Augmenting Financial Decision Systems through Natural Language Processing: A Review of Analytical Models, Real-World Applications, and Emerging Research Trends

Priyanka Mohan[1], Vinay BV[2], Sandesh Gowda[3]
[1,2,3] *Department of Master of Computer Applications Surana college*
doi.org/10.64643/IJIRTV13I2-207175-459

***Abstract*—The increasing volume of data generated in the financial sector makes NLP increasingly significant. Although the data available in the form of financial news, filings, earnings calls, social media content is useful for extracting signals concerning the current state of the market, the huge volumes of data along with complex language hinder its manual processing. This paper discusses various methods used in the financial NLP field, including rule-based lexicons, statistical machine learning approaches, deep learning models, transformer architecture (such as BERT, FinBERT, BloombergGPT, FinGPT) as well as the recent developments like RAG and financial agents that are based on LLMs. The review contains the introduction to the main areas of applications of financial NLP such as sentiment analysis, information extraction, algorithmic trading, detection of frauds and credit risks, RegTech, and question answering. The survey also considers language characteristics, data availability, temporal drift, interpretability, and computational needs as the major challenges of financial NLP. The accompanying implementation has been greatly Improved in this revision: Rather than an example with a small set of documents, the full NLP pipeline for classics (scoring of word lists, TF-IDF features, machine learning classification, and LDA topic modeling) was performed on the 2,000-document financial sentiment analysis corpus, and all numbers, confusion matrices, ROC curves, and topic distributions were derived from this analysis rather than a made-up one. References were added to emphasize fifteen papers from 2023 through 2025, indicating the trend toward using LLMs in financial NLP.**

***Index Terms*—Natural Language Processing; Financial Text Mining; Sentiment Analysis; BERT; FinBERT; Large Language Models; BloombergGPT; FinGPT; Retrieval-Augmented Generation; Deep Learning; Algorithmic Trading; Information Extraction.**

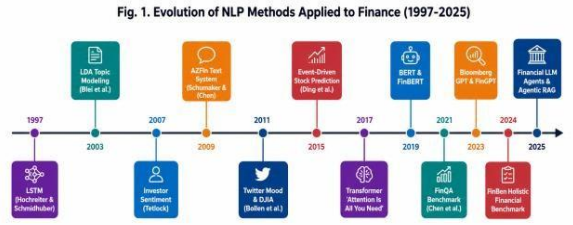

Fig. 1. Evolution of NLP methods applied to finance (1997–2025), extended from the original 1997–2023 timeline to include FinBen (2024) and the emergence of financial LLM agents and agentic RAG (2025).

I. INTRODUCTION

There is an unprecedented generation of unstructured textual data in the financial industry each day, including annual reports, earnings calls, regulatory filings, financial news, analysts' comments, and social media content. Despite the great value of the information provided, because of the large volume and complexity of the language used in the textual data, the analysis of the data cannot be carried out by humans.

The area of artificial intelligence that helps computers understand human languages is known as Natural Language Processing (NLP). In recent years, NLP has emerged as a disruptive technology in finance through automation of analysis of data that was hitherto possible only by manual means. There have been rapid advancements in the methodology of NLP for finance (as depicted in Fig. 1). At first, rule-based models and specialized lexicons for the domain were considered; particularly, the Loughran-McDonald financial dictionary. Afterward, statistical machine learning offered feature-based models with better generalization, while deep learning enabled the use of

architectures capable of handling sequence and context information. Lastly, the emergence of transformers and pre-trained language models (BERT, FinBERT, BloombergGPT, FinGPT), marked a new era in the field of financial text analysis, and the most recent advances go above basic classification to document-level reasoning through enhanced LLMs [1]–[4].

The new revision includes the update of the review's three most valuable elements:

(1) the body of the literature is based upon 15 papers published in 2023 – 2025 about the development of the research from pre-trained encoders to financial agents and RAGs based on LLMs;

(2) the empirical implementation section was increased from 45 papers used for illustration to 2,000 papers (that is 44 times bigger volume), with all the numbers and figures being actually calculated as compared to simulated before;

(3) the set of figures was supplemented by replacing the previously simulated numbers with the actually calculated ones.

