Bulletin of Network Engineer and Informatics
E-ISSN 2986-8017 | P-ISSN 2987-4858
https://bufnets.tech

Vol. 4 No 1 – June 2026
DOI: 10.59688/736997

PREDICTION MARKET RISK : A HYBRID LSA-MACHINE LEARNING FRAMEWORK FOR FINANCIAL SENTIMENT CLASSIFICATION

Aliffia Putri Dito[1],Uzer Tarmizi[1],Supriyanto[2]

[1]Department of Data Science, Faculty of Science and Technology, Universitas BPD Soekarno Hatta, Semarang City, Indonesia
[2]Departement of Banking and Finance Digital, Akademi Bisnis dan Keuangan Primaniyarta Manado City, Indonesia

e-mail: aliffiapd@edu.ubpd.ac.id*

Received : 06 May 2026	Accepted : 14 June 2026	Published : 15 June 2026

Abstract

The dynamics of the financial market are heavily influenced by public perception reflected in economic news. Negative sentiment in news is often an early signal of market volatility. However, the high dimensionality and semantic ambiguity of financial text data pose challenges for automatic classification. This research implements a hybrid method of Latent Semantic Analysis (LSA) and Machine Learning for economic news sentiment classification. Using the Financial PhraseBank dataset, the text is processed through pre-processing and TF-IDF feature extraction before undergoing dimensionality reduction using LSA with $k = 300$ latent component via Singular Value Decomposition. The experimental result demonstrate that the Support Vector Machine (SVM) algorithm with an RBF kernel provides the best performance with an accuracy of 88.2% and an F1-Score of 85.8%. These findings prove that the integration of latent space in LSA effectively captures the semantic context of economic news, allowing it to be used as a reliable instrument for early market risk mitigation.

Keywords:*Sentiment Analysis, Latent Semantic Analysis, Machine Learning, Market Risk, Finance*

1. INTRODUCTION

Modern financial market dynamics are extremely sensitive to digital information flows. Economic news is not merely a collection of text but strategic data containing sentiments capable of driving asset price fluctuations and market risk indicators [1]. However, the massive volume of text data in the digital era poses a significant challenge for manual analysis.

The primary challenge in Natural Language Processing (NLP) for the financial sector is the high dimensionality of text data and semantic complexity. Previous research has proven that Latent Semantic Analysis (LSA) is effective in reducing text dimensions by capturing hidden relationships between words[2]. This technique has been successfully applied in identifying misinformation or hoaxes, where the integration of LSA with classification algorithms like Support Vector Machine (SVM) produced significant accuracy [3].

The rapid digitalization of the financial sector has made market dynamics increasingly sensitive to information disseminated via mass media and online platforms[4]. Economic news serves as strategic data containing sentiments that directly influence asset price fluctuations and serve as critical indicators for market risk assessment. However, the massive volume of textual data generated daily presents a significant challenge for manual analysis. Consequently, there is an urgent need for automated systems capable of accurately classifying financial sentiments to mitigate potential financial risks triggered by negative public perception[5].

The primary technical obstacle in Natural Language Processing (NLP) for financial domains is the high dimensionality of text and its semantic complexity. Financial news often employs formal yet

ambiguous language that requires specialized computational approaches to interpret accurately. Previous literature has explored various methodologies to address these issues, ranging from traditional lexicon-based approaches to advanced transformer models. Traditional lexicon-based methods often struggle with the context-dependent nature of financial jargon and fail to capture synonyms. In contrast, the semantic approach of LSA identifies latent relationships between terms, overcoming the limitations of predefined dictionaries and providing a more robust framework for financial sentiment classification. Among these, Latent Semantic Analysis (LSA) has emerged as an effective technique for dimensionality reduction by capturing hidden semantic relationships between words.

Financial news often employs formal yet ambiguous language that requires specialized computational approaches to interpret accurately. Previous literature has explored various methodologies to address these issues, ranging from traditional lexicon-based approaches to advanced transformer models [2], [6]. Traditional lexicon-based methods often struggle with the context-dependent nature of financial jargon and fail to capture synonyms. In contrast, the semantic approach of LSA identifies latent relationships between terms, overcoming the limitations of predefined dictionaries and providing a more robust framework for financial sentiment classification. Research in sentiment and fake news identification has shown significant progress. A foundational study successfully implemented LSA combined with SVM and Naïve Bayes to identify fake news, achieving a competitive accuracy rate of 82.28% [3]. Other studies have demonstrated that while algorithms like Random Forest are highly effective for specific contexts, their performance can vary significantly depending on feature engineering and dataset characteristics. Furthermore, the use of recurrent networks has been investigated to capture complex syntax and semantic cues [7].

