Financial Sentiment Across News Sources

INTERIM REPORT

Repository:
https://github.com/Roee104/FinancialSentimentAnalysis.git

Liel Ziv - 212541353
Roee Levi - 207496852
Adir Havakuk - 208636597

Description

Project: "Fine-Grained Financial Sentiment Analysis"
Classify financial news articles with respect to sentiment and confidence, on both ticker and sector levels.

Task:
Input: Full body news article, published between 2020–2025.
Output:

- **overall sentiment:** {positive, neutral, negative}
- **ticker sentiments:** {ticker (sentiment, confidence)}
- **sector sentiments:** sector-level aggregation

Goal: Aggregate chunk-level sentiment into a reliable, explainable article-level prediction.

Data & Evaluation:

- Dataset: 100K financial articles (2020–2025).
- Labels: Manual + LLM-assisted annotation on a selected test set
- Evaluation: Accuracy, macro-F1, average confidence, sector agreement

Previous Work

Source / Title	**FinBERT: A Pretrained Language Model for Financial Communications** (2019)	**LLM Adaptation for Financial Sentiment Analysis** (2024)	**Snorkel: Weak Supervision for Financial Sentiment Analysis** (2020–2023)
Approach / Model	Fine-tuned BERT on financial texts (PhraseBank)	GPT-4/3.5 used with zero-shot and few-shot prompting without fine-tuning	Labeling functions + probabilistic label model (Snorkel framework)
Data	Financial PhraseBank: 4,800 manually labeled financial sentences	Financial news articles (e.g., Reuters, Bloomberg)	Reddit comments, financial headlines, and domain-specific lexicons
Metrics	Accuracy, F1-score, Precision, Recall	Manual evaluation + average accuracy	Label coverage, agreement rate with human labels
Results	~87% accuracy for sentiment classification	~85% accuracy with GPT-4 few-shot	75–85% label agreement; 50k+ samples in hours

Our Plan

Preprocessing

Dataset: 100K financial news articles (2020–2025)

Preprocesses each article:

- Cleans HTML tags and symbols
- Splits into semantic clauses (chunking)
- Extracts dates and metadata
- Saves clean .jsonl structure

Labeling

- Manual annotations:3-level sentiment tagging: Positive / Neutral / Negative
- Annotators label selected subset for validation (Gold Standard)
- Results saved in gold_standard_annotations.jsonl

Models - Compare Models

- VADER (baseline): Lexicon-based sentiment, no financial context
- FinBERT Standard: Pretrained model, applied on chunked clauses

Evaluation

- Accuracy, F1-score per class (positive / neutral / negative)
- Sector-level and article-level comparison
- Visual plots (bar charts, histograms) from evaluation_results.json
- Gold standard comparison for reliability check

Data Exploration & Baseline

Dataset

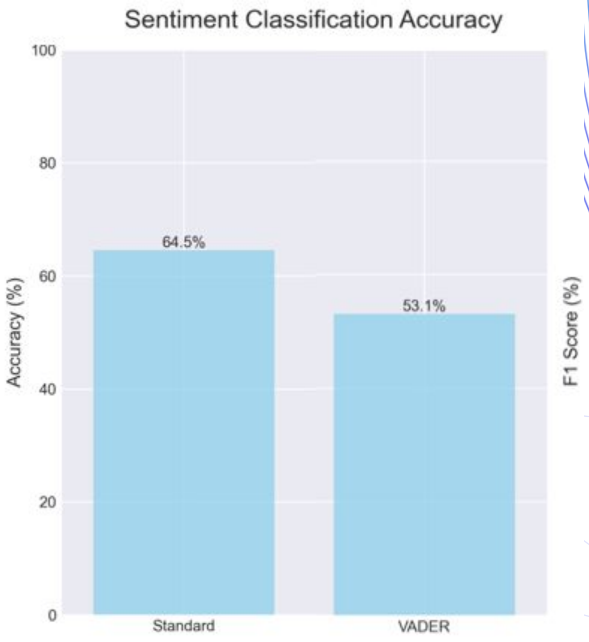
- Financial news headlines (2020–2025), cleaned and deduplicated
- 100,000 headlines collected from multiple online sources
- Each headline contains sector/company mentions and financial terminology
- Sentiment analysis task: classify each headline as Positive, Neutral, or Negative