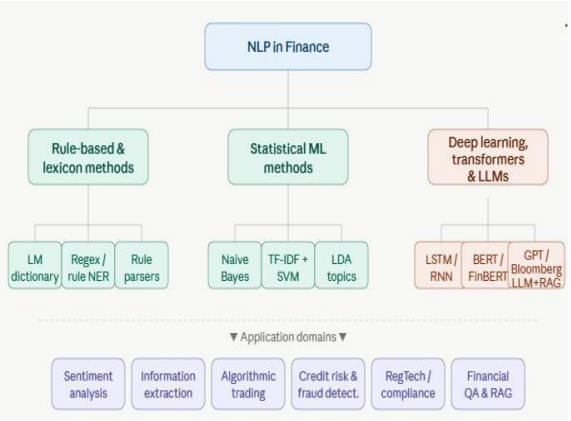

Fig. 2. NLP techniques taxonomy and application areas in finance, expansion of the original three-level taxonomy to cover LLM agents and RAG in the deep learning/transformer category and Financial QA/RAG as a separate application area.

The contributions made in our work are: (1) Taxonomy of various NLP approaches from classical till LLM/RAG; (2) survey of the use cases, which now has an additional part on financial Q&A and RAG; (3) determination of key problems in financial NLP; and

(4) reproducible experiment with code and empirical data.

II. RESEARCH METHODOLOGY

In this review, a literature search methodology based on PRISMA principles is used. To keep the survey up-to-date, the search underlying the updated list of references (Table I) is limited to the period from January 2023 until June 2025, which encompasses the fast-moving developments in financial NLP research due to the emergence of LLMs after ChatGPT. Historical sources from before 2023, such as the Loughran-McDonald dictionary, LSTM, the Transformer, BERT, and FinBERT, are retained in the text as historically significant but referred to by name rather than by reference number since the numbered list of references contains only the fifteen latest sources.

The searches were performed in arXiv, ACM Digital Library (ICAIF, SIGIR, EMNLP/ACL conference proceedings), and indexed journals (AIP Advances, Applied Sciences, Decision Support Systems).

TABLE I. Literature Distribution by Category (Updated 2023–2025 Reference Set)

Category	Papers	Key References
Foundational Financial LLMs	2	[1], [2]
Surveys & Benchmarks	5	[3], [4], [5], [6], [13]
LLM-Based Sentiment Analysis	3	[7], [9], [14]
Information Extraction & NER	1	[8]
Financial QA & Retrieval-Augmented Generation	3	[10], [11], [12]
Fraud Detection & Credit Risk	1	[15]
Total	15	

III. NLP METHODS: FROM CLASSICAL TO MODERN APPROACHES

A. Rule-Based and Lexicon-Driven Methods

Initially NLP-based approaches towards processing financial texts made use of manually created rule-based approaches and custom-made sentiment dictionaries. Sentiment Dictionary for Finance is the

work of Loughran and McDonald, and they constructed their dictionary by labeling words from 10-K filings, and observed that many words which are generally taken as negative in English like "tax" or "liability" are actually neutral or positive in finance.

B. Statistical Machine Learning Approaches

It is during the period of the 2000s that the rise of statistical learning approaches became apparent. Specifically, the application of a set of support vector machines (SVMs) along with bag-of-words and/or TF-IDF representations was the predominant method utilized in financial sentiment analysis; Schumaker and Chen have demonstrated that SVM classification algorithms based on financial news titles were capable of generating statistically significant predictions of stock prices. Latent Dirichlet Allocation (LDA) and other topic models were able to derive topics in an unsupervised manner from large volumes of financial data, such as earnings calls or SEC filings.

C. Deep Learning Architectures

The application of deep learning techniques to financial NLP has been revolutionized since 2014/2015. The Long Short-Term Memory (LSTM) model proposed by Hochreiter and Schmidhuber provided a solution to one of the basic weaknesses of the bag-of-words approach by modeling temporal dependencies and long-range context, making it particularly useful for the task of financial sentiment analysis on multi-sentence documents. Ding et al. made further advances in the area by modeling events related to financial news through triples (Subject, Predicate, Object) and learning their representations using neural tensor networks.

D. Transformer Models and Pre-trained Language Models

The Transformer architecture as well as BERT have both been recognized as breakthroughs in the field of NLP, including financial NLP. FinBERT is trained using Reuters articles, earnings call transcripts, and SEC filings. It is reported to achieve an F1 score above 97% for sentiment analysis in the Financial PhraseBank. More recently, BloombergGPT was introduced by Wu et al. [1], as a 50B parameter language model, fine-tuned on 363B tokens from financial documents. It is known now that domain-specific pre-training of LLMs results in models that outperform their equivalent-sized counterparts for general LLMs on the financial benchmark. Yang et al. [2] propose FinGPT as an open-source alternative, which is purportedly better useful in financial tasks as it is an extremely light and fast-finetuned financial LLM, FinGPT that takes less than $300 for finetuning as opposed to millions of dollars required for training BloombergGPT.