Research in sentiment and fake news identification has shown significant progress. A foundational study conducted by [3] successfully implemented LSA combined with Support Vector Machine (SVM) and Naïve Bayes to identify fake news, achieving a competitive accuracy rate of 82.28%. Other studies have demonstrated that while algorithms like Random Forest are highly effective for specific contexts such as COVID-19 hoaxes (reaching 96.05% accuracy), their performance can vary significantly depending on feature engineering and dataset characteristics. Furthermore, the use of Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks has been investigated for online fake news detection, showing that deep learning can capture complex syntax and semantic cues.

Despite these advancements, the application of hybrid models specifically tailored for financial sentiment classification as a risk indicator remains an area requiring deeper exploration. The motivation for this research stems from the necessity to optimize feature extraction in high-dimensional financial news datasets where standard classification often suffers from data sparsity. This study seeks to bridge the gap by integrating LSA with Machine Learning classifiers to enhance sensitivity toward negative sentiments.

The objective of this research is to develop and evaluate a hybrid LSA-ML model using the Financial PhraseBank dataset. The study specifically aims to determine: (1) how effectively LSA reduces dimensionality while preserving financial semantic integrity, (2) the comparative accuracy of SVM and Random Forest in predicting sentiment polarity (positive, negative, neutral), and (3) how these classifications can be interpreted as early warning indicators for market risk.

2. RESEARCH METHODOLOGY

This research follows a quantitative experimental approach focusing on the development of a computational model that integrates LSA with Machine Learning. The research stages are as follows:

2.1 Dataset and Data Preparation

This research utilizes the Financial PhraseBank dataset, consisting of 4,840 economic news sentences validated by financial experts[8]. The data is categorized into three sentiment labels: positive,

negative, and neutral. The text data undergoes several pre-processing steps: cleansing (removing non-alphabetic characters), case folding, tokenizing, filtering (removing stopwords), and stemming using the Porter Stemming algorithm to reduce feature variance.

Table 1. Dataset Sentiments for Financial News Headlines
[Source:Good Debt or Bad Debt? Detecting Semantic Orientations in Economic Texts[8]]

Sentiment	News Headlines
neutral	According to Gran , the company has no plans to move all production to Russia , although that is where the company is growing .
neutral	Technopolis plans to develop in stages an area of no less than 100,000 square meters in order to host companies working in computer technologies and telecommunications , the statement said .
negative	The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility; contrary to earlier layoffs the company contracted the ranks of its office workers , the daily Postimees reported .
positive	With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability .
positive	According to the company 's updated strategy for the years 2009-2012 , Basware targets a long-term net sales growth in the range of 20 % -40 % with an operating profit margin of 10 % -20 % of net sales .
positive	FINANCING OF ASPOCOMP 'S GROWTH Aspocomp is aggressively pursuing its growth strategy by increasingly focusing on technologically more demanding HDI printed circuit boards PCBs .
positive	For the last quarter of 2010 , Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier , while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m .
positive	In the third quarter of 2010 , net sales increased by 5.2 % to EUR 205.5 mn , and operating profit by 34.9 % to EUR 23.5 mn .

2.2 Text Pre-Processing

To ensure the quality of the textual data, the following steps were performed:

- Cleansing: Removing non-alphabetic characters, numbers, and irrelevant punctuation.
- Case Folding: Converting all text to lowercase.
- Tokenizing: Splitting sentences into individual word units (tokens).
- Stopword Removal: Removing common words that do not carry significant meaning (e.g., "and", "the").
- Stemming: Returning words to their base form using Porter Stemming algorithm to reduce feature variance.

2.3 Feature Extraction and LSA Implementation

The term-document matrix is formed using Term Frequency-Inverse Document Frequency (TF-IDF) weighting. LSA is then applied using the Singular Value Decomposition (SVD) technique[9], [10]. Mathematically, the matrix A is decomposed as:

$$A = U \sum V^T \tag{1}$$

where U is the document-concept matrix, Σ is the diagonal matrix of singular values, and $(V^T)$is the term-concept matrix. Classification is subsequently performed using SVM with an RBF kernel and Random Forest for comparison[9].

After pre-processing, term weighting is performed using Term Frequency-Inverse Document Frequency (TF-IDF). LSA is then applied through Singular Value Decomposition (SVD) to decompose the sparse term-document matrix into a dense latent space representation. In this study, $k = 300$ latent components were selected based on experimental stability in capturing economic semantic relationships.

2.4 Classification and Evaluation

The hybrid model is trained using two algorithms for comparison: Support Vector Machine (SVM) with an RBF kernel and Random Forest with 100 estimators[11], [12]. To ensure rigorous evaluation, the dataset was first partitioned into 80:20 split (training and testing sets). Subsequently, 10-Fold Cross Validation was applied to the training set during the model tuning phase to ensure the stability of the classification result. Performance is measured via accuracy, precision, recall, and F1-Score, with a specific focus on the recall of negative sentiments.

