Related Works

Financial Sentiment Analysis on the Financial PhraseBank Dataset

Selivankin Kirill
Group: BDA-2304

Introduction

This document provides an overview of related works in financial sentiment analysis, with a focus on studies that use the Financial PhraseBank dataset (Malo et al., 2014). The Financial PhraseBank is one of the most widely used benchmarks for sentiment classification in the financial domain, containing approximately 4,840 sentences from Finnish financial news, annotated by 16 domain experts into three classes: positive, neutral, and negative. The dataset provides subsets at different annotator agreement levels (50%, 66%, 75%, 100%), with the 75%-agreement subset (3,453 sentences) being the most commonly used variant.

The works below span from the original dataset paper (2014) through classical machine learning approaches, domain-specific transformer models (FinBERT variants), to recent large language model (LLM) evaluations, reflecting the rapid evolution of NLP methods applied to financial texts.

1. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts

Authors: Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, Pyry Takala

Year: 2014

Venue: Journal of the Association for Information Science and Technology, 65(4), 782-796

Summary: Introduces the Financial PhraseBank dataset and a Linearized Phrase Structure (LPS) model for detecting semantic orientations in financial text. The dataset contains ~4,840 English sentences from Finnish financial news (OMX Helsinki), annotated by 16 domain experts into positive, negative, or neutral sentiment. The LPS model accommodates phrase-structure information and domain-specific language, outperforming word-level polarity lexicons.

Contribution: Creation of the Financial PhraseBank benchmark dataset, which became the standard evaluation resource for financial sentiment analysis. The dataset provides multiple agreement-level subsets (50%, 66%, 75%, 100%).

2. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

Authors: Dogu Araci

Year: 2019

Venue: arXiv preprint; AAAI 2020 Workshop

Summary: Further pre-trained BERT on a financial corpus (Reuters TRC2) and fine-tuned for sentiment

classification on Financial PhraseBank. Achieved ~86% accuracy on the 75%-agreement subset (15 percentage points above previous state-of-the-art) and ~97% on the all-agree subset. Error analysis showed 73% of misclassifications were between positive and neutral labels.

Contribution: First application of BERT to the financial domain. The ProsusAI/finbert model on HuggingFace became the most widely used off-the-shelf financial sentiment classifier.

3. FinBERT: A Pretrained Language Model for Financial Communications

Authors: Yi Yang, Mark Christopher Siy Uy, Allen Huang

Year: 2020

Venue: arXiv preprint; Contemporary Accounting Research (2023)

Summary: Pre-trained BERT on a 4.9B-token financial corpus comprising corporate filings (10-K, 10-Q), analyst reports, and earnings call transcripts. Fine-tuned on 10,000 manually annotated sentences. Achieved 88.2% accuracy, compared to 62.1% for Loughran-McDonald dictionary and 85.0% for vanilla BERT.

Contribution: A distinct FinBERT variant focusing on accounting and corporate disclosure texts. Demonstrated that domain-specific pre-training on corporate communications improves over general BERT, especially for negative sentiment detection (89.7% accuracy).

4. FinancialBERT - A Pretrained Language Model for Financial Text Mining

Authors: Ahmed Rachid Hazourli

Year: 2022

Venue: ResearchGate preprint

Summary: Pre-trained BERT on four large-scale financial corpora: TRC2 financial subset, Bloomberg Financial News, EDGAR Corporate Reports, and Earnings Call Transcripts. Reported accuracy of 0.99 and F1 of 0.98 on the PhraseBank sentiment task (50%-agreement subset), outperforming vanilla BERT and ProsusAI FinBERT.

Contribution: Showed that broader and more diverse financial pre-training corpora can push performance further. Note: the 99% result uses the 50%-agreement subset and should be compared cautiously with results on other agreement levels.

5. Financial Sentiment Analysis: Classic Methods vs. Deep Learning Models

Authors: Aikaterini Karanikola, Gregory Davrazos, Charalampos M. Liapis, Sotiris Kotsiantis

Year: 2023

Venue: Intelligent Decision Technologies (IOS Press)

Summary: Comprehensive comparison of traditional ML methods (SVM, Logistic Regression, Random Forest with TF-IDF) against deep learning and transformer models on a merged FiQA + Financial PhraseBank dataset. SVM with TF-IDF achieved up to ~90.76% accuracy; RoBERTa showed ~7% improvement over Logistic Regression.

Contribution: Systematic empirical comparison establishing quantitative gaps between traditional ML and transformer approaches. Confirmed that SVM+TF-IDF provides a strong baseline (~88-91%), but transformers offer meaningful improvements.

6. Financial Sentiment Analysis and Classification: A Comparative Study of Fine-Tuned Deep Learning Models

Authors: Applied AI Research Lab

Year: 2025

Venue: MDPI Finances, 13(2), 75

Summary: Compared fine-tuned GPT-4o, GPT-4o-mini, BERT, and FinBERT on the combined FiQA + Financial PhraseBank dataset. Used Bayesian optimization across 100 trials for hyperparameter tuning. FinBERT achieved 92% accuracy, significantly outperforming SVM (82%) and Logistic Regression (79%).

Contribution: First head-to-head comparison of GPT-4 family models against fine-tuned FinBERT on financial sentiment. Showed that domain-specific fine-tuned models remain competitive against much larger general-purpose LLMs.

7. Pre-trained Large Language Models for Financial Sentiment Analysis

Authors: Wei Luo, Dihong Gong et al.

Year: 2024

Venue: arXiv preprint (2401.05215)

Summary: Fine-tuned LLaMA2-7B using few-shot learning, further pre-training, and supervised fine-tuning on Financial PhraseBank. Claimed state-of-the-art results by adapting the autoregressive LLM paradigm for financial sentiment classification.