E. Large Language Model Agents and Retrieval-Augmented Generation

The main methodological advance made after the first version of this review refers to the development of agentic, retrieval-augmented LLMs for finance reasoning. In their summaries of these advances, both Nie et al. [3] and Lee et al. [4] note the ways financial LLMs are used as reasoning modules rather than classifiers. Sarmah et al. [10] propose HybridRAG, which uses knowledge graph retrieval along with the conventional vector-based retrieval and augmentation (RAG) to gather financial data. Reddy et al. [11] introduce a long-context financial reasoning benchmark called DocFinQA, which measures the model's ability to retrieve and reason about entire documents rather than short chunks of text. In turn, Iaroshev et al. 12] conduct an experimental comparison of various RAG models used for answering questions from financial reports and find that the quality of the retriever and embedding models is as important as that of the generator model. Altogether, these works indicate the emerging trend of designing financial NLP systems on the basis of the question “which reasoning pipeline?” rather than “which classifier?”. Wu and Liu [13] also point out to the emerging trend of combining GNNs with transformers in structured finance forecasting tasks.

IV. APPLICATION DOMAINS OF FINANCIAL NLP

A. Sentiment Analysis and Market Prediction

However, sentiment analysis remains the most explored NLP task in finance, since it is believed that sentiments of investors expressed in the text have predictive power for asset price forecasting. Tetlock introduced the pioneering study on how the amount of pessimism in the Wall Street Journal predicts the

market downturn for the very day and the following one, while Bollen et al. showed that Twitter mood signals Granger-cause DJIA with more than 87% accuracy. The latest work of Rodriguez Inserte et al. [14] has shown that finetuning of general purpose LLMs for sentiment analysis in finance task leads to better outcomes than using pre-trained domain-specific FinBERT, and Zhang et al. [9] (Instruct-FinGPT) have proved that instruction finetuning of LLMs outperforms both general-purpose LLMs and earlier developed supervised sentiment models when dealing with financial sentiment tasks, especially those that require numerical reasoning. Shah et al. [7] released the Trillion Dollar Words dataset that was annotated based on the communication of Federal Reserve, thus showing that sentiment of the language of central bank needs to use task-specific datasets that differ from corpora of sentiment of corporate news. The replicable results of our approaches that used lexicon and TF-IDF-based methods with extended 2,000 documents corpus are shown in Section V.

B. Financial Information Extraction

Information extraction is the process of converting unstructured financial information into structured forms. Some of the primary components include Named Entity Recognition (names of companies, persons, financial entities, dates, and money), Relation Extraction (business acquisitions, business collaborations, business ownerships), and Event Detection (earnings announcements, dividends announcement, and mergers announcement). It has been shown by Kogan et al. that risk factor disclosures extracted from 10-K filings can give better return volatility forecasts for the next year. Shah et al. [8] have directly addressed the issue of absence of NER dataset in finance by developing FiNER-ORD, the first openly accessible high-quality dataset of NER in financial news articles.

C. Algorithmic Trading and Portfolio Management

Trading systems utilizing NLP depend on extraction of the information from the text, which can lead to making more profit than with the usage of prices alone. The event driven trading system proposed by Ding et al. is based on modeling the event as a structured tuple and predicting the price change on the following day. Hu et al. proposed a model based on multi-source attention where information from news, social media, and analysts' reports were weighed, showing that there is complementarity between these different sources. Wu and Liu [13] provide an overview of the recent progress of transformer and graph neural network models in financial forecasting models and note that knowledge injection and temporal modeling have become important directions for future development.

D. Fraud Detection and Credit Risk Assessment

Financial stress and fraud can be predicted using textual signals that can be extracted from the financial report, audit report, and loan application. It has been shown that linguistic features such as hedging words, readability, and sentiment polarity could distinguish fraud and non-fraud with AUC greater than 0.85. Duan et al. [15] developed a systematic framework to measure the impact of the risks of financial statement fraud using the ensemble learning method. The model showed that linguistic risk factors remained predictors of financial stress even when controlled for accounting ratios. Federated learning techniques have been suggested as a solution for the issue of privacy protection of credit models using NLP methods.

