ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

A Practical Analysis of Large Language Models Empowering Investment Management from the Financial Text Perspective

Yuying Zhao

The University of Lancaster, Lancaster, LA1 4YW, UK

Abstract

This paper investigates how large language models can assist in investment management through the processing of financial text. It is a discussion of ideas and applications, not an empirical investigation, and no original datasets, backtests, portfolios or performance data have been provided. The development path from dictionary-based textual indicators to Transformers, domain-adapted language models and retrieval-assisted workflows is presented. Next, it presents some applications of research-intake systems, document comparison, event extraction, risk monitoring, investment-memo drafting and client communication. Pay attention to the evidence supporting the results of the model. Although the generated summary and classification can increase the efficiency of the workflow, they are not verified facts, forecasts or investment recommendations unless supported by official data and professional reviews. A controlled operating model based on source provenance, retrieval, structured prompts, numerical verification, human approval, access control, audit log and continuous evaluation is proposed in this paper. Therefore, large language models are more suitable for the applications of supervised analysis interfaces than for autonomous portfolio management systems.

Keywords

Large language models, financial text, investment management, textual analysis, retrieval-augmented generation, model risk.

1. Introduction

Investment Management is based on both numbers and words. All the above are sources of information for events that are not fully disclosed in price data or accounting ratios, such as annual reports, regulatory filings, earnings call transcripts, policy statements, research notes, news and stewardship documents. There is a problem of the volume, speed and complexity of the language; relevant statements need to be identified, interpreted, linked with previous disclosures, and included in the documented decision process.

A New Interface for this work is Provided by LLMs. They can summarise, categorize, compare, extract, translate and generate text according to the instructions. These functions can reduce the cost of search and help analysts get to the point more quickly. The Fluent output may contain unsupported claims, old information, ambiguous references or numerical errors. A good answer in terms of language may not be suitable for investment purposes.

This paper studies the practical paths for large language models to support investment management from a financial-textual perspective. It is only the theory part and does not contain any original databases, samples, experiments, models, trading strategies or investment results. Published studies and public frameworks serve as external support; however, the proposed workflow is only a design idea and has not been verified by tests. In short, the values of language models are generated through regulated, source-grounded processes, and they remain subject to control by data verification, fiduciary judgment and accountable human approval.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

2. From Textual Measures to Language Models

Financial textual analysis developed before modern generative systems. Tetlock has quantified the impact of pessimistic media language on stock market declines in [1]. Based on the above study, although financial language can be used as a research object, it has not been found that tone alone is a stable or readily tradable signal. Its more extensive practical implication is that the interpretation of attention and sentiment in text is affected by time, context, and research design.

Loughran and McDonald have noted that general-purpose dictionaries may not be suitable for financial terms and have thus developed domain-specific word lists of regulatory filings [2]. Liability is a negative word in daily life, but it is also used often in accounting. Based on the above study, generic sentiment labels are not suitable for evaluating the financial repercussions of investment workflows. Specify the type of document, industry, accounting standards and purpose of use for this division in the definition.

Transformer replaced recurrence with attention-based processing to address the problem of modelling long-distance dependencies in sequential data in parallel [3]. This technology will be used to enhance the functionality of classification, extraction and question-answering in financial text through context propagation across multiple modules; however, it will not automate financial literacy. If the filing period is too long, it is impractical to carry out fieldwork and disperse samples for retrieval.

BERT showed that bidirectional pre-training on unlabeled data can be fine-tuned for various language-understanding tasks [4]. Araci's FinBERT conducted domain adaptation for financial sentiment analysis and improved performance on the evaluation set [5]. The development above provides a multi-level view of ability: general linguistic representations offer the foundation, and financial corpora, task definitions and labelled examples enhance domain adaptation. Good results in a benchmark do not guarantee reliability in other markets, languages, disclosure systems, or the future.

