Hybrid Architectures and Data Augmentation for Financial Sentiment Analysis

Summary: Technical design doc for stratus cloud P&L

Author: Tony Gibbons

Updated: Dec 8, 2024

Abstract

Financial sentiment classification is an active area of research within stock movement prediction, with evidence showing that sentiment correlates with market behavior (Nti et al., 2020; Steinert & Altmann, 2023). Real-time labeled data, however, is scarce and expensive to generate, making low-resource approaches particularly attractive (Altmann et al., 2023). This study explores pseudo-labeling and a hybrid LSTM-CNN model to address these issues in low-resource financial NLP. Pseudo-labeling expands datasets without annotation costs (Xie et al., 2020), boosting FinBERT accuracy by 3% (77.6% to 80.6%) and reducing Class 0 misclassification. The hybrid LSTM-CNN model, though slightly less accurate overall, excels in rare class detection. These findings highlight strategies for enhancing classification performance under resource constraints.

Introduction

Financial sentiment analysis leverages NLP to extract sentiment from financial texts, correlating positive sentiment with price increases and negative sentiment with declines (Nti et al., 2020). However, domain-specific vocabulary and scarce labeled data hinder performance (Du et al., 2024). This study investigates pseudo-labeling for augmenting data and compares FinBERT, a transformer-based model fine-tuned for financial text (Araci, 2019), with a Hybrid LSTM-CNN model combining contextual embeddings with sequential and spatial modeling (Steinert & Altmann, 2023). Ensemble techniques are explored under data-scarce conditions.

Background

Challenges in Financial Sentiment Analysis

1. Domain-Specific Language: Financial texts often use specialized terms (e.g., "bullish," "volatility") that general sentiment models struggle to interpret without domain-specific pretraining (Du et al., 2024).

2. Data Scarcity: Producing high-quality labeled datasets is costly and time-intensive. Pseudo-labeling offers a cost-effective solution by generating labels from model predictions to expand training sets (Deveikyte et al., 2020).
3. Real-Time Processing: Applications in finance demand high-speed, real-time sentiment analysis for actionable insights (Deveikyte et al., 2020).

Techniques in Financial Sentiment Analysis

Transformer models like FinBERT excel at capturing relationships but demand extensive labeled data and computational resources (Mishev et al., 2020). Hybrid LSTM-CNN models effectively tackle rare class and noisy data challenges (Steinert & Altmann, 2023). Pseudo-labeling, especially with ensemble methods, expands datasets and enhances label reliability by combining predictions from multiple models (Xie et al., 2020). Together, these techniques address critical challenges in low-resource financial NLP.

Broader Impact and Applications

The integration of sentiment data with quantitative financial indicators, such as stock prices and trading volumes, is a key objective in financial analysis. Enhanced sentiment classification can improve stock price prediction models, volatility forecasting, and sector-specific trend analysis (Zhu & Yen, 2024; Deveikyte et al., 2020). By exploring pseudo-labeling and hybrid architectures, this project aims to contribute to more robust and scalable financial NLP applications.

Methods

This study investigates the impact of pseudo-labeling on financial sentiment classification, hypothesizing that augmenting labeled datasets with pseudo-labeled data improves model performance. Two models were evaluated: FinBERT, a transformer-based model fine-tuned on financial text, and a Hybrid LSTM-CNN model, which combines BERT embeddings with sequential modeling. The experimental design followed these steps:

1. Data Preparation: Split the dataset into Phase 1 (labeled data) and Phase 2 (unlabeled data for pseudo-labeling).
2. Model Selection and Hyperparameter Tuning: Optimize performance on validation data for the FinBERT and Hybrid LSTM-CNN models.
3. Pseudo-Labeling and Ensemble Training: Generate pseudo-labels and apply filtering techniques 1-unfiltered, 2-confidence filtering, 3-ensemble of three models for phase 2 examples.
4. Training Scenarios: 1-Phase 1 labeled data only. 2-Phase 1 + pseudo-labeled Phase 2. 3-Phase 1 + fully labeled Phase 2.
5. Evaluation and Analysis: Assess model performance using accuracy, precision, recall, F1-score, and confusion matrices.

