SentiCraft: A Robust Architecture for Explainable Sentiment Analysis in Finance

Journal:	*IT Professional*
Manuscript ID	ITPro-2025-09-0162
Manuscript Type:	General Interest
Date Submitted by the Author:	29-Sep-2025

Complete List of Authors:	Khosla, Aashish; University of Massachusetts Amherst Manning College of Information & Computer Sciences Sikdar, Debjyoti; John Nash Research Program, John Nash Research Program Sutar, Adyant; John Nash Research Program, John Nash Research Program

Keywords:	I.2 Artificial Intelligence < I Computing Methodologies, I.2.7 Natural Language Processing < I.2 Artificial Intelligence < I Computing Methodologies, O.1.2.a Sentiment analysis < O.1.2 Natural language < O.1 Affect sensing and analysis < O Affective Computing

Article Type: Original Research

SentiCraft: A Robust Architecture for Explainable Sentiment Analysis in Finance

Aashish Khosla, *University of Massachusetts - Amherst, 01003, USA*

Debjyoti Sikdar, *John Nash Research Program, Kolkata, 700060, India*

Adyant Sutar, *John Nash Research Program, Bangalore, 560001, India*

Abstract—SentiCraft is a proposed sentiment analysis architecture that helps to mitigate the existing limitations of models in domain-specific contexts, specifically finance. Widely used models such as VADER rely on lexicons, and FinBERT is influenced by pretrained financial text; both of them face challenges in adapting to emerging events and capturing nuanced market sentiment. SentiCraft is a hybrid architecture that integrates contextual embeddings with sentiment-specific feature filtration, allowing it to score better on interpretability and adaptability. This framework makes SentiCraft well-suited for applications where small variations in language–such as credit downgrade versus market correction—carry disproportionately strong effects. Our paper outlines the architecture of the SentiCraft model, contrasts it with existing models, and suggests future directions in multimodal integration, including combining textual sentiment with paralinguistic signals such as speech tone and nonverbal signals. SentiCraft aims to advance sentiment analysis research by bridging the gap between sentiment analysis accuracy, contextual depth, and practical applications.

The discernment and extraction of emotional tone or opinion conveyed in text through approaches in computational linguistics, machine learning, and natural language processing is commonly referred to as sentiment analysis or opinion mining.[1] This subdiscipline has witnessed significant advancements in techniques and applications, ranging from the adoption of ML and ensemble models to their applications in areas like portfolio management and market predictions.[2] In contemporary finance, businesses and other organizations have access to a plethora of mass media platforms that act as massive repositories of public and expert opinions; consequently, organizations utilize these various sources for gathering insights and attitudes that concern them.

Financial text is different from everyday text because it is denser and has domain-specific words (words that make sense in a sentence only in the financial domain). On the other hand, everyday language is usually broader, colloquial, and more casual. There are also many financial phrases that make little to no sense when thought of in everyday language. For example, *catching a falling knife*. In the financial world, it means buying an asset when its price is dropping rapidly. On the other hand, this seems like dangerous advice to a layperson, as it could be taken at face value. The above sentence gives a glimpse of how phrases in the financial domain have a completely unique meaning when thought of in everyday language. Loughran and McDonald show that negative word lists created for non-financial disciplines, specifically the nonproprietary Harvard Psychosociological Dictionary's negative word list (H4N), misclassify words when analysing the tones of financial texts, and create six finance-specific word lists for the better classification of tone.

Our proposed model, SentiCraft, uses more sophisticated ensemble schemes by incorporating deep learning-based sentiment analysis, followed by a lexicon- or rule-based override for robustness. Concise language with appropriate use of jargon is essential for the efficient communication of complex information.[2] Du et al. also point out that financial language makes heavy use of idioms such as *riding the bull* or *in the red*, often includes quantitative data that must also be considered for sentiment analysis, and is often direction-dependent, as the tone of words like *profit* can be either positive or negative depending on the directional word (e.g., *rise* or *decline*) used in conjunction. Such tricky, idiosyncratic language features can distort measures of sentiment if not properly accounted for; this supports the use of finance-specific lexicons and machine learning (ML) and deep learning models trained on financial data (popularly through transfer learning).[5-6] Popular financial sentiment analysis (FSA) methods include lexicon-based approaches such as the Loughran-McDonald sentiment word lists, developed from a sample of 50,115 10-K filings from 1994 to 2008, ML algorithms like Support Vector Machines (SVM), domain-specific pre-trained language models like FinBERT, and autoregressive decoder architectures like the Generative Pre-trained Transformer (GPT). By aggregating sentiment measurements over time, our proposed model would create sentiment dashboards and feed results into trading systems or analysis tools. Such forms of post-analysis are necessary for the explainability of FSA, which is of tremendous importance given the potentially substantial ramifications of sentiment analysis-driven financial decisions.