Large autoregressive models extended this idea to perform tasks based on natural-language instructions and a small number of examples. Brown and his colleagues have achieved good few-shot results in many language tasks; however, there are also some shortcomings [6]. Instruction-based interaction in investment management reduces the technical barrier to exploration but also enables informal operations. Suggestions can lead to different outcomes, and a reasonable-looking answer may not be based on facts. Therefore, ease of use should increase the demands for validation, not reduce them.

3. Financial Text as an Investment Input

First, in practice, the sources need to be identified. Regulatory filings are the company's formal disclosures, but they can be long, legally complex, and historical. Earnings-call transcripts contain prepared statements and Q&A; although convenient for research, there are also problems of transcription and attribution. News is timely, but it may lack verification and frequently report the same incident. Analyst research adds interpretation but may be licensed, opinionated or restricted from model training and redistribution.

The Units of meaning in financial texts differ. The Investment team may need the entities, dates, accounting periods, currencies, guidance ranges, causal statements, management commitments or amendments to the above information. Ask the model to "summarize the filing", and the result is the same. A safer workflow has been established to record the claim, speaker, document, publication time, reporting period, location of supporting quotation, uncertainty, and whether the item is a fact, forecast or opinion/model inference.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Opportunities and requirements for data design in domain-specific large models. BloombergGPT was trained on a combination of financial and general-purpose corpora, and tested for financial and other language tasks [7]. FinGPT introduced an open-source, data-centric approach with a financial data pipeline and light adaptation, and has shown promise for robo-advisory and algorithmic trading [8]. Although they are different development models, neither a name nor an address will be suitable for any specific school. Consider the scope of data rights, coverage, timeliness, evaluation mechanisms and operating conditions for governments.

Provenance refers to the origin and time of modification for a particular sentence, and thus its meaning may have changed. The company's disclosure, the journalist's analysis and the anonymous social media post are not the same. The system should store document identifiers and time stamps, deduplicate duplicate articles, separate primary and secondary sources, and avoid adding materials published after the investment decision to the historical reconstruction.

Time is also a kind of order. A backtest that allows for modified filings, corrected transcripts or later news of an earlier information set introduces look-ahead bias. Although this paper does not perform backtesting, a defensible implementation would record the version and availability time of all sources. The model's output should contain this metadata, and then analysts can determine what information was actually available at the time of the relevant decision.

4. Applications in the Investment Process

The first application is research entry. A retrieval system can distribute new documents according to the issuer, industry, subject, event type, etc., and then a language model generates a source-linked summary. Do not read the entire document; the digest will inform you of any changes in risk factors, altered policies, abnormal accounting, litigation, capital allocation or modifications to the management. Open the relevant support text and decide whether to reject or modify the interpretation of the model.

The second is an in-depth research. Models can compare the current filing or call with previous periods to highlight added, deleted or modified statements. It is convenient; often, the investments are for changes rather than for large amounts of money. Distinguish between alterations of text and enhancements to meaning. Although the deletion of an object may be irregular, we need to know whether it is due to a change in policy or a modification to the draft.

The third kind is structured event extraction. A set prompt will automatically populate the fields for guidance, acquisition, financing, product launch and regulatory action, etc. The above fields may be the analyst's dashboard, and the numerical indicators need to be verified with the tables and the original documents. Validate the unit, sign, fiscal period and currency independently. Do not use a language model to perform arithmetic by having it describe the steps of calculation.

The fourth application is investment-memo support. The model can organise the verified evidence into the categories of thesis, catalysts, valuation assumptions, counterarguments and risks. It may also be in conflict with the main idea or irrelevant to the thesis. Authorship belongs to the investment professional. The final memo should present the collected data and analysis results of the model professionally, and no other party is allowed to add their own views.

