A Survey on Financial Sentiment Analysis Using Natural Language Processing

Ms. S. Abinaya

Assistant Professor, Department of Information Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India.

abinayas@skacas.ac.in

Ms. D. Mariyam Beevi

Department of Information Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamil Nadu, India.

mariyambeevi619@gmail.com

Abstract:

Financial sentiment analysis has emerged as a crucial research area due to the growing influence of textual data on financial decision-making processes. With the rapid expansion of online financial news, social media platforms, analyst reports, and corporate disclosures, Natural Language Processing (NLP) techniques play a significant role in extracting sentiment-driven insights from unstructured financial text. This survey presents a comprehensive overview of existing approaches used in financial sentiment analysis, highlighting traditional methods, machine learning models, and recent deep learning-based techniques. The paper discusses commonly used datasets, domain-specific sentiment lexicons, and evaluation metrics employed in financial contexts. Furthermore, it analyzes the challenges unique to financial language, such as domain ambiguity, sarcasm, temporal relevance, and data imbalance. Recent advancements involving transformer-based models and hybrid approaches are also examined to understand their impact on prediction accuracy and robustness. Finally, the survey outlines open research issues and future directions, emphasizing the need for explainable models and real-time sentiment

analysis systems for practical financial applications. This work aims to provide researchers and practitioners with a structured understanding of the current state and potential growth of financial sentiment analysis using NLP.

Keywords: Financial Sentiment Analysis, NLP, Machine Learning, Deep Learning.

Introduction:

The financial domain is highly sensitive to information, where public opinion, news events, and market narratives can significantly influence investment decisions and market behavior. With the exponential growth of digital financial content—such as news articles, earnings reports, social media discussions, and analyst commentaries—there is an increasing need for automated techniques to analyze and interpret large volumes of unstructured textual data. Financial sentiment analysis aims to identify and quantify opinions or emotions expressed in financial text and assess their impact on market movements and economic indicators.

Natural Language Processing (NLP) has become a key enabling technology for extracting meaningful insights from financial text. Early sentiment analysis approaches relied on manually crafted lexicons and rule-based methods; however, these techniques often struggle to capture the contextual and domain-specific nuances of financial language. To address these limitations, machine learning and deep learning models have been widely adopted, offering improved adaptability and performance. More recently, transformer-based architectures such as BERT and its financial variants have demonstrated significant advancements in understanding complex linguistic patterns within financial documents.

Despite these developments, financial sentiment analysis remains challenging due to factors such as ambiguous terminology, rapidly changing market contexts, and the presence of noise in user-generated content. This survey systematically reviews existing methodologies, datasets, and evaluation techniques used in financial sentiment analysis, highlighting their strengths and limitations. By synthesizing current research trends and identifying open challenges, this paper aims to provide a comprehensive reference for researchers and practitioners working in the intersection of finance and natural language processing.

Financial Text Characteristics:

Financial text exhibits distinct linguistic and structural properties that differentiate it from general-domain textual data. These characteristics pose unique challenges for sentiment analysis and require specialized Natural Language Processing (NLP) techniques for accurate interpretation. One of the primary features of financial text is its **domain-specific vocabulary**, which includes technical terms, abbreviations, and jargon whose sentiment polarity may differ from common usage. For instance, words such as *"liability," "volatile,"* or *"risk"* do not necessarily convey negative sentiment in financial contexts but instead describe standard market conditions.

Another notable characteristic is **contextual and temporal dependency**. The sentiment of financial information is often influenced by time-sensitive events such as earnings announcements, policy changes, or geopolitical developments. As a result, the same textual expression may have varying sentiment implications depending on the prevailing market conditions. Additionally, financial texts frequently contain **implicit sentiment**, where opinions are conveyed through numerical indicators, comparisons, or forecasts rather than explicit emotional language.

Financial data sources are also highly diverse, ranging from formal documents like annual reports and regulatory filings to informal content such as social media posts and online forums. This diversity leads to significant variation in writing style, tone, and reliability. Moreover, financial texts may include **speculative, uncertain, or forward-looking statements**, introducing ambiguity and increasing the complexity of sentiment classification.