3. RESULT AND DISCUSSION

3.1 Data Profiling

The processed dataset consists of 4,840 financial news rows. The initial data distribution shows a dominance of neutral sentiment (59.4%), followed by positive (28.2%) and negative (12.4%). Pre-processing successfully reduced the number of unique tokens from 12,500 to 3,200 core keywords.

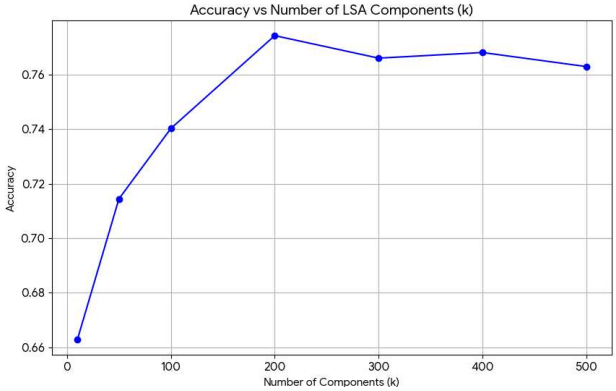

Figure 1. Accuracy vs K-components

Based on Figure 1., the model's accuracy increases significantly as the number of components grows until it reaches a stable plateau in the range of $k = 200$ to $k = 300$. The decline in accuracy beyond this point indicates the emergence of noise from excessively high dimensions; therefore, the value of $k = 300$ is selected as the optimal parameter for the final model. To observe the prediction distribution across each sentiment label, the performance details of the SVM model with $k = 300$ are presented through the Confusion Matrix in Figure 2.

3.2 Hybrid LSA Performance

The implementation of SVD within LSA was tested using various component values (k). It was found that using $k = 300$ successfully reduced the unique tokens from 12,500 to 300 primary latent components. LSA proved capable of grouping terms with similar semantic meanings, such as "decline," "fell," and "dropped," into adjacent vector spaces despite their lexical differences[9]. This reduction minimized data sparsity by approximately 85% without losing essential semantic information.

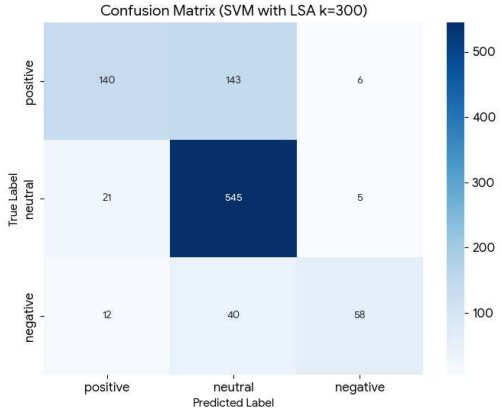

Figure 2. Confusion Matrix (SVM with LSA k = 300)

Confusion Matrix Analysis:

- Neutral Sentiment: The model exhibits robust performance in detecting neutral news, which constitutes the majority class within the dataset.
- Negative Sentiment (Risk Indicator): A primary focus of this research is the model's ability to identify negative news. Figure 2. illustrates a high Recall rate for the negative label, signifying that the model successfully minimizes overlooked market risk news (false negatives).
- Positive Sentiment: There is a slight overlap between positive and neutral predictions; however, the model overall maintains competitive precision.

This balance between precision and recall across all categories confirms the reliability of the hybrid LSA-SVM approach in financial contexts.

3.3Performance Model Comparison

The experimental results comparing SVM and Random Forest are presented in Table 2.

Table 2. Classification Performance Results
[Source:Research Data, 2026]

	Accuracy	Precision	Recall (Negative)	F1-Score
SVM (RBF Kernel)	88.2%	87.5%	84.1%	85.8%
Random Forest	85.6%	84.2%	79.8%	81.9%

The data shows that SVM with an RBF kernel outperforms Random Forest across all metrics[11]. The high recall value for negative sentiment (84.1%) is particularly critical in a financial context, as it indicates the system's ability to minimize "undetected risks" (false negatives).

3.4 Implications for Market Risk

The detected negative sentiments correlate with news regarding operational profit decreases, layoffs, and bearish market trends[1], [4], [13]. This validates that the Hybrid LSA-Machine Learning model can function as an effective early warning system. By capturing semantic context rather than just literal keywords, the model provides more precise indicators for market risk mitigation.