The fifth application is Monitoring. A language model categorizes the incoming text based on the risk taxonomy of the portfolio, such as refinancing, governance, cyber incidents, customer concentration, policy exposure, etc. Calibrate the alert function to help prioritise, do not trigger a trade automatically. False positives lead to fatigue; false negatives result in false comfort. Based on the primary documents, use market, accounting, legal or technical expertise to confirm material decisions.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

The sixth is a client and relationship manager. The model can generate simple-language explanations of the approved portfolio records and prepare issue summaries for contacting issuers. Evaluate the suitability, legality, confidentiality and conformity with the plan, etc. Personalised Language should not be mistaken for the model knowing about a client's life. It is a design mode under a supervised consultation process.

5. Evidence, Evaluation and Misleading Statements

Assess according to the actual situation. Evaluation of sentiment classification can be performed using labelled data, extraction by field-level precision and recall, and question-answering based on answer correctness and citation support. Quality assurance of summaries is to ensure they are not incomplete and consistent. High-level aggregation of scores may not show significant errors in infrequent but material circumstances; therefore, the test cases need to include stress cases of negation, tables, amended filings, conflicting sources, and ambiguous time references.

FinanceBench is an open-book financial question-and-answer benchmark, and its questions, answers, and evidence strings have been obtained from public company documents [9]. Based on the evaluation, some of the models had serious defects, such as hallucinations and wrong or refused answers. The benchmark meets the demand for evidence-based testing, but its performance in the sample cannot be generalised to all errors of the enterprise. Add tests for the tools, documents, languages and controls of the institution.

Lopez-Lira and Tang have studied whether the model classification of news headlines is related to changes in stock prices [10]. The above results support the hypothesis that advanced models can process financial news, but this does not guarantee the generation of profits through deployment. The predicted results of the different samples, models, prompts, data periods, transaction costs and market adaptations will also differ. This paper does not present the results of that study or an investment plan.

Often, the false claim is the combination of workflow advantages and portfolio performance. A short study of the documents cannot guarantee a high return; an improvement in benchmark accuracy is not fitness for life administration; and a reasonable explanation does not equate to causal awareness. Do not use language that suggests an LLM "predicts the market", "removes analyst bias", or "automatically guarantees compliance"; instead, provide specific statements about the task, evidence, evaluation period and human supervision.

Model comparison should also consider the effects of time. Hosted systems are not always reproducible in the same manner, and open models can vary due to changes in weights, fine-tuning data, quantisation, etc. All evaluated systems and the corresponding prompts, retrieval settings, temperatures, dates and test sets must be versioned. If an exact reproduction is not possible, this deficiency should be stated in the performance information rather than omitted.

6. Risk Controls and Governance

The four components of the NIST Artificial Intelligence Risk Management Framework are: Govern, Map, Measure, and Manage [11]. Application to Investment Management: Governance determines ownership and approval rights; Mapping identifies users and affected parties, data flows, failure consequences, etc.; Measurement tests reliability, privacy, security, and bias; and Management selects controls and monitors residual risk. The system is voluntary and non-binding; therefore, the company will not be held legally or under fiduciary duty for any damage caused.

A controlled architecture can integrate retrieval-augmented generation with permissioned document stores. Retrieval reduces the evidence provided to the model and enables citation;

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

however, it does not eliminate the error. Search may not have the main content, rank the old edition higher, or provide sufficient information. Operate the access control before collection, and display the source, date and passage for all answers. Discard the unsupported responses or mark them as hypotheses.

The human review should be reasonable and reasonable. The reviewer should have some knowledge of the field and relevant data to validate the results. High-impact applications, such as trade recommendations, client advice and valuations, or regulatory communications, require a higher-level approval than low-impact applications, such as formatting internal notes. Segregation of duties can prevent one person from both setting up the system and approving the corresponding transactions without any independent checks.

The U.S. The Securities and Exchange Commission's 2023 proposal on predictive data analytics addressed conflicts that may arise from broker-dealers' or investment advisors' use of covered technologies in investor communications [12]. As it was only a proposal, it should not be published as a new rule. The system can be designed to promote the interests of the company rather than those of its shareholders. The company should check the current laws and relevant jurisdictions itself, not use this paper for legal advice.