Furthermore, financial sentiment datasets often suffer from **class imbalance**, with neutral sentiment dominating over positive and negative categories. This imbalance can bias learning models and reduce classification performance. Understanding these unique characteristics is essential for designing robust sentiment analysis systems tailored to the financial domain.

Types of Financial Sentiment Analysis:

Financial sentiment analysis can be categorized into several types based on the level of granularity, data sources, and analytical objectives. Each type serves a specific purpose in understanding market

behavior and supporting financial decision-making.

1. Document-Level Sentiment Analysis

Document-level sentiment analysis determines the overall sentiment expressed in an entire financial document, such as news articles, earnings reports, or analyst notes. This approach assumes that the document conveys a single dominant sentiment and is commonly used to assess the general market outlook or corporate performance. While effective for high-level analysis, it may overlook sentiment variations within different sections of the text.

2. Sentence-Level Sentiment Analysis

Sentence-level analysis focuses on identifying sentiment at the individual sentence level. This method is particularly useful for financial documents that contain mixed opinions, such as annual reports or market commentaries. By analyzing sentiment sentence by sentence, finer-grained insights can be obtained compared to document-level analysis.

3. Aspect-Based Financial Sentiment Analysis

Aspect-based sentiment analysis examines sentiment toward specific financial entities or aspects, such as company performance, revenue growth, risk factors, or stock price movements. This type enables a more detailed understanding of how particular financial attributes influence sentiment, making it valuable for investors and analysts seeking targeted insights.

4. Entity-Based Sentiment Analysis

Entity-based sentiment analysis focuses on identifying sentiment associated with specific financial entities, including companies, stocks, indices, or economic indicators. This approach is widely used in stock market prediction and portfolio management, where sentiment toward individual entities plays a critical role.

5. Fine-Grained Sentiment Classification

Unlike binary or ternary classification (positive, negative, neutral), fine-grained sentiment analysis assigns multiple sentiment levels, such as very positive or slightly negative. This approach provides more nuanced sentiment interpretation, which is beneficial in financial risk assessment and trend analysis.

6. Emotion-Oriented Financial Sentiment Analysis

This type goes beyond polarity detection to identify specific emotions such as optimism, fear, uncertainty, or confidence expressed in financial text. Emotion-based analysis is particularly useful for understanding market volatility and investor behavior during critical economic events.

Table I: Types of Financial Sentiment Analysis

Source	Data	Usage
News Media	Articles	Market trend prediction
Social Media	Tweets	Investor sentiment
Earnings Calls	Transcripts	Corporate outlook

Methodologies:

Financial sentiment analysis employs a variety of methodologies ranging from traditional lexicon-based approaches to advanced machine learning and deep learning techniques. Each methodology offers distinct advantages and faces specific challenges when applied to the financial domain.

1. Lexicon-Based Approaches

Lexicon-based methods rely on pre-defined dictionaries of sentiment words with associated polarity scores. In the financial domain, domain-specific lexicons such as **Loughran-McDonald** or **Financial Sentiment Lexicon** are commonly used. These approaches compute sentiment by aggregating scores of words or phrases in a text. While lexicon-based methods are interpretable and simple to implement, they often fail to capture context, sarcasm, or nuanced financial language.

2. Machine Learning-Based Approaches

Machine learning (ML) techniques leverage features extracted from text to train classifiers for sentiment prediction. Common algorithms include **Support Vector Machines (SVM), Naive Bayes, Random Forests**, and **Logistic Regression**. Feature engineering plays a crucial role in ML-based

methods, with techniques such as **Bag-of-Words, TF-IDF, and n-grams** frequently applied. ML models generally outperform lexicon-based methods but require labeled datasets and careful handling of class imbalance in financial data.

3. Deep Learning-Based Approaches

Deep learning models automatically learn hierarchical representations of textual data and have shown significant improvements in financial sentiment prediction. Techniques such as **Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks**, and **Gated Recurrent Units (GRUs)** are widely used. These models capture contextual information and sequential dependencies, making them suitable for complex financial texts.