PROBLEM STATEMENT

As mentioned earlier, certain features of financial language cloud true sentiment and can cause FSA to fail.[2] Thus, general-purpose sentiment lexicons often fail to accurately extract sentiment from financial texts; for instance, words like *vice* and *liability,* which may be considered negative in general lexicons like the Harvard Psychosociological Dictionary, are not negative in the financial domain.[5] Additionally, most deep learning models are considered *black boxes*, as their decision-making processes are opaque.[8] Given the aforementioned profound risks of FSA-driven decisions, the prospect of model interpretability must be explored. Du et al. mention attempts at explainability through knowledge-graph visualizations, feature relevance, and simplification. Kumar et al. proposed a visualization framework called CLEAR-Trade for stock-market prediction models, utilizing a deconvolution process with unpooling to calculate response maps.[12] SentiCraft, as mentioned earlier, includes sentiment dashboards for explainability and analysis. For instance, it could produce line graphs to visualize trends in overall sentiment toward a certain company. Lexicon-based approaches alone fail to interpret elements such as negation and sarcasm, as they interpret isolated words without context.[13] These limitations can be overcome using deep learning models. On the other hand, complications such as overfitting in deep learning models can be mitigated using the proposed lexicon-based override. SentiCraft's proposed ensemble architecture aims to resolve these FSA problems through mutual mitigation. The purpose of SentiCraft's layered (modular) architecture is to enable reusability and independent development of each module. We consider SentiCraft to be a natural next step in technique-driven FSA research, as its architecture integrates the aforementioned ensemble approach and the interpretability or explainability features recommended by systematic reviews of FSA literature.[2]

FINANCIAL SENTIMENT ANALYSIS – KEY CHALLENGES

Analyzing the emotion of finance-related text is more complex than general sentiment analysis, as the language used is highly domain-specific. This includes the prevalence of jargon, mixed sentiments, and idiomatic expressions that may convey a different meaning if interpreted in a general context (illustrated through examples in table 1). A few of these challenges are captured below:

- *Domain-specific jargon* — Terms such as *bullish*, *bearish*, *long*, *short*, or *overweight* have specialized meanings in the financial context, which differ from their everyday usage and are often misinterpreted by general-purpose models.
- *Mixed sentiment in a single sentence* — Financial statements frequently include both positive and negative information in the same sentence (e.g., *Revenue grew 25%, but operating margins declined*), making it difficult to derive an overall sentiment.
- *Figurative and idiomatic expressions* — Phrases like *dead cat bounce* and *catching a falling knife* require domain knowledge to interpret correctly, as they have no literal meaning in everyday language.
- *Limitations of general-purpose language models* — While they can interpret some common financial terms, they often misclassify context-sensitive terms (e.g., EBITDA, dead cat bounce). Domain-specific models like FinBERT and BloomGPT generally achieve higher accuracy in financial sentiment analysis due to training on finance-specific corpora.

TABLE 1. Sentiment Analysis between General Purpose and Fine Tuned Models

Sentence	General Purpose Model Interpretation	Finance Tuned Model Interpretation	Correct Sentiment
The stock experienced a dead cat bounce after last week's sell-off.	Reads dead cat bounce literally; may think it's about an animal or unrelated event; sentiment is unclear.	Understands *dead cat bounce* as short-lived recovery in a downtrend; negative sentiment.	Negative
The company's earnings beat expectations, leading to a rally in the stock price.	Reads *beat expectations* as a vague phrase; might not connect it to positive outcome; sentiment is unclear or neutral.	Recognizes *earnings beat expectations* as positive financial performance; positive sentiment.	Positive
Investors are worried about the looming debt crisis in the sector.	Focuses on *worried* but may miss specific financial risk context; sentiment negative but might be mild.	Understands *debt crisis* as a serious financial threat; assign strong negative sentiment.	Negative
The merger talks have stalled, causing uncertainty among shareholders.	Interprets *stalled talks* as just a delay; may miss implication of negative impact; sentiment possibly neutral.	Understands stalled merger talks as a negative event leading to uncertainty; sentiment negative.	Negative