Other controls are vendor due diligence, contractually limited data retention, encryption, incident response, timely injection testing, model change review and full audit logs. Do not save confidential data, client information, proprietary research or material non-public information in an unauthorised external system. Periodically conduct tests of both the normal quality and severe failure modes, and set clear thresholds for suspension, rollback and escalation.

Thus, a feasible operating model is to add a language model to the extended decision system. The foundation includes the main evidence, verified calculations, portfolio constraints, risk system, responsible staff, etc. Automation can prepare, route and challenge information, but it does not have the power to act. This Design can use the advantages of natural language processing and will not have errors due to fluent but unsupported text.

7. Implementation Path

First, build a narrow, reversible use case for generating source-linked filing digests in an internal study. Set up users, approved documents, prohibited data, expected output, error category and escalation rules for the team before deployment. A foundation is provided by the existing manual operation process. The indicators of success are citation accuracy, omission rate of materials, correction rate by reviewers, processing time and user satisfaction; however, the actual results will vary from institution to institution.

Then, a selected knowledge base and organised prompts will be supplied. Templates need to have the model cite the evidence, specify when the information is unavailable, keep units and periods, and separate extraction from interpretation. Deterministic Software will perform arithmetic and portfolio calculation. The reviewer can view the answer and the retrieved context, and the system will save the feedback in the managed test repository instead of automatically re-tuning for unverified corrections.

Expansion will only occur after error analysis. The addition of a new document type, language, asset class or client-facing application will change the risk and need to be tested. Model upgrades should be considered material system changes if they alter behaviour. Parallel execution, approval checkpoints and a rollback function will reduce the service disruption. Independent validation needs to test not only the mean accuracy but also data leakage, temporal inconsistency, conflict of interest, and the impact of plausible fabrication.

Governance needs to determine whether the tool has enhanced the investment process and, in the absence of other evidence, not attribute performance improvements to it. The advantages are good source traceability, wider coverage, early detection of inconsistent disclosures, and

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

improved documentation of uncertainty. They are practical and knowable benefits. They offer a reference but are not absolute, and the fault for this investment institution cannot be attributed to the model.

8. Conclusion

Large Language Models can help to organise the financial data in an investment management firm for easy query and comparison. The first few applications of these are document classification, evidence extraction, longitudinal analysis, risk supervision and memorandum writing. All models provide support for and interpret the world of organisations in different ways.

The actual application of these systems is more dependent on the surrounding information architecture than on fluent generation. Reliable Use requires licensed and time-stamped sources, retrieval with provenance, explicit schemas, numerical verification, task-specific evaluation, access control, audit records and qualified human approval. A model that cannot present the evidence for a claim should not be used to make consequential investment decisions.

No data have been generated; no portfolios tested; and no higher returns achieved. Based on the published research and public governance documents, a cautious implementation path has been proposed. Therefore, the appropriate conclusion is limited: Large Language Models can improve the processing of financial text and the discipline of investment research when used as supervised tools, but they do not eliminate model risk, market uncertainty, conflicts of interest, or professional responsibility.

References

[1] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[2] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[5] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063. https://doi.org/10.48550/arXiv.1908.10063

[6] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[7] Wu, S., Irsoy, O., Lu, S., et al. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564. https://doi.org/10.48550/arXiv.2303.17564

[8] Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. arXiv:2306.06031. https://doi.org/10.48550/arXiv.2306.06031

[9] Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A new benchmark for financial question answering. arXiv:2311.11944. https://doi.org/10.48550/arXiv.2311.11944

[10] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv:2304.07619. https://doi.org/10.48550/arXiv.2304.07619

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[11] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

[12] U.S. Securities and Exchange Commission. (2023). Conflicts of interest associated with the use of predictive data analytics by broker-dealers and investment advisers (Release Nos. 34-97990; IA-6353). https://www.sec.gov/files/rules/proposed/2023/34-97990.pdf