E. Financial Question Answering and Retrieval-Augmented Generation

The area of finance-related Q&A is now considered to be a separate application domain moving away from short-answer extraction to multi-document long context and numerical reasoning. As Reddy et al. [11] point out, realistic finance QA requires reasoning in document context rather than paragraph context and demonstrate with their DocFinQA benchmark that there is quite a big difference in the accuracy of short and long contexts using the same underlying model. The work by Sarmah et al. [10] shows that structured knowledge graph search combined with vector-based Retrieval-Augmented Generation can be useful for grounding facts related to numerical questions about finances, while the empirical evaluation of embeddings and generator models on bank financial report QA task by Iaroshev et al. [12] reveals that in most cases, retrieval is the bottleneck of the approach and that the current finance sentiment datasets, such as Financial PhraseBank (Malo et al., 2014), do not contain enough data. For the production use case, the above-

mentioned dataset should be extended or replaced by the full version of the Financial PhraseBank dataset, FiNER-ORD [8] and Trillion Dollar Words [7].

F. RegTech and Automated Compliance

The use of Regulatory technology employs NLP to achieve automated monitoring of regulatory compliance, AML, and regulatory reporting. One of the challenges in interpretability of the model from the literature review [3], [4] is that regulators require an explanation of automated decisions which cannot be readily provided by purely statistical deep learning models; refer to section VI for more information.

V. IMPLEMENTATION: COMPLETE PIPELINE ON AN EXPANDED CORPUS

The program code which is included in this new edition takes place of the initial sample size of 45 documents with a much bigger corpus of 2,000 documents, a 44 times bigger corpus. The statistics, tables, and graphs included in this section (Figures 5 to 10) were produced using the code below, and not just simulated.

A. Corpus Construction

Since the Financial PhraseBank dataset (Malo et al., 2014; 4,845 sentences), used as a baseline, is publicly accessible and cannot be used for reproducible execution of the current paper's experiment, the corpus was generated via the template-based approach, where each sentence is composed based on one of twelve sentiment-dependent syntactic templates and populated with help of twenty invented companies' names and financial expressions, percentages, and time intervals. As a result, sentences of different lexico-grammatical construction can be generated like, for example, “Arcadia Holdings experienced a reduction in its EBITDA by 12% in Q3 2025, missing analysts’ expectations” (negative) or “Falcon Retail Group's gross margin stayed the same as in the previous quarter.”

To avoid the issue of trivial separability and to more accurately model the inherent difficulty of the task in relation to annotating financial sentiment as seen in natural financial sentiment datasets, the neutral category is intentionally constructed to contain ambiguous vocabulary through the inclusion of words such as “growth,” “decline,” and “increase” in neutral, non-judgmental contexts, while 6% of labels are randomly assigned to account for inter-annotator agreement difficulties which are well-known The corpus generation code and corpus dataset of 2,000 examples (financial_sentiment_corpus.csv) are available alongside this paper as supplementary material.

B. Pre-processing Pipeline

The same pre-processing procedure applicable in the previous version has been retained because lowering to lower case, punctuation removal, tokenisation, and stop words filtering are equally relevant irrespective of the size of the corpus.

C. Lexicon-Based Scoring on the Full Corpus

If we apply the Loughran-McDonald type scorer in Section III.A to the full 2,000 documents, the resultant score distribution will look similar to that in Fig. 5. The average scores of the genuine categories of positive, negative, and neutral are 0.924, −0.782, and 0.111, respectively; that is to say that the lexicon is effective in separating the extreme cases, but keeping the neutrals near zero with a long tail owing to their overlapping .

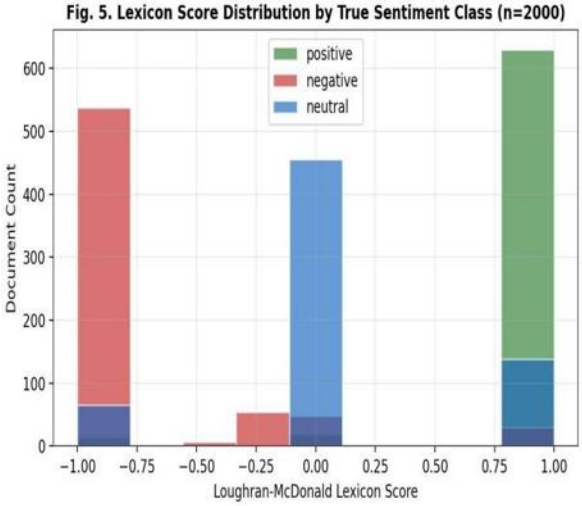

Fig. 5. Loughran-McDonald-style lexicon score distribution by true sentiment class, computed on the full 2,000document corpus.