RELATED WORK

Early work in financial sentiment analysis focused on highlighting the limitations of general-purpose lexical resources when applied to financial documents. Loughran and McDonald showed that negative word lists, such as the Harvard IV dictionary, lead to substantial misclassification when used on 10-K filings and therefore introduced finance-specific lexicons that better capture the tonal nuance of financial language.[1] Subsequent studies further demonstrated the need for domain-specific approaches by developing entity-level sentiment datasets such as FinEntity,[20] which annotate the spans and the associated sentiments of entities in financial texts to facilitate more fine-grained downstream analysis.[6] In a similar direction, Takale proposed a mixed CNN–GRU architecture to capture long-term dependencies and complex market signals for improved sentiment-based forecasting.[5] With the advent of large language models (LLMs), research has increasingly shifted towards using contextualised representations for financial sentiment. Araci introduced FinBERT and showed that fine-tuning BERT on large-scale financial corpora considerably improves classification performance on benchmarks such as the Financial PhraseBank and FiQA.[2-4] More recent work has extended this model family to include other PLMs such as FinancialBERT and RoBERTa, with comparative studies showing that general-purpose LLMs such as RoBERTa can, in some cases, outperform finance-specific models.[4] In addition to transfer learning and pre-training techniques, ensemble methods have also been shown to improve sentiment classification performance across different feature representations (e.g., Bag of Words and TF–IDF) and class imbalance handling strategies (e.g., SMOTE).[4] Beyond technique-driven research, several studies examine how financial sentiment can be used in market prediction and decision support systems. Kelvin Du et al. classified financial sentiment research into technique-driven and application-driven streams, and argued that the two streams interact in an iterative way where improved models lead to better financial downstream applications (such as market prediction and portfolio management).[3]

PROPOSED METHODOLOGY

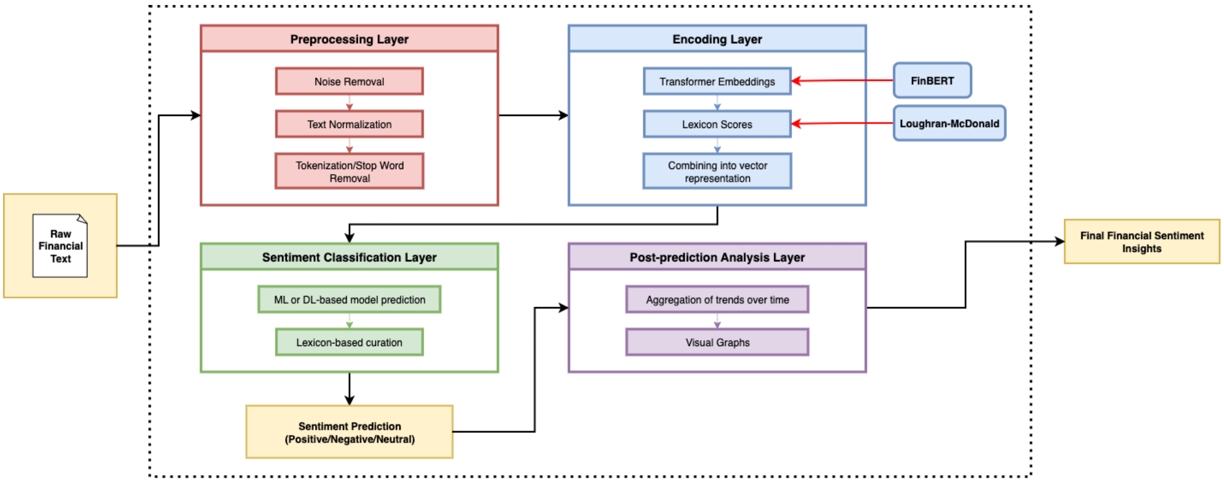
Given the unique challenges of financial sentiment analysis, we propose a new architecture called 'SentiCraft'. Figure 1 demonstrates the multilayer approach, which is explained further below.

Input: Raw Financial Text

The input of raw financial texts includes articles, social media posts, blogs, balance sheets, and every kind of finance document exclusively in English. This unprocessed textual data serves as an initial foundation for our SentiCraft architecture.