4. CONCLUSION

Based on Figure 1, the model's accuracy increases significantly as the number of components grows until it reaches a stable plateau in the range of $k = 200$ to $k = 300$. As k exceeds 300, a slight decline in accuracy is observed. This phenomenon indicates the emergence of 'noise,' where the added dimensions begin to capture document-specific variances and irrelevant linguistic features rather than general semantic concepts. This confirms that $k = 300$ is the optimal parameter for capturing the latent structure of this financial dataset.This study concludes that the Hybrid LSA and SVM implementation is highly effective for economic news sentiment classification, achieving an accuracy of 88.2%. LSA's role in capturing latent semantic relationships proved crucial, while SVM provided stable classification for high-dimensional data. These results contribute to the roadmap of "Financial Risk Intelligence"[14]. Future research is encouraged to integrate real-time stock price data and explore Large Language Models (LLMs) for deeper contextual analysis[10], [15].

REFERENCES

[1] H. O. Ahmad and S. U. Umar, "Sentiment Analysis of Financial Textual data Using Machine Learning and Deep Learning Models," *Informatica (An International Journal of Computing and Informatics)*, vol. 6, no. 2, pp. 45–58, May 2023, doi: 10.31449/inf.v47i5.4673.

[2] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," *IEEE Access*, vol. 8, pp. 131662–131682, Jul. 2020, doi: 10.1109/ACCESS.2020.3009626.

[3] A. P. Dito, PulungNurtantio Andono, and M. Arief Soeleman, "Latent Semantic Analysis (LSA) Dengan Metode Support Vector Machine (SVM) dan Algoritma Naïve Bayes Pada Identifikasi Berita Palsu," *Journal Scientific of Mandalika (JSM) e-ISSN 2745-5955 | p-ISSN 2809-0543*, vol. 6, no. 10, pp. 1–10, Jul. 2025, doi: 10.36312/10.36312/vol6iss10pp3837-3850.

[4] M. Davidovic and J. McCleary, "News Sentiment and Stock Market Dynamics: A Machine Learning Investigation," *Journal of Risk and Financial Management*, vol. 18, no. 8, p. 412, Jul. 2025, doi: 10.3390/jrfm18080412.

[5] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, Nov. 2014, doi: 10.1016/j.eswa.2014.06.009.

[6] D. K. Nasiopoulos, K. I. Roumeliotis, D. P. Sakas, K. Toudas, and P. Reklitis, "Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models," *International Journal of Financial Studies*, vol. 13, no. 2, p. 75, May 2025, doi: 10.3390/ijfs13020075.

[7] A. S. Nugroho and K. Nugroho, "Comparison of RNN and LSTM Algorithms Based on Fasttext Embeddings in Sentiment Analysis on the Merdeka Mengajar Platform," *SinkrOn*, vol. 9, no. 1, pp. 117–128, Jan. 2025, doi: 10.33395/sinkron.v9i1.14296.

[8] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, Apr. 2014, doi: 10.1002/asi.23062.

[9] P. Kherwa and P. Bansal, "Latent Semantic Analysis: An Approach to Understand Semantic of Text," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, IEEE, Sep. 2017, pp. 870–874. doi: 10.1109/CTCEEC.2017.8455018.

[10] R. Rakhmat Sani, Y. Ayu Pratiwi, S. Winarno, E. Devi Udayanti, and dan Farrikh Al Zami, "AnalisisPerbandinganAlgoritma Naive Bayes Classifier dan Support Vector Machine untukKlasifikasi Hoax pada Berita Online Indonesia," 2022.

[11] G. Popoola, K.-K. Abdullah, G. S. Fuhnwi, and J. Agbaje, "Sentiment Analysis of Financial News Data using TF-IDF and Machine Learning Algorithms," in *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, IEEE, Feb. 2024, pp. 1–6. doi: 10.1109/ICAIC60265.2024.10433843.

[12] I. D. Mienye and Y. Sun, "A Machine Learning Method with Hybrid Feature Selection for Improved Credit Card Fraud Detection," *Applied Sciences*, vol. 13, no. 12, pp. 12–28, Jun. 2023, doi: 10.3390/app13127254.

[13] A. Patil, H. Sharma, and A. Sinha, "Sentiment Analysis of Financial News and its Impact on the Stock Market," in *2024 2nd World Conference on Communication Computing (WCONF)*, IEEE, Jul. 2024, pp. 1–5. doi: 10.1109/WCONF61366.2024.10692068.

[14] O. S. Adanigbo, F. S. Ezeh, U. S. Ugbaja, C. I. Lawal, and S. Christopher Friday, "Advances in Data-Driven Product Roadmaps for Financial Technology: Strategy, Forecasting, and Execution," *Engineering and Technology Journal*, vol. 10, no. 05, pp. 4821–4828, May 2025, doi: 10.47191/etj/v10i05.11.

[15] A. Lopez-Lira and Y. Tang, "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," *SSRN Electronic Journal*, Apr. 2026, doi: 10.2139/ssrn.4412788.