4. Transformer-Based Approaches

Recent advancements in transformer architectures, such as **BERT, FinBERT, RoBERTa, and GPT-based models**, have revolutionized financial sentiment analysis. These models leverage attention mechanisms to capture long-range dependencies and contextual nuances, achieving state-of-the-art performance on various financial datasets. Fine-tuning pre-trained transformer models on domain-specific corpora further enhances their accuracy and robustness.

5. Hybrid Approaches

Hybrid methodologies combine lexicon-based, ML, and deep learning techniques to leverage the strengths of each. For example, sentiment scores from lexicons can be used as additional features in machine learning models or to guide deep learning architectures. Hybrid approaches aim to improve interpretability, accuracy, and adaptability in diverse financial scenarios.

6. Evaluation Metrics

Evaluation of financial sentiment models typically employs metrics such as **accuracy, precision, recall, F1-score, and Area Under the Curve (AUC)**. Domain-specific challenges, such as class imbalance and temporal relevance, may require adjusted metrics or validation strategies to ensure reliable performance.

Datasets:

Datasets play a crucial role in training, validating, and benchmarking financial sentiment analysis models. Unlike general-domain sentiment datasets, financial datasets require domain-specific text with appropriate annotations to capture market-related

sentiment accurately. They can be broadly categorized based on their sources, content type, and annotation style.

1. Financial News Datasets

Financial news datasets consist of news articles, press releases, and corporate announcements. Popular datasets include:

- **Financial PhraseBank**: Contains 4,841 sentences from financial news, manually labeled as positive, negative, or neutral.
- **Reuters News Dataset**: Provides historical news articles with stock price labels, often used for supervised sentiment and market prediction tasks.

2. Social Media and Microblogging Datasets

Investor opinions on social media platforms can influence market trends. Notable datasets include:

- **StockTwits Dataset**: Comprises messages from StockTwits, labeled for sentiment related to specific stocks.
- **Twitter Financial Dataset**: Contains tweets mentioning companies or financial events, annotated for positive, negative, or neutral sentiment.

3. Analyst Reports and Earnings Call Transcripts

These datasets include financial reports, conference call transcripts, and earnings announcements. They are valuable for understanding expert opinions and forward-looking statements:

- **ENRON Email Dataset**: A collection of corporate emails with potential sentiment signals related to company performance.
- **Earnings Call Transcripts**: Texts of quarterly earnings discussions, often annotated manually or using semi-supervised methods for sentiment polarity.

Applications:

Applications of financial sentiment analysis include stock price prediction, portfolio optimization, algorithmic trading, fraud detection, and risk management.Applications of financial sentiment analysis include stock price prediction, portfolio optimization, algorithmic trading, fraud detection, and risk management.Applications of financial

sentiment analysis include stock price prediction, portfolio optimization, algorithmic trading, fraud detection, and risk management.Applications of financial sentiment analysis include stock price prediction, portfolio optimization, algorithmic trading, fraud detection, and risk management.

Challenges:

Financial sentiment analysis faces several unique challenges due to the complex nature of financial language, data, and market dynamics. Key challenges include:

1. Domain-Specific Vocabulary

Financial texts contain technical terms, abbreviations, and jargon whose sentiment may differ from general language. Words like *"liability"*, *"depreciation"*, or *"volatile"* may not inherently indicate negative sentiment, making standard sentiment tools less effective.

2. Contextual Ambiguity

The sentiment of financial content often depends on the surrounding context. The same statement can be positive or negative depending on market conditions, company performance, or temporal factors.

3. Implicit Sentiment

Financial sentiment is frequently conveyed indirectly through numbers, forecasts, or comparisons rather than explicit emotional words. Detecting these subtleties requires sophisticated models capable of interpreting numerical and textual information together.

4. Data Scarcity and Imbalance

High-quality, annotated financial datasets are limited. Additionally, neutral sentiment often dominates over positive and negative categories, leading to class imbalance that affects model performance.

5. Noisy and Informal Data

Social media and microblogging platforms contain informal language, abbreviations, and sarcasm. Noise in such data can reduce the accuracy of sentiment analysis.

6. Multilingual and Cross-Market Challenges

Financial information is often available in multiple languages, and sentiment expression can vary across cultures and markets. This complicates multilingual or cross-market sentiment analysis.

