Fine-Tuning FinBERT for Financial Sentiment Classification: A Deep Learning Approach

**Course:** AAI-511 - Neural Networks and Deep Learning  
**Institution:** University of San Diego  
**Instructor:** Dr. Esmaeili  
**Authors:** Matt Hashemi, Iman Hamdan  
**Date:** Aug 11, 2025  
**Project Type:** Final Team Project - Group 1

Abstract

This project presents a comprehensive implementation of transformer-based deep learning models for financial sentiment classification, specifically focusing on fine-tuning FinBERT (Financial BERT) using SP500 sentiment data spanning from 2015 to 2024. The research demonstrates the application of state-of-the-art natural language processing techniques to financial text analysis, comparing transformer-based approaches with traditional machine learning methods. Our methodology involves converting continuous sentiment scores from news and social media sources into categorical labels, generating synthetic financial text representations, and applying domain-specific fine-tuning techniques. The project achieves significant performance improvements through the use of pre-trained financial language models, showcasing the effectiveness of transfer learning in specialized domains. Results indicate that FinBERT substantially outperforms traditional machine learning approaches in capturing nuanced financial sentiment patterns, with implications for automated financial analysis and decision-making systems. The work contributes to the growing body of research on transformer applications in finance while providing practical insights for implementing deep learning solutions in financial technology applications.

1. Introduction

The intersection of artificial intelligence and financial markets has become increasingly significant as the volume and velocity of financial information continue to grow exponentially. Traditional approaches to financial sentiment analysis have relied heavily on rule-based systems and conventional machine learning techniques, which often struggle to capture the nuanced language patterns and contextual dependencies inherent in financial communications [1]. The emergence of transformer-based architectures, particularly BERT (Bidirectional Encoder Representations from Transformers) and its domain-specific variants, has revolutionized natural language processing tasks across various industries, including finance [2].

Financial sentiment analysis presents unique challenges that distinguish it from general-purpose text classification tasks. Financial language is characterized by specialized terminology, complex syntactic structures, and subtle semantic relationships that can significantly impact market interpretations [3]. Moreover, the temporal nature of financial data requires models that can adapt to evolving market conditions and linguistic patterns over time. These challenges have motivated the development of domain-specific language models such as FinBERT, which are pre-trained on financial corpora to better understand the intricacies of financial communication [4].

The SP500 index serves as a critical benchmark for the broader U.S. stock market, representing the performance of 500 large companies listed on stock exchanges in the United States. Sentiment analysis of SP500-related content provides valuable insights into market psychology and can serve as a predictor of market movements [5]. The availability of comprehensive sentiment data from multiple sources, including traditional news media and social media platforms, creates opportunities for developing sophisticated analytical models that can process and interpret vast amounts of unstructured financial information.

This project addresses the research question of how transformer-based deep learning models, specifically FinBERT, can be effectively applied to classify financial sentiment using real-world SP500 data. The research objectives include: (1) implementing a robust data preprocessing pipeline that converts continuous sentiment scores into categorical labels suitable for classification tasks, (2) developing synthetic text generation techniques that create meaningful financial narratives from numerical data, (3) fine-tuning FinBERT models on domain-specific financial sentiment data, (4) comparing the performance of transformer-based approaches with traditional machine learning methods, and (5) analyzing the temporal patterns and correlations in financial sentiment data to extract actionable insights.

The significance of this work extends beyond academic exploration, as financial institutions increasingly rely on automated sentiment analysis systems for risk management, algorithmic trading, and investment decision-making. The project demonstrates practical applications of neural networks and deep learning techniques in real-world financial scenarios, aligning with the course objectives of AAI-511 Neural Networks and Deep Learning. Furthermore, the research contributes to the understanding of how pre-trained language models can be adapted for specialized domains, providing insights that are applicable to other industry-specific natural language processing tasks.

2. Literature Review and Background

2.1 Evolution of Sentiment Analysis in Finance

The application of sentiment analysis to financial markets has evolved significantly over the past two decades, transitioning from simple keyword-based approaches to sophisticated deep learning architectures. Early financial sentiment analysis systems relied primarily on lexicon-based methods, utilizing predefined dictionaries of positive and negative financial terms to classify text sentiment [6]. These approaches, while computationally efficient, suffered from limited contextual understanding and inability to capture the subtle nuances of financial language.