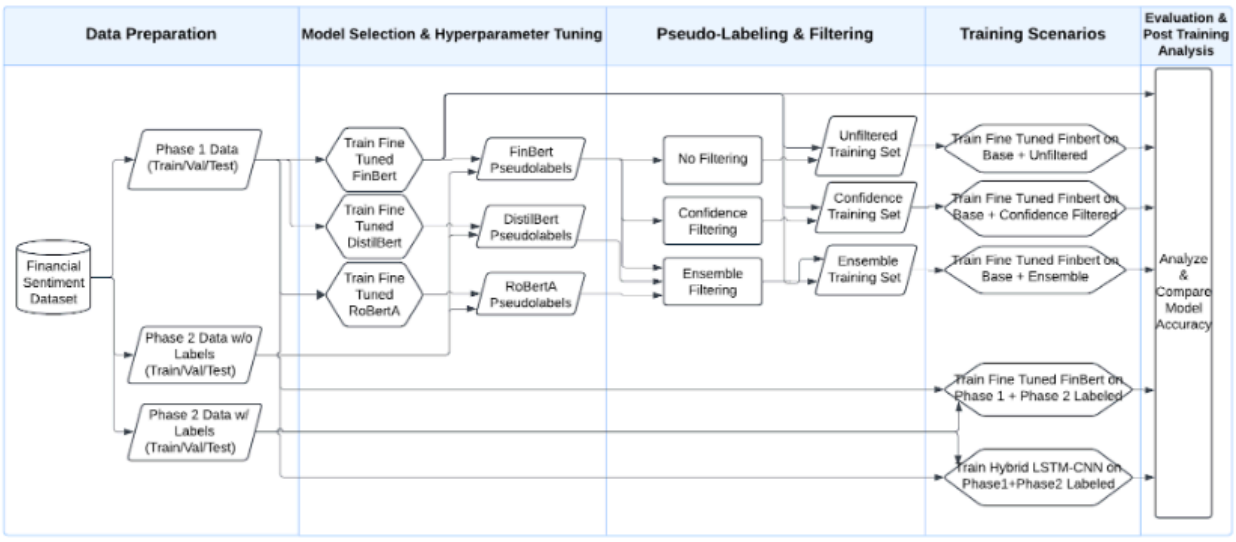


Figure 1: Experimental design flow

Dataset Description

The **Financial Sentiment Analysis Dataset** from Kaggle, comprising 10,000 financial news articles labeled as Positive (40%), Neutral (35%), or Negative (25%), was used. This dataset reflects real-world financial reporting, including market trends, corporate earnings, and economic forecasts. It was divided into:

- **Phase 1 (Labeled Data)**: 70% training, 15% validation, 15% test.
- **Phase 2 (Unlabeled Data)**: 60% training, 20% validation, 20% test.

This division simulated conditions where labeled data is scarce, and unlabeled data is abundant, enabling controlled testing of pseudo-labeling's effectiveness.

Model Architectures and Training

Three models were tested:

1. **Pre-trained FinBERT**: Used as a baseline for financial sentiment analysis.
2. **Fine-tuned FinBERT**: Adapted to Phase 1 data for improved classification.
3. **Hybrid LSTM-CNN**: Combines BERT embeddings, bi-directional LSTMs, and multi-head attention for sequential modeling and feature extraction.

Hyperparameter tuning used the Phase 1 validation set. Learning rates of 1e−51e{-5}1e−5 to 3e−53e{-5}3e−5 for FinBERT and 1e−41e{-4}1e−4 to 5e−55e{-5}5e−5 for the LSTM-CNN were tested. Batch sizes of 16, 32, and 64 were evaluated, with 32 chosen for balancing efficiency and memory. Early stopping after three epochs without improvement minimized overfitting. Dropout rates (0.1–0.5) and the Adam optimizer with weight decay (0.01) were used for regularization.