Contribution: Demonstrated that decoder-only large language models (LLaMA family) can be effectively adapted for financial sentiment classification, challenging the dominance of encoder-based models like FinBERT.

8. Reasoning or Overthinking: Evaluating LLMs on Financial Sentiment Analysis

Authors: Multiple authors

Year: 2025

Venue: arXiv preprint (2506.04574)

Summary: Evaluated GPT-4o, GPT-4.1, and o3-mini in zero-shot settings on Financial PhraseBank under different prompting paradigms simulating System 1 (fast/intuitive) vs. System 2 (slow/deliberate) thinking. Found that chain-of-thought reasoning does NOT improve performance; GPT-4o without CoT was most accurate.

Contribution: Counter-intuitive finding that elaborate reasoning strategies (CoT) can hurt financial sentiment classification. Suggests that financial sentiment is a pattern-recognition task where overthinking leads to worse results.

9. FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading

Authors: Multiple authors

Year: 2024

Venue: ICAIF '24 (ACM International Conference on AI in Finance)

Summary: Built a finance-specific LLM framework on LLaMA-2 7B as a generator-discriminator scheme that both classifies sentiment valence and quantifies its strength. Training data combined Financial PhraseBank, FiQA, Twitter Financial News, and GPT-labeled data (34,180 total samples).

Contribution: Extended financial sentiment analysis beyond classification into sentiment strength quantification. Demonstrated practical downstream application of sentiment scores in algorithmic trading strategies.

10. Financial Sentiment Analysis: Techniques and Applications

Authors: Kelvin Du, Frank Xing, Rui Mao, Erik Cambria

Year: 2024

Venue: ACM Computing Surveys, Volume 56, Issue 9, Article 220, pp. 1-42

Summary: Comprehensive survey covering the evolution of financial sentiment analysis from lexicon-based methods through conventional ML (SVM, Logistic Regression) to deep learning and transformer models. Reviews benchmark datasets (PhraseBank, SemEval 2017 Task 5, FiQA Task 1), learning approaches, pre-trained language models, and evaluation methods.

Contribution: The most comprehensive survey bridging FSA techniques and their financial market applications. Covers connections between NLP methodology and financial theory across computer science, information systems, and finance disciplines.

Summary Comparison

#	Paper	Year	Approach	Best Accuracy (PhraseBank)
1	Malo et al.	2014	LPS model + lexicons	Baseline
2	FinBERT (ProsusAI)	2019	BERT + fin. pre-training	~86% (75%-agree)
3	FinBERT (Yang et al.)	2020	BERT + corporate text	88.2% (analyst reports)
4	FinancialBERT	2022	BERT + multi-corpus	~99% (50%-agree, caution)
5	Karanikola et al.	2023	SVM/LR vs. transformers	SVM ~91%, RoBERTa higher
6	MDPI Comp. Study	2025	GPT-4o vs. FinBERT	FinBERT ~92%
7	Luo et al. (LLaMA2)	2024	LLaMA2-7B fine-tuned	Claimed SOTA
8	Reasoning/Overthinking	2025	GPT-4o/4.1 zero-shot	GPT-4o (no CoT) best
9	FinLlama	2024	LLaMA-2 7B + multi-data	Trading-oriented
10	Du et al. Survey	2024	Comprehensive survey	N/A

Note: Direct accuracy comparison across papers is difficult because they use different agreement-level subsets (50%, 66%, 75%, 100%), different train/test splits, and different evaluation protocols. Results should be interpreted within the context of each study's experimental setup.

Our Results in Context

In the context of related works, our project achieves competitive results on the Financial PhraseBank (75%-agreement subset, 518-sample held-out test set with stratified 70/15/15 split):

Model	Accuracy	F1 (macro)	Notes
FinBERT (fine-tuned)	95.56%	0.950	Best overall model
FinBERT (base, no FT)	94.79%	0.935	Pre-trained head only
XLM-RoBERTa (fine-tuned)	91.70%	0.895	Multilingual, 100+ lang
RoBERTa (fine-tuned)	91.12%	0.909	General-purpose baseline
Qwen2.5-3B (zero-shot)	81.85%	0.781	Local LLM via Ollama
RoBERTa (zero-shot NLI)	39.96%	0.415	No financial knowledge

Key findings relative to the literature:

- Our FinBERT fine-tuned accuracy (95.56%) significantly exceeds the original FinBERT paper (~86% on 75%-agree, Araci 2019). This improvement comes from data augmentation (training set balanced from 2,417 to 4,506 samples), class-weighted loss, and careful hyperparameter tuning.
- Consistent with Karanikola et al. (2023), we confirm that transformer models outperform traditional ML baselines. Our RoBERTa fine-tuned (91.12%) aligns with reported SVM+TF-IDF performance (~91%), while FinBERT provides an additional ~4.4% improvement.
- Our LLM evaluation (Qwen2.5-3B zero-shot at 81.85%) supports the finding from the MDPI 2025 study that domain-specific fine-tuned models remain competitive against larger general-purpose LLMs. The gap is especially pronounced on the positive class (LLM F1: 0.589 vs FinBERT: 0.930).

- Consistent with the 'Reasoning or Overthinking' paper (2025), financial sentiment appears to be a pattern-recognition task where domain-specific knowledge (from pre-training) is more valuable than model size or reasoning capabilities.
- Our cross-lingual results with XLM-RoBERTa (80% on Spanish with zero-shot transfer) demonstrate the practical potential of multilingual models for financial NLP in non-English markets, complementing the primarily English-focused related works.