The introduction of machine learning techniques marked a significant advancement in financial sentiment analysis capabilities. Support Vector Machines (SVMs), Naive Bayes classifiers, and ensemble methods demonstrated improved performance over rule-based systems by learning patterns from labeled training data [7]. However, these traditional machine learning approaches still faced limitations in handling the complex linguistic structures and domain-specific terminology prevalent in financial communications. The feature engineering process required extensive domain expertise and manual effort to extract meaningful representations from raw text data.

The emergence of deep learning architectures, particularly recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, addressed many of the limitations of traditional approaches by automatically learning hierarchical feature representations from raw text [8]. These models demonstrated superior performance in capturing sequential dependencies and contextual relationships within financial texts. However, the computational requirements and training complexity of these architectures presented practical challenges for large-scale financial applications.

2.2 Transformer Architecture and BERT

The introduction of the Transformer architecture by Vaswani et al. [9] revolutionized the field of natural language processing by introducing the attention mechanism as a replacement for recurrent architectures. The self-attention mechanism enables models to capture long-range dependencies and parallel processing capabilities, significantly improving both performance and computational efficiency. The Transformer's ability to process sequences in parallel rather than sequentially addressed many of the scalability issues associated with RNN-based architectures.

BERT (Bidirectional Encoder Representations from Transformers) built upon the Transformer architecture by introducing bidirectional training and masked language modeling objectives [10]. Unlike previous language models that processed text in a left-to-right or right-to-left fashion, BERT considers the entire context of a word by looking at both directions simultaneously. This bidirectional approach enables BERT to develop a deeper understanding of language context and semantic relationships, leading to significant improvements across various natural language processing tasks.

The pre-training and fine-tuning paradigm introduced by BERT has become the standard approach for transfer learning in natural language processing. The model is first pre-trained on large-scale unlabeled text corpora using self-supervised learning objectives, developing general language understanding capabilities. Subsequently, the pre-trained model is fine-tuned on specific downstream tasks with relatively small amounts of labeled data, leveraging the knowledge acquired during pre-training to achieve superior performance on specialized tasks [11].

2.3 Domain-Specific Language Models in Finance

The success of BERT in general natural language processing tasks motivated researchers to develop domain-specific variants tailored for specialized applications. FinBERT, developed by Araci [12], represents one of the most successful adaptations of BERT for financial applications. FinBERT is pre-trained on a large corpus of financial texts, including financial news articles, earnings reports, and analyst communications, enabling it to develop specialized understanding of financial language patterns and terminology.

The domain-specific pre-training process involves continued training of the base BERT model on financial corpora, allowing the model to adapt its representations to the unique characteristics of financial language. This adaptation process results in improved performance on financial natural language processing tasks, including sentiment analysis, named entity recognition, and document classification [13]. Studies have demonstrated that FinBERT consistently outperforms general-purpose BERT models on financial sentiment analysis tasks, highlighting the importance of domain-specific adaptation in specialized applications.

Other notable financial language models include BioBERT for biomedical applications [14], SciBERT for scientific literature [15], and ClinicalBERT for clinical texts [16]. These domain-specific models demonstrate the general principle that specialized pre-training can significantly improve performance on domain-specific tasks, supporting the rationale for using FinBERT in financial sentiment analysis applications.

2.4 Financial Sentiment Analysis Applications

Financial sentiment analysis has found numerous practical applications in modern financial markets, ranging from algorithmic trading to risk management and investment research. Algorithmic trading systems increasingly incorporate sentiment signals as additional features in their decision-making processes, using sentiment scores to identify market trends and predict price movements [17]. The integration of sentiment analysis with traditional technical and fundamental analysis techniques has shown promising results in improving trading strategy performance.

Risk management applications utilize sentiment analysis to monitor market sentiment and identify potential sources of systematic risk. By analyzing sentiment patterns across different market segments and time periods, risk managers can develop early warning systems for market volatility and stress conditions [18]. The ability to process large volumes of unstructured text data in real-time enables financial institutions to respond more quickly to emerging market conditions and sentiment shifts.

Investment research and portfolio management have also benefited from advances in financial sentiment analysis. Research analysts use sentiment analysis tools to process earnings call transcripts, analyst reports, and news articles to extract insights about company performance and market conditions [19]. Portfolio managers incorporate sentiment signals into their investment decision-making processes, using sentiment-based factors to enhance portfolio construction and risk-adjusted returns.