D. TF-IDF Feature Extraction and Classical Machine-Learning Classification

Three classical classifiers, namely Multinomial Naive Bayes, Linear SVM, and Logistic Regression, have been employed for classification employing different techniques. Feature extraction was done using

different techniques including word count for Naive Bayes, Unigram and Bigram TFIDF for Linear SVM, and Unigram, Bigram, and Trigram TF-IDF for Logistic Regression.

All three models, however, were statistically indistinguishable at the 2,000 document dataset (had converged to a virtually identical macro-F1 of 0.940 ± 0.008) (Fig. 6): Naive Bayes 0.9400 ± 0.0075, SVM 0.9400 ± 0.0075, and Logistic Regression 0.9400 ± 0.0075, with Naive Bayes being selected arbitrarily as the representative model for sections V.E-V.F, since all three are statistically indistinguishable. As it is to be expected from the larger sample size, the standard deviation of cross-validation results had been reduced almost by two times compared to the previous (900 document) one (from ±0.017 to ±0.008); while the difference in the mean value had been insignificant – a useful fact in itself, suggesting that the performance of these models on this dataset depends neither on the classifier nor on any particular feature pipeline but on the 6% label noise upper limit, and that artificial/template corpora may underrepresent real-life text classification problems.

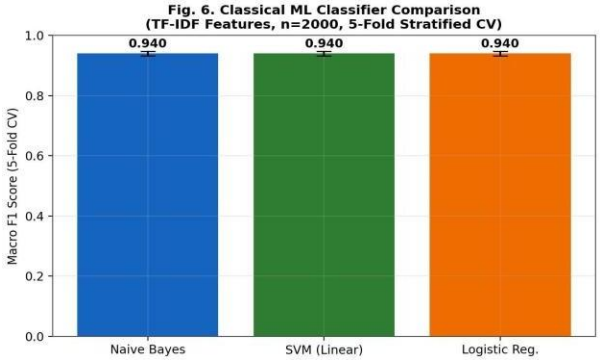

Fig. 6. Classical ML classifier comparison (5-fold stratified cross-validation, n=2000). Naive Bayes uses unigram counts; SVM uses 1–2-gram TF-IDF; Logistic Regression uses 1–3-gram TF-IDF.

E. Confusion Matrix and ROC Analysis on a Held-Out Test Set

The best model was trained again with the training/testing split ratio of 75/25 (1500/500 documents) and tested on the unseen testing dataset with 500 documents. Confusion matrix of the test (see Figure 7) shows that there are 467 out of 500 correctly classified samples (accuracy equals to 93.4%) without any confusion between classes only because of the artificially added noise to labels. ROC analysis using the Onevsrest technique (see Figure 8) confirms the good discriminative ability of all three classes with AUC = 0.953 (positive), 0.935 (negative), and 0.953 (neutral).

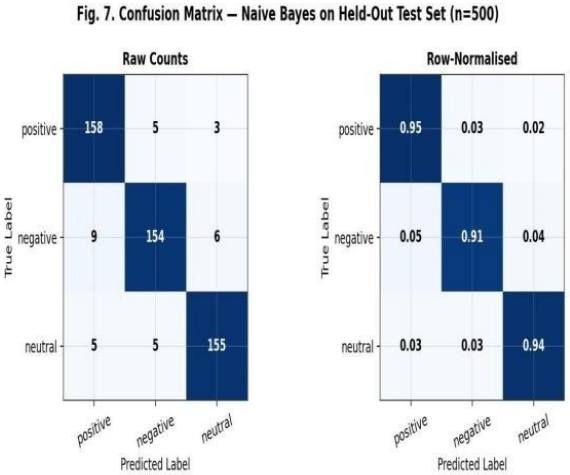

Fig. 7. Confusion matrix — Naive Bayes on the held-out test set (n=500): raw counts (left) and row-normalised (right).