Pre-processing Layer

We begin with noise removal. Financial texts and data often contain unimportant and extraneous elements such as emojis, hashtags, metadata, hyperlinks, and other non-linguistic symbols. These additional terms do not



SentiCraft

FIGURE 1. The SentiCraft architecture featuring five distinct layers to aptly pre-process the data, encode it, classify its emotion with two approaches and aggregate trends over time.

contribute to the overall sentiment of the text and may create noise in the embeddings. The noise removal step systematically filters out these unwanted terms, ensuring that the text is reduced to only those terms that contribute to the sentiment of the text. Removing such components ensures that irrelevant tokens do not interfere with downstream processing and hence do not hinder the accuracy of the model.

After removing noise, we then normalize the text. Normalization formats the text in a standard structure by converting all the words to lowercase, handling punctuations, and unifying number formats (e.g., converting 2M to 2 million). This brings consistency to the textual data and reduces lexical variability.

The normalized data is then broken into finer units or words called tokens, for sentiment analysis. Next, stop words (e.g., the, is, of, etc.) – the words that do not contribute to the overall sentiment of the text – are removed. Word-level tokenization (breaking text into a set of individual words called tokens) helps the model to tokenize important words and prioritize them for sentiment analysis. This step not only reduces the computational complexity but also helps in preserving the tokens, which play a pivotal role in sentiment analysis.

Encoding Layer

With pre-processing complete, we proceed to encode the text using transformer embeddings. We utilize FinBERT, a finance domain-specific fine-tuned transformer model of BERT, that generates contextual embeddings to capture subtle sentiment patterns specific to the financial domain. This enables the model to recognize faint signals like negative or positive market signals, market optimism, distinguishing between *bear market* (negative) and *bull market* (positive), and many more.

We supplement this with the Loughran-McDonald lexicon, which provides lists of financial words categorized into sentiment categories like Positive, Negative, Uncertain, Litigious, and Constraining. The output is a sentiment score that portrays the polarity of the text with respect to the financial domain. These sentiment scores, derived from the Loughran-McDonald lexicon, act as specific sentiment indicators; this complements the contextual embeddings with domain-specific understanding.

Finally, we merge the transformer embeddings and lexicon scores together into a unified vector. This hybrid representation balances contextual semantics from transformer embeddings with explicit polarity scores from lexicons, this provides the downstream classifier with a more refined foundation for accurate sentiment analysis.

Sentiment Classification Layer

The combined vector representation is given as input into a classification model (Machine Learning or Deep Learning). This model predicts the sentiment of the input text. Then, these outputs are categorized as Positive, Negative, or Neutral, hinged onto the overall polarity of the textual data.

There may be cases where the lexicon-based polarity cues are strong; they can be used to adjust or change the model's predictions. FinBERT is extremely good at analyzing contextual meaning, but it can seldom underrepresented highly domain-specific cues. For example, terms like *lawsuit* or *credit downgrade* carry an extremely negative effect in the financial domain; this may not be emphasized by a purely context-driven model. But by integrating domain-specific lexicon scoring with transformer embeddings, our architecture helps mitigate this problem. By fusing the Loughran-McDonald lexicon scores along with contextual embeddings, the model ensures that the domain-specific signals are not overshadowed by broader contextual interpretation.

This hybrid representation acts as a correction mechanism, where lexicon-based polarity acts as a grounding factor against the infrequent misclassification by the classification models.

Sentiment Prediction

The model finalizes the text in one of the three categories as a final output: Positive, Negative or Neutral. The model comes to this outcome after interpreting every text's polarity on both contextual semantics and lexicon-based refinement.

Post Prediction Analysis Layer

Financial markets are highly volatile, and the significance of sentiment analysis lies not in a single instance but in the analysis patterns. A single negative article may not influence a trader's decision, but a cluster of articles, blogs, and headlines can. Sentiment analysis can be aggregated across different time windows(hourly, daily, weekly):

- ➔ *Short-term aggregation (minutes to hours)*—This type of sentiment analysis can be used to detect immediate market reaction to positive and negative news.
- ➔ *Medium-term aggregation (days to weeks)*—This type of sentiment analysis can be helpful to identify if negative/positive sentiments about a company are sustained over multiple news cycles. Also helps to identify if short-term shocks are temporary or the start of a bigger trend.
- ➔ *Long-term aggregation (quarters to years)*— This type of sentiment analysis can be helpful to analyze how overall sentiment correlates with reputation and long-term stock performance.

This step transforms sentiment analysis into time-series data, allowing it to be applicable in various fields like risk management, forecasting, and many more. The collective sentiment scores are charted into visual graphs or diagrams to enable analysts to understand the changes in trends over time. For example:

- ➔ *Line charts*— Monitor sentiment polarity over time, marking upward or downward trends.
- ➔ *Heat maps*—Plot the density of sentiments in various industries (e.g., technology, banking, and medicine).
- ➔ *Bar charts or pie charts*—Summarize sentiment distribution for a specific company or market index.