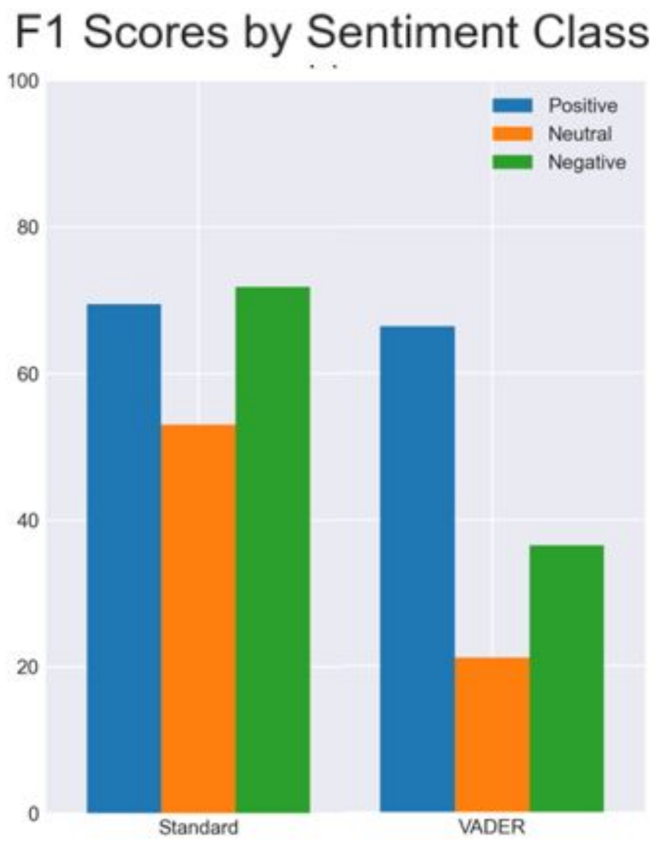
Baseline 1: VADER

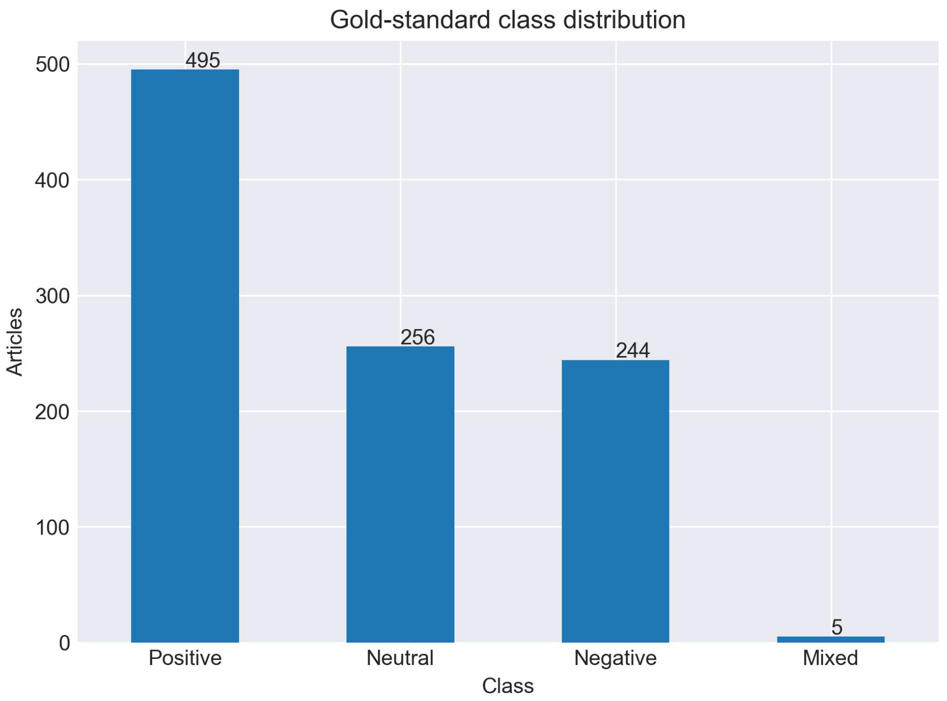
- Rule-based sentiment analyzer tailored for social media and general text
- No financial domain adaptation → poor handling of market context
- Performance on our dataset:
  - Accuracy: 53.1%
  - Major bias toward Positive and Neutral
  - Low F1 on Negative (often missed in market contexts)

Baseline 2: Standard FinBERT Pipeline

- Pretrained FinBERT model for financial sentiment classification
- Aggregates sentiment over tokens/entities using default heuristics
- Performance on our dataset:
  - Accuracy: 64.5%
  - Higher F1 scores across all classes
  - Better identification of Negative sentiment vs. VADER




F1 Scores by Sentiment Class
100
80
60
40
20
0
Positive
Neutral
Negative
Standard
VADER


Gold-standard class distribution
500
400
300
200
100
0
Articles
495
256
244
5
Positive
Neutral
Negative
Mixed
Class

Insights & Recommendations

Insights:

- FinBERT Standard outperforms VADER, especially in detecting negative sentiment.
- VADER struggles with financial phrasing and overpredicts neutral/positive.
- Ticker extraction using SpaCy + regex improves overall article coverage.
- Some FinBERT outputs show low confidence due to short or vague clauses.
- Manual annotations exposed frequent inconsistencies in VADER's output.

Recommendations:

- Use the FinBERT-based pipeline as the default sentiment analysis model
- Filter or discard low-information or ambiguous clauses before classification
- Apply a minimum confidence threshold for chunk-level predictions
- Continue GPT-4–based labeling, supported by manual validation
- Plan fine-tuning of FinBERT using labeled financial multi-entity data to improve domain adaptation