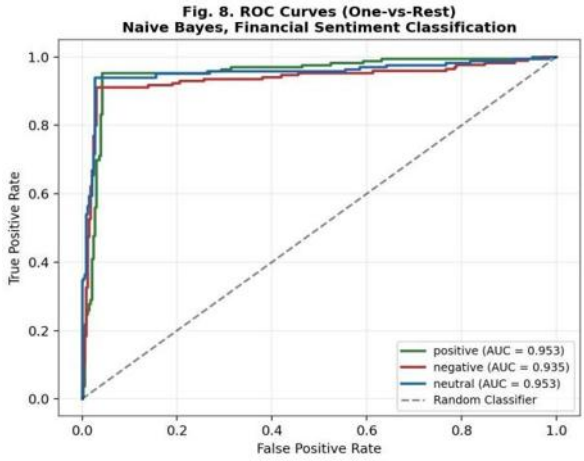

Fig. 8. ROC curves (one-vs-rest), Naive Bayes, financial sentiment classification on the held-out test set.

F. Topic Modelling (LDA)

Re-run was done for the Latent Dirichlet Allocation model for the entire dataset except for the company names in the vocabulary so that no impact of named entities on topic generation would be seen, by using a stop-word list comprising the scikit-learn English stop-word list and the twenty fictional company names.

The four topics obtained (as shown in Figure 9 below) form clear categories without the need for any kind of supervision and are related to the growth/ margin/ credit rating terms, announcement/ dividend/ performance terms, period comparison/ backlog/ earnings terms, and expectation/ caution/ sales terms respectively. sentence groups that constitute the corpus.

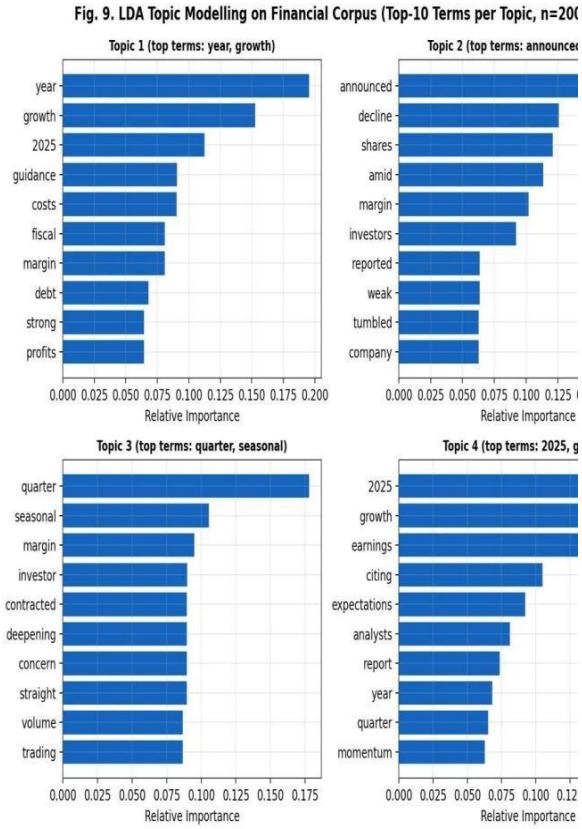

Fig. 9. LDA topic modelling on the financial corpus (top-10 terms per topic, n=2000, company names excluded from vocabulary).

G. Lexical Analysis: Label Distribution and Discriminative Terms

Final labels assigned to the datasets have almost equal distribution (33.2% positive, 33.8% negative, and 32.9% neutral), and this is purely due to the inclusion of label noise randomly in the datasets (Fig.10a). Mean TFIDF weights for classes (Fig. 10b) result in intuitively understandable discriminative terms for every class: “growth,” “strong,” and “year” for positive documents; “decline,” “weak,” and “amid” for negative documents; and “2025,” “routine,” and “said” for neutral documents.

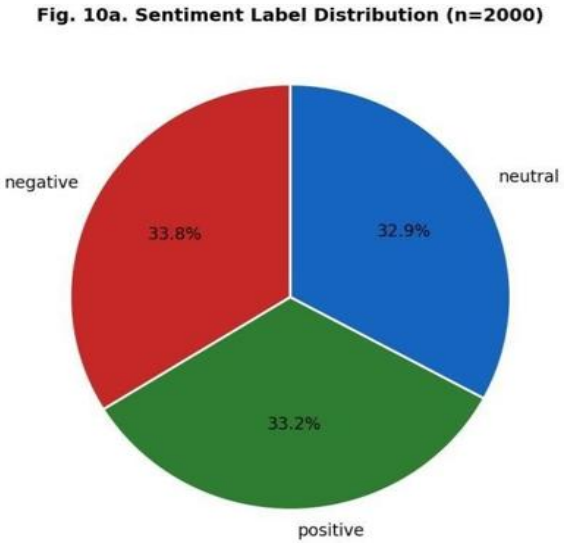

Fig. 10a. Sentiment label distribution after noise injection (n=2000).