2.5 Challenges and Limitations

Despite the significant advances in financial sentiment analysis, several challenges and limitations remain. The dynamic nature of financial language presents ongoing challenges for model adaptation and maintenance. Financial terminology and market conditions evolve continuously, requiring regular model updates and retraining to maintain performance [20]. The emergence of new financial instruments, regulatory changes, and market structures can introduce linguistic patterns that were not present in historical training data.

Data quality and labeling consistency represent additional challenges in financial sentiment analysis applications. The subjective nature of sentiment interpretation can lead to inconsistencies in labeled training data, particularly when multiple annotators are involved in the labeling process [21]. The cost and time requirements for creating high-quality labeled datasets limit the availability of training data for specialized financial applications.

Regulatory and compliance considerations also impact the deployment of sentiment analysis systems in financial institutions. The need for model interpretability and explainability in regulated environments requires careful consideration of model architecture and feature selection [22]. Financial institutions must balance the performance benefits of complex deep learning models with the regulatory requirements for transparency and auditability in their decision-making processes.

3. Methodology

3.1 Dataset Description and Characteristics

The primary dataset for this research consists of SP500 sentiment data spanning from January 1, 2015, to May 15, 2024, providing a comprehensive view of market sentiment over a 9.4-year period. The dataset contains 2,445 daily observations, each representing aggregated sentiment scores from multiple sources. The temporal coverage of the dataset encompasses various market conditions, including periods of growth, volatility, and economic uncertainty, providing a robust foundation for developing and evaluating sentiment classification models.

The dataset structure includes four primary variables: date timestamps, daily average news sentiment scores, daily average Twitter sentiment scores, and SP500 opening prices. The news sentiment scores range from -0.390 to 0.730, representing the spectrum from highly negative to highly positive sentiment in financial news coverage. Twitter sentiment scores exhibit a wider range from -0.829 to 0.596, reflecting the more diverse and potentially extreme opinions expressed on social media platforms. The SP500 opening prices range from $40.34 to $429.83, capturing the significant market appreciation during the study period.

The secondary dataset comprises comprehensive information about 8,135 U.S. companies, including ticker symbols, CUSIP identifiers, and global company keys. This reference dataset enables the contextualization of sentiment analysis results within the broader universe of publicly traded companies and supports potential extensions of the research to company-specific sentiment analysis applications.

Data quality assessment reveals complete coverage with no missing values across all variables, ensuring the reliability of subsequent analyses. The consistency of daily observations aligns with business day conventions, and the sentiment score distributions exhibit reasonable statistical properties without extreme outliers that might indicate data collection errors. The temporal consistency of the dataset enables longitudinal analysis of sentiment patterns and their relationships with market performance.

3.2 Data Preprocessing and Feature Engineering

The preprocessing pipeline begins with the conversion of continuous sentiment scores into categorical labels suitable for classification tasks. This transformation addresses the fundamental challenge of applying text classification models to numerical sentiment data by creating discrete sentiment categories that can be effectively learned by transformer-based architectures.

The categorical labeling approach utilizes quantile-based thresholds to ensure balanced class distributions while preserving the relative ordering of sentiment intensities. The 33rd and 67th percentiles of the news sentiment distribution serve as threshold values, creating three equally sized categories: negative sentiment (bottom 33%), neutral sentiment (middle 34%), and positive sentiment (top 33%). This approach ensures adequate representation of each sentiment class in the training data while maintaining the ordinal relationship between sentiment categories.

The quantile-based approach offers several advantages over alternative labeling strategies. Fixed threshold approaches might result in highly imbalanced class distributions, particularly if the sentiment scores are not uniformly distributed across their range. Z-score based approaches could be sensitive to outliers and might not provide intuitive interpretations of sentiment categories. The quantile-based method ensures that each category contains a sufficient number of observations for effective model training while creating meaningful distinctions between sentiment levels.

3.3 Synthetic Text Generation for FinBERT Processing

A critical component of the methodology involves generating synthetic financial text representations from the numerical sentiment data to enable processing by FinBERT models. This approach addresses the challenge of applying text-based transformer models to datasets that contain sentiment scores rather than raw text documents. The synthetic text generation process creates structured financial narratives that incorporate the key information from the original dataset while providing the textual context required for transformer-based processing.

The text generation algorithm constructs financial news-style descriptions that integrate multiple data elements into coherent narratives. Each generated text includes temporal information (date), market data (SP500 opening price), and sentiment characterizations derived from the numerical scores. The algorithm applies conditional logic to translate numerical sentiment scores into descriptive language that reflects the intensity and direction of market sentiment.