Future Directions:

Advancing financial sentiment analysis requires addressing existing challenges while

leveraging emerging technologies. The following research directions offer promising avenues for improvement:

1. Hybrid Modeling Approaches:

Combining lexicon-based methods with machine learning and transformer-based models can enhance both accuracy and interpretability. Hybrid approaches can leverage domain knowledge while benefiting from deep contextual representations.

2. Multimodal Sentiment Analysis

Integrating textual data with numerical financial indicators, charts, or news images can provide richer insights into market sentiment. Multimodal analysis can capture information that is missed when considering text alone.

3. Real-Time Sentiment Monitoring

Developing models capable of processing streaming financial data from news, social media, and market feeds in real-time can provide immediate insights for efficient real-time systems remain an open research area.

4. Explainable and Transparent Models

The demand for interpretability in financial applications is growing. Research on explainable AI (XAI) can help build trust in automated sentiment predictions, especially for regulatory compliance and high-stakes decision-making.

5. Domain Adaptation and Transfer Learning

Adapting pre-trained models to specific financial domains or markets can overcome data scarcity and improve generalization. Transfer learning across markets, languages, or asset classes is a key area for future research.

6. Multilingual and Cross-Market Analysis

Global financial markets generate data in multiple languages and formats. Developing models that can understand and transfer sentiment across languages and regions can improve the scalability of financial sentiment systems.

Conclusion:

Financial sentiment analysis has become a critical tool for understanding market trends, investor behavior, and corporate performance. This survey presented a comprehensive overview of the field, including the unique characteristics of financial text, types of sentiment analysis, datasets, and methodologies ranging from

lexicon-based approaches to advanced transformer-based models. Despite significant progress, the field continues to face challenges such as domain-specific language, implicit sentiment, data scarcity, and the need for model interpretability.

Future research directions, including hybrid modeling, multimodal analysis, real-time monitoring, and explainable AI, offer promising opportunities to enhance the accuracy, scalability, and applicability of financial sentiment analysis. By addressing these challenges and pursuing emerging methodologies, researchers and practitioners can develop more reliable and actionable systems that support investment decisions, risk assessment, and market prediction. This survey serves as a reference for understanding the current landscape and guiding future advancements in the intersection of finance and natural language processing.

Acknowledgement:

Ms S Abinaya, B.Sc., MCA., Assistant Professor of Information Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore. She has 7 months of teaching experience. She has organized various Competitions, conducted Quiz. She enriched her teaching career by attending several Faculty Development Programme, Webinar, Seminar etc.

I, MARIYAM BEEVI D pursuing a Bachelor of Science in Information Technology at Sri Krishna Adithya College of Arts and Science. I presented many papers in various colleges and attended many workshops.

References:

[1] D. T. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” *arXiv preprint*, Aug. 2019.

[2] “Natural language processing in finance: A survey,” *Information Fusion*, vol. 115, Mar. 2025.

[3] S. Bansal, B. K. Singh, and M. K. Jain, “An Analysis of Different Sentiment Analysis Models on Financial Text Using Transformer,” *Proc. ICAIEHS*, Oct. 2025.

[4] A. Rodriguez Inserte, M. Nakhlé, R. Qader, G. Caillaut, and J. Liu, “Large Language Model Adaptation for Financial Sentiment Analysis,” *arXiv*, Jan. 2024.

[5] O. Shobayo, S. Adeyemi-Longe, O. Popoola, and B. Ogunleye, “Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4 and Logistic

Regression: A Data-Driven Approach," *arXiv*, Dec. 2024.

[6] A. Atsiwo, "Financial Sentiment Analysis: Leveraging Actual and Synthetic Data for Supervised Fine-tuning," *arXiv*, Dec. 2024.

[7] T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 2011 (Loughran-McDonald sentiment word lists).

[8] **Financial PhraseBank dataset**, Hugging Face (Financial PhraseBank used extensively in financial sentiment work).

[9] J. Lopez Roldan, *Sentiment Analysis in the Financial PhraseBank Dataset: An Approach to FinBERT*, ResearchGate, Apr. 2025.

[10] S. *et al.*, "Prediction of the Stock Market From Linguistic Phrases," *Journal of Database Management*, Jan. 2023.