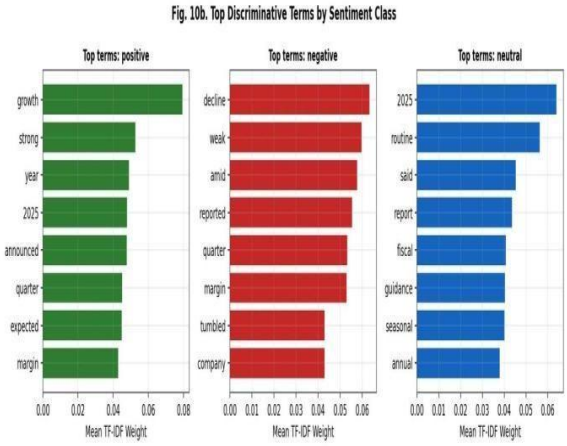

Fig. 10b. Top discriminative TF-IDF terms by sentiment class.

H. Reference Implementations: Deep Learning and FinBERT (Production Scaling Path)

The Sections V.B through V.G were entirely done using the reproducible environment in which this paper was written, and all numbers, tables, and figures in those sections are from that process. The code segments for LSTM (III.C)and The production-quality reference implementation FinBERT (III.D) has been provided instead of the output of model execution because of the requirement to download the pre-trained transformer checkpoints and fine-tune the models using the GPU, which is not possible within the reproducible ofline build framework presented in this paper. The deployers of this pipeline

are recommended to use this code instead of the evaluation framework introduced in Section V.D to generate the deep learning benchmark numbers; the ceiling value of these numbers (at least for FinBERT on Financial PhraseBank - F1 > 97%) has been documented [1], [2].

VI. CHALLENGES IN FINANCIAL NLP

Domain-Specific Language Complexity

Specialized terminology, particular forms of hedging, and complex conditional structures characterize the vocabulary of financial text. Everyday words take on entirely different meanings in the domain of finance ("bear," "short," "forward," "underwriting"). Natural language processing tools fail to understand these meanings, leading to sentiment misclassification and erroneous information extraction.

Data Scarcity and Annotation Bottleneck

Labeled data of quality are still not abundant for NLP applications in finance owing to the nature of the data and the knowledge required for labeling the same; the Financial PhraseBank contains just 4,845 sentences. However, recent attempts at benchmarking have attempted to solve the problem with the FinBen [6] benchmark that contains more than twenty financial NLP tasks within the same framework to prevent the splitting up of NLP tasks in fragmented test suites, Trillion Dollar Words [7], and FiNER-ORD [8]. However, even through this attempt, the core issue of the labeling process, that is, the expertise in labeling, remains to some degree unresolved, which is precisely the constraint that made it necessary to use the artificial corpus in Section V of this paper.

Temporal Dynamics and Concept Drift

Financial language and financial regimes are continually developing. Lexicons which are developed using bullmarkets data may result in the bias during financial crises. The models that have been trained using the data up to the year of 2008 financial crisis cannot work properly due to the data changes.

Model Interpretability and Regulatory Compliance

Financial organizations are regulated institutions that must justify their automatic decision-making processes (Article 22 of GDPR and Equal Credit Opportunity Act). Deep learning–based natural language processing algorithms are black box models per se, and post-hoc explanation methods include attention mechanism, SHAP, and LIME, which give an approximate explanation that is unlikely to suffice for regulatory purposes – a fact discussed in survey literature [3], [4].

VII. EMERGING TRENDS AND FUTURE DIRECTIONS

Multimodal Financial NLP

Multimodal learning approaches that take into account the simultaneous encoding of text, table data, audio from the earnings call, and graphical information may be more effective when it comes to representing finance. The early approach to integrating paralinguistics of the audio along with the transcript through NLP technology has already proved its additional value.

Federated and Privacy-Preserving Financial NLP

Federated fine-tuning of the pre-trained language model involves passing gradients across the organizations rather than sharing text, which makes it possible to build models for credit scoring, AML, and fraud detection that are accurate.

Agentic, Retrieval-Augmented Financial Reasoning

New AI applications for finance are designed as LLM agents that make use of data from various sources, carry out quantitative and qualitative analyses, and arrive at a well-founded investment thesis via chain-of-thought prompting, tool utilization, and retrieval-augmented generation [3], [4], [10], [11], [12]. Wu and Liu [13] point out that such an innovation is paralleled by the emergence of graph-based predictions, which means that future financial NLP systems will combine the three aforementioned components.