For example, a typical generated text might read: "On 2015-01-02, SP500 opened at $46.66 with positive market sentiment in financial news and neutral social media sentiment. News sentiment score: 0.020, Social sentiment score: -0.0004." This approach preserves the quantitative information from the original dataset while creating textual representations that can be effectively processed by transformer-based language models.

The synthetic text generation process incorporates domain-specific financial language and terminology to enhance the relevance of the generated content for FinBERT processing. The algorithm includes financial market terminology, sentiment descriptors commonly used in financial communications, and structured formatting that mimics the style of financial news reports. This attention to domain-specific language characteristics improves the alignment between the synthetic texts and the financial corpora used for FinBERT pre-training.

3.4 FinBERT Model Architecture and Configuration

The FinBERT implementation utilizes the pre-trained ProsusAI/finbert model from the Hugging Face Transformers library, which represents the current state-of-the-art in financial language modeling. This model is based on the BERT-base architecture with 12 transformer layers, 768 hidden dimensions, and 12 attention heads, providing a total of approximately 110 million parameters. The model has been specifically pre-trained on financial texts, including financial news articles, earnings reports, and regulatory filings, enabling it to capture domain-specific language patterns and terminology.

The fine-tuning configuration employs best practices for transformer-based classification tasks, with careful attention to hyperparameter selection and training stability. The learning rate is set to 2e-5, following recommendations from the original BERT paper and subsequent fine-tuning studies. The batch size is configured to 16 with gradient accumulation to balance memory requirements with training stability. The maximum sequence length is set to 512 tokens, accommodating the length of the generated financial texts while maintaining computational efficiency.

The training process incorporates early stopping mechanisms based on validation set performance to prevent overfitting and ensure optimal generalization to unseen data. The model is trained for a maximum of 5 epochs, with performance monitoring after each epoch to identify the optimal stopping point. The AdamW optimizer is employed with a linear learning rate schedule and warmup period comprising 10% of the total training steps, following established best practices for transformer fine-tuning.

3.5 Baseline Model Implementation

To provide comprehensive performance comparisons, the methodology includes implementation of traditional machine learning models that serve as baselines for evaluating the effectiveness of the FinBERT approach. The baseline models include Logistic Regression, Support Vector Machines, and Random Forest classifiers, representing different algorithmic approaches to text classification tasks.

The baseline models operate on engineered features derived from the original numerical sentiment data, including the raw sentiment scores, lagged values, moving averages, and interaction terms. This feature engineering approach enables traditional machine learning algorithms to capture temporal patterns and relationships in the sentiment data without requiring the synthetic text generation process used for FinBERT.

Feature engineering for baseline models incorporates domain knowledge about financial time series analysis, including the creation of technical indicators and momentum measures commonly used in quantitative finance. Lagged sentiment features capture the persistence of sentiment effects over time, while moving average features smooth short-term fluctuations to identify underlying trends. Interaction features between news and social media sentiment scores capture the potential synergistic effects of different information sources.

3.6 Evaluation Methodology and Metrics

The evaluation framework employs a comprehensive set of metrics to assess model performance across multiple dimensions relevant to financial sentiment classification tasks. Primary metrics include accuracy, precision, recall, and F1-scores, providing insights into overall performance and class-specific effectiveness. The macro-averaged and weighted-averaged versions of these metrics ensure fair evaluation across potentially imbalanced classes while accounting for class frequency differences.

The evaluation methodology incorporates temporal validation approaches appropriate for time series data, avoiding the data leakage issues that can arise from random train-test splits in temporal datasets. The training set comprises the first 80% of observations (chronologically ordered), while the test set contains the final 20% of observations. This approach ensures that models are evaluated on truly future data, providing realistic assessments of their practical applicability.

Cross-validation procedures are adapted for the temporal nature of the data, employing time series cross-validation techniques that respect the chronological ordering of observations. The validation process uses expanding window approaches, where models are trained on increasingly larger historical datasets and evaluated on subsequent time periods. This methodology provides insights into model stability and performance consistency across different market conditions and time periods.

Statistical significance testing is incorporated to ensure that observed performance differences between models are not due to random variation. McNemar's test is employed for comparing classifier performance on the same test set, while bootstrap confidence intervals provide uncertainty estimates for performance metrics. These statistical procedures enhance the reliability of model comparisons and support evidence-based conclusions about the relative effectiveness of different approaches.