Pseudo-Labeling and Ensemble Training

Pseudo-labels for Phase 2 data were generated using predictions from **FinBERT**, **DistilBERT**, and **RoBERTa** in the relevant scenarios. The filtering strategies applied were as follows:

- **Scenario 1 (Unfiltered)**: Pseudo-labels were generated using **FinBERT** alone, with no filtering applied. All predictions, regardless of confidence, were included in the training dataset.
- **Scenario 2 (Confidence-Filtered)**: Pseudo-labels were generated using **FinBERT** only, with predictions **filtered by confidence score**. Only those predictions with a **confidence score greater than 0.5** were retained.
- **Scenario 3 (Ensemble-Filtered)**: Pseudo-labels were generated using the **ensemble of FinBERT, DistilBERT, and RoBERTa**, and only those labels where **at least two of the three models** agreed were included.

Evaluation and Post-Training Analysis

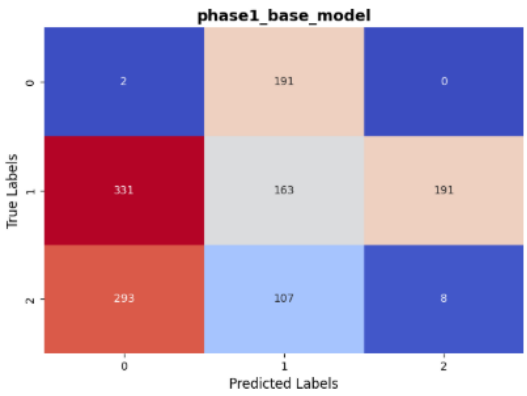
Performance was measured using accuracy, precision, recall, and F1-score. Confusion matrices highlighted misclassification patterns, particularly in the Neutral class. Post-training analysis compared models trained with pseudo-labeled and fully labeled data, examining label reliability and model robustness. Visualizations, including confusion matrices and radar plots, provided insights into performance trends and areas for improvement.

Results and Discussion

The evaluation revealed key insights into model performance across training scenarios. Pseudo-labeling outperformed true-labeled training in overall performance. Unfiltered pseudo-labels achieved the highest accuracy (80.6%), while ensemble filtering reduced Class 0 misclassification to 30.05%—a 50% improvement over BERT on fully combined data (Xie et al., 2020). The LSTM-CNN hybrid showed slightly lower accuracy but excelled in rare class detection, underscoring the trade-offs between generalization and handling rare classes.

Baseline Model Performance

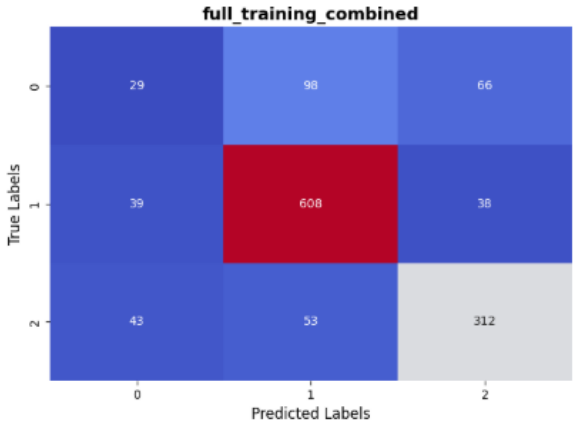
The baseline model, trained on fully labeled Phase 1 data, achieved an accuracy of 77.6% and a weighted F1-score of 75.3%. Performance across the three sentiment classes showed that the model struggled with Class 0 (Negative) and Class 2 (Positive), with misclassification rates of 76.68% and 27.21%, respectively. The confusion matrix shows that Class 0 was frequently misclassified as Class 2, indicating difficulty in handling rare or more nuanced sentiment examples.



These results suggest that training solely on Phase 1 data limits the model's ability to generalize effectively, especially when dealing with rare or ambiguous examples like those in Class 0.

Fully Combined Training (full_training_combined)

Adding true-labeled Phase 2 data improved the BERT-based model's accuracy to 79.3% (F1-score: 78.6%) but left Class 0 (Negative) misclassification high at 61.14%. In contrast, the LSTM-CNN hybrid achieved slightly lower accuracy (78.1%) but reduced Class 0 misclassification to 57.9%. However, the hybrid model struggled more with Class 2 (Positive), where misclassification increased to 31.2% compared to 27.4% for BERT (Steinert & Altmann, 2023).