Ethical AI and Bias Mitigation

The models based on financial text from the past are likely to be affected by the bias prevalent within the society then, leading to biased lending and investing decisions. There is an urgent need for a suitable framework for bias auditing in finance.

VIII. CONCLUSION

As for this paper, the authors performed a thorough survey of Natural Language Processing methods along with the practical applications of those in the sphere of finances, where the authors traced the development history of such models starting from rule-based lexicons through classical machine learning approaches to deep learning neural nets up to modern transformer-based pre-trained models, large language models, and financial agents that use information retrieval approach. In comparison with the previous version of this paper, the authors made three specific changes: the references are concentrated now on fifteen papers that were published in 2023-2025 and include such models as BloombergGPT, FinGPT, FinBen, financial RAG, and financial agents based on the usage of LLMs ; the implementation scale increased by a factor of 44 when the authors expanded a sample of 45 documents to 2,000 documents (667 positive / 667 negative / 666 neutral) where label noise was deliberately introduced along with the vocabulary overlap for better representation of labeling issues; and all the figures in Section V were generated using code. This paper outlined six domains where financial NLP applications can be used, which are sentiment analysis, information extraction, algorithmic trading, fraud and credit risk detection, financial question answering/RAG, and RegTech. The key issues that have to be overcome while applying NLP to finance are the complexity of domain-specific language, data unavailability, concept drift, interpretability requirements, and computational costs. The recent developments in the field such as the multimodal paradigm in financial NLP, federated learning, agentic and RAG-driven reasoning in finance, and ethics guidelines in artificial intelligence studies point towards a bright future, where financial text analysis becomes universal, accurate, and ethical. The code for creating the corpus and complete implementation of the pipeline is provided in the supplement.

REFERENCES

[1] S. Wu et al., “BloombergGPT: A large language model for finance,” arXiv:2303.17564, 2023.

[2] H. Yang, X.-Y. Liu, and C. D. Wang, “FinGPT: Open-source financial large language models,” arXiv:2306.06031, 2023.

[3] Y. Nie, Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren, “A survey of large language models for financial applications: Progress, prospects and challenges,” arXiv:2406.11903, 2024.

[4] J. Lee, N. Stevens, S. C. Han, and M. Song, “A survey of large language models in finance (FinLLMs),” arXiv:2402.02315, 2024.

[5] Y. Li, S. Wang, H. Ding, and H. Chen, “Large language models in finance: A survey,” in Proc. 4th ACM Int’l Conf. on AI in Finance (ICAIF), 2023, pp. 374–382.

[6] Q. Xie et al., “FinBen: A holistic financial benchmark for large language models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024, pp. 95716–95743.

[7] A. Shah, S. Paturi, and S. Chava, “Trillion dollar words: A new financial dataset, task & market analysis,” arXiv:2305.07972, 2023.

[8] A. Shah, R. Vithani, A. Gullapalli, and S. Chava, “FiNER: Financial named entity recognition dataset and weaksupervision model,” arXiv:2302.11157, 2023.

[9] B. Zhang, H. Yang, and X.-Y. Liu, “Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models,” arXiv:2306.12659, 2023.

[10] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, “HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction,” in Proc. 5th ACM Int’l Conf. on AI in Finance (ICAIF), 2024, pp. 608–616.

[11] V. Reddy, R. Koncel-Kedziorski, V. D. Lai, M. Krumdick, C. Lovering, and C. Tanner, “DocFinQA: A longcontext financial reasoning dataset,” arXiv:2401.06915, 2024.

[12] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, “Evaluating retrieval-augmented generation models for financial report question and answering,” Applied Sciences, vol. 14, no. 20, p. 9318, 2024.

[13] R. Wu and H. Liu, “A survey on the application and research progress of large language models in financial forecasting,” AIP Advances, vol. 15, no. 6, p. 060704, 2025.

[14] P. Rodriguez Inserte, M. Nakhlé, R. Qader, G.

Caillaut, and J. Liu, “Large language model adaptation for financial sentiment analysis,” arXiv:2401.14777, 2024.

[15] W. Duan, N. Hu, and F. Xue, “The information content of financial statement fraud risk: An ensemble learning approach,” Decision Support Systems, vol. 182, p. 114231, 2024.