This data visualization can either be used by human analysts to identify emerging risks/opportunities quickly, or by automated systems to detect abnormal spikes that may signal market volatility.

Final Financial Sentiment Insights

The Final Financial Sentiment Insights unify every stage of SentiCraft – from preprocessing to aggregation – into actionable intelligence that balances contextual nuance, domain-specific cues, and temporal trends for informed financial decision-making.

COMPARISON WITH EXISTING APPROACHES

We propose SentiCraft as an alternative, improved architecture for financial sentiment analysis over existing approaches like VADER and FinBERT. In this section, we compare the three approaches based on their methodology, domain suitability, and trade-offs.

In terms of methodology, VADER is a lexicon and rule-based model designed specifically for social media and general-purpose text. FinBERT, by contrast, is a transformer-based model based on BERT, fine-tuned on the financial domain. It learns domain-specific sentiment semantics by contextualizing words, and its output is a probability distribution over sentiment classes (positive, negative, neutral). SentiCraft is a hybrid system that combines context-driven word embeddings with domain-specific knowledge, leveraging attention mechanisms to capture subtle sentiment signals.

Looking at domain suitability, VADER is optimized for general English texts such as tweets, product reviews, and forums. It struggles with domain-specific terms, though it works well in real-time applications due to low computational cost. FinBERT is specifically designed for the financial domain, capturing subtle financial sentiment shifts and outperforming generic sentiment analyzers on financial datasets. SentiCraft is a domain-adapted architecture, specifically designed for financial sentiment analysis, and it provides more fine-grained interpretability compared to FinBERT and VADER.

Considering strengths, VADER is a simple, fast, and beginner-friendly tool that requires no training data and is extremely effective for short, informal text like social media and customer reviews. FinBERT is a context-aware tool that can handle ambiguity in financial language. It achieves state-of-the-art accuracy on financial sentiment benchmarks and is adaptable by fine-tuning on new financial corpora. SentiCraft, in turn, has enhanced domain knowledge, with multi-level aggregation that captures sentiment persistence over time, and it is more explainable than black-box models.

Finally, examining limitations, VADER is weak on long, complex, or domain-specific text, requiring lexicon updates for specialized fields and failing to capture contextual meanings. FinBERT is extremely resource-intensive, requiring GPUs for training and testing, and it has a black-box nature that focuses mainly on short text snippets rather than temporal aggregation (combining sentiment across time to see persistence rather than analyzing each text in isolation). SentiCraft, meanwhile, is still an emerging framework with higher complexity than FinBERT and VADER, requiring curated financial databases, which may limit its scalability.

This analysis underscores why SentiCraft is a worthwhile addition to the space of financial sentiment analysis. It bridges the speed and simplicity of VADER with the conceptual depth of FinBERT, offering an approach that balances interpretability and domain-specific adaptability. Its hybrid architecture contextually grounds sentiment signals while reinforcing them lexically. Although still emerging, the framework's emphasis on interpretability and adaptability positions SentiCraft as a valuable addition to research and real-world applications.

CONCLUSION

For further research, we mainly recommend considering alternatives such as RoBERTa (which outperformed FinBERT in a previously mentioned study)[7] or BERT for the encoding and sentiment classification layers, besides considering special datasets like FinEntity, which annotates entity spans and associated sentiments. It is also crucial to consider the need of multilingual, cross-market sentiment analysis solutions when modifying our proposed architecture for future research. Lastly, we recommend further research on adapting the proposed architecture to analyse multimodal data (text, audio, video, images, etc., including earnings calls and video news releases), which would enable comprehensive consideration of a wide variety of financial media, particularly for content obtained from social media platforms. We believe that our SentiCraft model provides a strong foundation for further development of technique-oriented FSA research. Its flexible architecture can easily be adapted to fit a variety of raw data and FSA algorithms, and its theorized outputs fit both the need for explainable opinion mining and the direct feeding of sentiment features into external applications for downstream tasks such as portfolio management.

ACKNOWLEDGMENTS

We thank Prof. Nirali Arora of the A.P. Shah Institute of Technology, Mumbai, for her support in the drafting, editing, and finalization of our paper. We also thank the John Nash Research Program for the resources provided. This work was supported independently by the authors.

