the Fed's website. Despite their market relevance, such texts remain underused in NLP-based forecasting. Shah et al. (2023) addresses this by introducing an annotated dataset of FOMC communications and proposing a sentence-level classification task (hawkish, dovish, neutral) using fine-tuned transformers. Economic indicators like GDP, CPI, and unemployment figures – issued by the BEA (U.S. BEA, 2025), FRED (Federal Reserve Bank of St. Louis, 2024), Eurostat (Eurostat, 2025), IMF (International Monetary Fund, 2025), OECD (Organisation for Economic Co-operation and Development, 2025), and the World Bank (The World Bank, 2025) – are mostly numerical but often include press commentary, all freely available via official portals.

**Analyst Reports** These are proprietary equity research reports with buy/sell calls and sector outlooks. Data providers such as Thomson Reuters, FactSet, Bloomberg, and S&P I/B/E/S aggregate these reports, typically through subscription services. All-in-one platforms like the Bloomberg Terminal (Bloomberg L.P., 2025b) and S&P Capital IQ (S&P Global, 2025) provide access to equity research, earnings estimates, and corporate event data as part of their premium offerings. There is no comprehensive free source; however, excerpts or key takeaways occasionally surface in news articles or on platforms like Seeking Alpha.

| Category | Conferences and Workshops |
|---|---|
| **Financial Forecast (Category I)** | ACL, Australasian Language Technology Association Workshop, CCL, EACL, EcoNLP Workshop, EMNLP, EMNLP (Industry), Financial Technology and Natural Language Processing (FinNLP) Workshop, ICCL, LREC-COLING, NAACL, NAACL (Industry), SRW |
| **Financial Resources (Category II)** | ACL, ACL (Industry), CoNLL, EACL, EcoNLP Workshop, EMNLP, EMNLP-IJCNLP, e-Commerce and NLP Workshop, Financial Narrative Processing and MultiLing Financial Summarisation Workshop, Financial Technology and Natural Language Processing (FinNLP) Workshop, GWC, ICON, ICCL, LREC, LREC-COLING, NAACL, NAACL (Industry), NLP4PI Workshop, Pattern-based Approaches to NLP in the Age of Deep Learning Workshop, SIGHUM Workshop |
| **Financial Applications (Category III)** | ACL, ATALA, Bridging Human–Computer Interaction and Natural Language Processing Workshop, CCL, CoLM, Computational Approaches to Subjectivity, Sentiment and Social Media Analysis Workshop, DeeLIO Workshop, EACL, EcoNLP Workshop, EMNLP, EMNLP (Industry), EMNLP-IJCNLP, Financial Narrative Processing and MultiLing Financial Summarisation Workshop, Financial Technology and Natural Language Processing (FinNLP) Workshop, ICON, ICCL, INLG, LREC, LREC-COLING, NAACL, NAACL (Industry), Natural Legal Language Processing Workshop, News Media Content Analysis and Automated Report Generation Workshop, Pattern-based Approaches to NLP in the Age of Deep Learning Workshop, Safety4ConvAI Workshop, Structured Prediction for Natural Language Processing Workshop, TextGraphs |
| **Finance Related (Category IV)** | ACL, ACL-IJCNLP, BlackboxNLP Workshop, EMNLP, Financial Technology and Natural Language Processing (FinNLP) Workshop, LREC-COLING, SRW, TrustNLP Workshop |

Table 9: Conference and Workshop Categorization

| Year | 2024 | 2023 | 2022 | 2021 | 2020 | 2019 | 2018 | 2017 | 2016 | 2015 | 2014 | 2012 | 2010 | 2008 | 2006 | 2002 | 2001 | 1976 | 1975 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Papers | 127 | 53 | 90 | 32 | 57 | 15 | 14 | 21 | 6 | 1 | 5 | 1 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |

Table 10: Number of papers satisfying our filtration criterion per year.

| Metric | Definition |
|---|---|
| Sharpe Ratio | Measures how much return an investment gives for each unit of risk, compared to a risk-free asset (Investopedia, 2025a). |
| Maximum Drawdown | The biggest drop from a peak to a low point in a portfolio's value before it recovers. |
| Cumulative Return | The total profit or loss from an investment over time. |

Table 11: Key Definitions of financial terms used in our paper.

| Document type | Sources | Accessibility (incl. access method) | Modalities | Update cadence |
|---|---|---|---|---|
| Regulatory filings | SEC EDGAR; other registries: SEDAR (Canada), EDINET (Japan), Companies House (UK) | Public via web UI; free EDGAR APIs and RSS (U.S. SEC, 2024); other registries public but less commonly used | Text/PDF | Event-driven; periodic (annual/quarterly) |
| Earnings calls (transcripts, slides, audio) | Seeking Alpha (Seeking Alpha, 2025); The Motley Fool (The Motley Fool, 2025); FactSet (FactSet Insight, 2025); Thomson Reuters (Thomson Reuters, 2025); S&P Global (S&P Global Market Intelligence, 2025); company IR sites | Mixed: many providers behind paywalls; IR sites often public; access via web UI and provider APIs (where offered) | Text transcripts; audio | Quarterly around earnings; event-driven (scheduled call dates) |
| Financial news | Reuters, Bloomberg, Dow Jones, CNBC, FT, WSJ; free aggregators like Yahoo Finance (Yahoo Finance, 2025) and Google News (Google News, 2025) | Mixed: premium (e.g., Bloomberg) and free sources, including static datasets (UC Irvine, 2023; Zhou et al., 2024); access via web UIs and vendor APIs (e.g., Bloomberg API (Bloomberg L.P., 2025a)) | Text (articles; headlines) | Real-time / continuous |
| Social media | X (Twitter); Reddit; sentiment datasets for X (Kaggle Contributor, 2025; Chaudhary, 2020; Xu and Cohen, 2018b) and Reddit (Kaggle Contributor, 2021; Wang et al., 2024b) | X API with limited free and paid tiers (X (formerly Twitter), 2025); StockTwits free/commercial API (StockTwits, 2025); public web UIs | Short-text posts; sentiment tags (StockTwits Bullish/Bearish) | Real-time |
| Macroeconomic & monetary communications | U.S. Federal Reserve (FOMC) statements and transcripts (Federal Reserve System, 2025) | Public via official web portals; archives available | Text (policy statements, minutes, press releases) | Scheduled (policy meetings, release calendars) |
| Economic indicators | BEA (U.S. BEA, 2025), FRED (Federal Reserve Bank of St. Louis, 2024), the World Bank (The World Bank, 2025) | Public via official portals and APIs | Numeric time series + accompanying text commentary | Scheduled (monthly / quarterly) |
| Analyst reports | Bloomberg Terminal (Bloomberg L.P., 2025b); S&P Capital IQ (S&P Global, 2025) | Paid/proprietary subscription platforms; enterprise access/APIs; occasional public excerpts in news/Seeking Alpha | Text/PDF (reports, notes) | Ongoing; event-driven (coverage initiations, earnings, sector updates) |

Table 12: Financial text/data sources by document type with sources, accessibility (and access method), modalities, and update cadence.