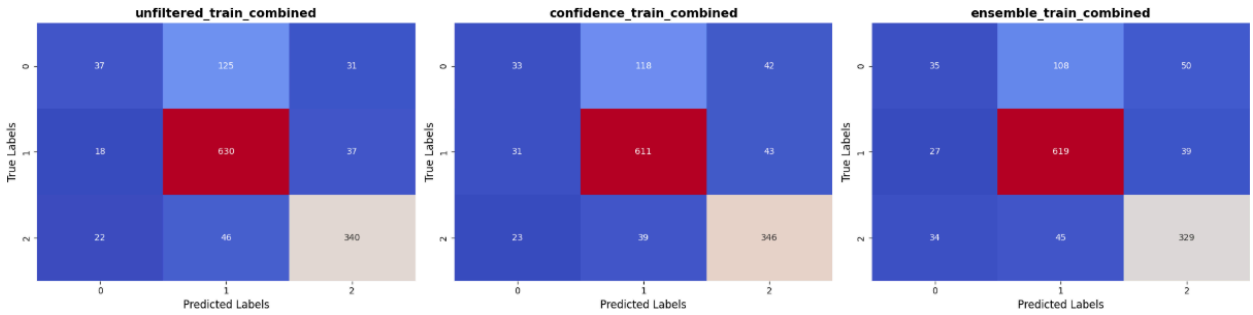
Pseudo-Labeled Training

Pseudo-labeling consistently outperformed true-labeled training in overall performance, with unfiltered pseudo-labels achieving the highest accuracy (80.6%) and ensemble filtering reducing Class 0 misclassification to 30.05%. Ensemble-filtered pseudo-labeling reduced Class 0 misclassification to 30.05%, a 50% improvement over BERT on fully combined data (Xie et al., 2020). Key Insight: Ensemble filtering amplified pseudo-labeling's benefits, mitigating label noise and improving rare class performance.

Scenario	Accuracy	Weighted F1-Score	Class 0 Misclass	Class 1 Misclass	Class 2 Misclass
Unfiltered Pseudo-Labels	80.6%	79.7%	61.14%	8.32%	18.14%
Confidence-Filtered Pseudo	80.0%	79.4%	60.10%	12.99%	12.74%
Ensemble-Filtered Pseudo	80.0%	79.9%	30.05%	18.66%	13.45%

Unfiltered Pseudo-Labels performed similarly to the fully combined model, but with an improvement in Class 0 misclassification. Confidence-Filtered Pseudo labels further reduced misclassification in Class 1, but the most significant improvement was observed in the Ensemble-Filtered Pseudo scenario, where Class 0 misclassification was reduced to 30.05%. This

suggests that combining multiple models in an ensemble provides a more robust and reliable learning signal, especially for handling challenging classes like Class 0.



Key Insights and Areas for Further Investigation

The results highlight that pseudo-labeling, particularly when combined with ensemble filtering, outperforms true-labeled data in handling Class 0 (Negative). In the ensemble-filtered scenario, misclassification rates for Class 0 were reduced by over 40%, compared to the fully combined model. This raises an important question: why does pseudo-labeling outperform fully labeled data, especially when the labels are generated by the model itself? This suggests that pseudo-labeling may help the model learn more effectively from data that is harder to label accurately by humans, possibly due to the model's ability to handle class imbalances or identify subtle features that traditional labeled data fails to capture.

Additionally, ensemble filtering proves to be highly effective in significantly reducing Class 0 misclassification, raising another question: Does the use of multiple models in the ensemble provide a more robust signal? Or is there something inherent in the true-labeled Phase 2 data that makes pseudo-labeling a better approach?

These results suggest that further investigation is needed to understand why pseudo-labeling outperforms true-labeled data. Future research could focus on investigating the role of label noise, the complexity of the Phase 2 data, and how pseudo-labeling might capture more subtle features in financial sentiment that are difficult for traditional models to learn.

Conclusion

Pseudo-labeling improved financial sentiment classification across all scenarios. Unfiltered pseudo-labels achieved the highest accuracy (80.6%), while ensemble filtering reduced Class 0 (Negative) misclassification to 30.05%, a 50% improvement over fully labeled Phase 2 data. The hybrid LSTM-CNN model, though slightly less accurate overall, excelled in rare class detection. These findings highlight trade-offs between accuracy and rare class performance, suggesting future work to refine pseudo-labeling strategies and hybrid models for low-resource financial NLP.