REFERENCES

1. C. Sammut and G. I. Webb, Encyclopedia of Machine Learning. Springer, 2011. (Book)
2. K. Du, F. Xing, R. Mao, and E. Cambria, “Financial Sentiment Analysis: Techniques and Applications,” ACM Computing Surveys, 2024. (Journal)
3. J. Cui, “Behavioral Finance: The Impact of Investor Expectation on Financial Decision-Making,” Advances in Economics, Management, and Political Sciences, 2024. (Journal)
4. A. W. Lo, “The Adaptive Markets Hypothesis,” The Journal of Portfolio Management, 2004. (Journal)
5. T. Loughran and B. McDonald, “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks,” The Journal of Finance, 2011. (Journal)
6. D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” arXiv preprint arXiv:1908.10063, 2019. (Preprint)
7. A. Karanikola, G. Davrazos, C. M. Liapis, and S. Kotsiantis, “Financial sentiment analysis: Classic methods vs. deep learning models,” Intelligent Decision Technologies, 2023. (Journal)
8. Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, “A Survey on Neural Network Interpretability,” IEEE Transactions on Emerging Topics in Computational Intelligence, 2021. (Journal)
9. E. F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work,” The Journal of Finance, vol. 25, no. 2, pp. 383–417, 1970. doi:10.2307/2325486. (Journal)
10. R. J. Shiller, “From efficient markets theory to behavioral finance,” Journal of Economic Perspectives, vol. 17, no. 1, pp. 83–104, 2003. (Journal)
11. B. G. Malkiel, “The efficient market hypothesis and its critics,” Journal of Economic Perspectives, vol. 17, no. 1, pp. 59–82, 2003. (Journal)
12. D. Kumar, G. W. Taylor, and A. Wong, “Opening the black box of financial AI with clear-trade: A class-enhanced attentive response approach for explaining and visualizing deep learning-driven stock market prediction,” arXiv preprint arXiv:1709.01574, 2017. (Preprint)
13. I. Cero, J. Luo, and J. M. Falligant, “Lexicon-based sentiment analysis in behavioral research,” Perspectives on Behavior Science, vol. 47, no. 1, pp. 283–310, 2024. (Journal)
14. K. Al Sharou, Z. Li, and L. Specia, “Towards a better understanding of noise in natural language processing,” in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Sept. 2021, pp. 53–62. (Conference proceedings)
15. Y. Zou and D. Herremans, “PreBit—A multimodal model with Twitter FinBERT embeddings for extreme price movement prediction of Bitcoin,”

Expert Systems with Applications, vol. 233, p. 120838, 2023. (Journal)
16. I. Mutinda, W. Mwangi, and G. Okeyo, “Sentiment analysis of text reviews using lexicon-enhanced BERT embedding (LeBERT) model with convolutional neural network,” Applied Sciences, vol. 13, no. 3, p. 1445, 2023. (Journal)
17. Y. Lei, T. Shen, Y. Cao, and A. Yates, “Enhancing Lexicon-Based Text Embeddings with Large Language Models,” arXiv preprint arXiv:2501.09749, 2025. (Preprint)
18. A. Kumar, “Hybrid Framework Integrating Lexicon and Learning Methods for Enhancing Sentiment Analysis Based on Patients’ Tweets on Medicines,” Journal of Electrical Systems, vol. 20, pp. 2786–2810, May 2024. doi:10.52783/jes.3287. (Journal)
19. I. Polignano, V. Basile, P. Basile, G. Gabrieli, M. Vassallo, and C. Bosco, “A hybrid lexicon-based and neural approach for explainable polarity detection,” Information Processing & Management, vol. 59, no. 5, p. 103058, 2022. (Journal)
20. Y. Tang, Y. Yang, A. H. Huang, A. Tam, and J. Z. Tang, “FinEntity: Entity-level sentiment classification for financial texts,” arXiv preprint arXiv:2310.12406, 2023. (Preprint)

Aashish Khosla is a researcher in the Department of Computer Science at the University of Massachusetts Amherst. His research interests include natural language processing, sentiment analysis, machine learning, and quantitative finance.

Debjyoti Sikdar is a researcher affiliated with the John Nash Research Program in Kolkata, India. His research interests include computational linguistics, language revitalization, machine learning and natural language processing.

Adyant Sutar is a researcher affiliated with the John Nash Fellowship in Bangalore, India, focusing on creating AI to perform sentiment analysis in the financial domain, exploring how artificial intelligence can robustly interpret emotions and opinions in financial markets.