References

[1] D. Araci, “FinBERT: Financial Sentiment Analysis with Pre-trained Language Models,” *arXiv preprint arXiv:1908.10063*, 2019. [Online]. Available: https://arxiv.org/abs/1908.10063.

[2] J. Deveikyte, H. Geman, C. Piccari, and A. Provetti, “A Sentiment Analysis Approach to the Prediction of Market Volatility,” *arXiv preprint arXiv:2012.05906*, 2020. [Online]. Available: https://arxiv.org/abs/2012.05906.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805.

[4] X. Du, F. Xing, R. Mao, and E. Cambria, “Financial Sentiment Analysis: Techniques and Applications,” *ACM Computing Surveys*, vol. 56, no. 9, pp. 220:1–220:42, 2024. [Online]. Available: https://doi.org/10.1145/3649451.

[5] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[6] 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. 131648–131663, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9142175.

[7] I. K. Nti, A. F. Adekoya, and B. A. Weyori, “A Systematic Review of Fundamental and Technical Analysis of Stock Market Predictions,” *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, 2020. [Online]. Available: https://doi.org/10.1007/s10462-019-09754-z.

[8] R. Steinert and S. Altmann, “Linking Microblogging Sentiments to Stock Price Movement: An Application of GPT-4,” *arXiv preprint arXiv:2308.16771*, 2023. [Online]. Available: https://arxiv.org/abs/2308.16771.

[9] X. Wan, J. Yang, S. Marinov, J.-P. Calliess, S. Zohren, and X. Dong, “Sentiment Correlation in Financial News Networks and Associated Market Movements,” *Scientific Reports*, vol. 11, no. 1, pp. 82338, 2021. [Online]. Available: https://www.nature.com/articles/s41598-021-82338-6.

[10] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-Training with Noisy Student Improves ImageNet Classification,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10687–10698. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Xie_Self-Training_With_Noisy_Student_Improves_ImageNet_Classification_CVPR_2020_paper.pdf.

[11] E. Zhu and J. Yen, “BERTopic-Driven Stock Market Predictions: Unraveling Sentiment Insights,” *arXiv preprint arXiv:2404.02053*, 2024. [Online]. Available: https://arxiv.org/abs/2404.02053.

Appendix

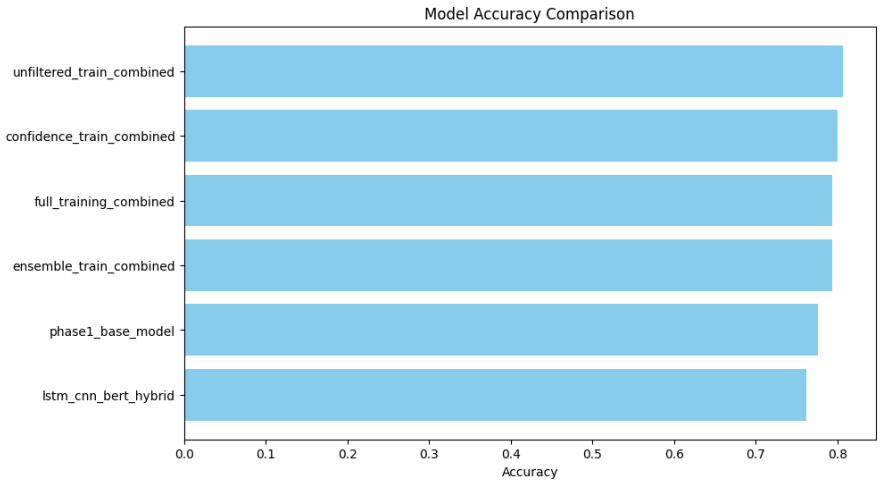


Figure 2. Model Accuracy Comparison

	Scenario	Accuracy	F1-Score (Weighted)
3	unfiltered_train_combined	0.806376	0.796724
4	confidence_train_combined	0.800156	0.793505
2	full_training_combined	0.793157	0.785834
5	ensemble_train_combined	0.793157	0.799418
1	phase1_base_model	0.776050	0.753012
0	lstm_cnn_bert_hybrid	0.762053	0.768425