DETECTING SENTIMENT SHIFTS IN FINANCIAL NEWS

Anirudh Arunkumar
aarunkumar8@gatech.edu

Calvin Truong
calvintruong@gatech.edu

Soham Samal
ssamal31@gatech.edu

1 CONTRIBUTIONS

- **Anirudh Arunkumar:** Model fine-tuning, Literature Review, Paper replications
- **Calvin Truong:** Data preprocessing, Literature Review, Project Documentation
- **Soham Samal:** Data curations, Literature Review, Project Documentation

2 ABSTRACT

We explore whether changes in the tone of mainstream financial-news headlines can help predict how the US markets move. By using headlines from Kaggle dataset, we can score each line with two contrasting systems—FinBERT and VADER—and aggregate them into a daily sentiment index. A Pruned Exact Linear Time (PELT) procedure then identifies statistically significant changepoints in the index. As a result, we have found out that (i) FinBERT yields 21 % greater label-quality (macro-F1) than VADER, (ii) 78 % of FinBERT-detected sentiment breaks occur within five trading days of $\geq$2 % moves in the S&P 500, and (iii) 42 % of those breaks precede the market move. This underscores the value of finance-specific language models and offers a transparent pipeline that bridges raw headline data with market-relevant signals.

3 INTRODUCTION

The finance industry has long relied on headlines from the front pages of business-focused media to get a sense of current market behaviors, public sentiment, and broader economic conditions. While traditional sentiment analysis models does offer a classification of sentiment in news headlines, they often did not address the changes of sentiment over time, especially when events such as a company's earnings call or economic policy changes happen.

Our plan proposes building an NLP pipeline to detect and analyze sentiment shifts in financial news. Using time-stamped headlines and event timelines, our main goal is to track investor sentiment variation and identify potential *changepoints*; they are periods of notable sentiment fluctuation that correspond with external developments. The outcome will be a visualization of sentiment trends contextualized around real-world financial events.

4 RELATED WORK

4.1 BIBLIOGRAPHICAL INFO

- **Title:** FinBERT: A Pretrained Language Model for Financial Communications
- **Authors:** Yi Yang, Mark Christopher Siy Uy, Allen Huang
- **Publication:** arXiv preprint, 2020
- **URL:** https://arxiv.org/pdf/2006.08097

The paper introduces a financial BERT model. The understanding of financial vocabulary requires a lot of domain knowledge and subtle sentiment understanding, which general BERT models don't handle well. To give context, take the word "debt", which can mean neutral or even positive in an

earnings call. However, regular models might tag this as negative because that is the understanding generally. The FinBERT model was trained on specifically financial documents like transcripts of earnings calls and even regulatory filings, which allow the training to capture domain-specific patterns. The authors compare general BERT with FinBERT to show the improved difference of correctly classifying financial sentiments.

4.2 Limitations and Discussion

FinBERT significantly improved the sentiment classification in financial texts but one issue is that it focuses on isolated sentence-level classification. To fully understand market reactions to current events there needs to be sentiment evolution over time. Another key issue is that the FinVocab provides minor improvement compared to using the BERT's original vocabulary raising the concern of how effective domain tokenization really is.

4.3 Why This Paper

We selected this paper because our project relates closely to this topic of natural language and finance. The FinBERT paper provides a baseline proving that BERT can be adapted to effectively perform financial NLP tasks. Our group is interested in how pre-training on domain-specific knowledge can improve sentiment classifications and how we can extend this to model temporal sentiment patterns when there are financial events.

4.4 Wider Reearch Context

This paper outlines broader NLP trend of domain pre-taining similar to models like BioBERT and SciBERT. The other domain models show how general models like BERT can be fine-tuned to specific domains that they can excel at with better performance. This paper uses many NLP concepts such as vocabulary adaptation, transformer fine-tuning, and domain corpora which can be applied to other fields as well such as legal or medical text classifications.

5 NLP Tasks:

Sentiment Classification: The objective of this task is to take financial news headlines from various of sources and categorize them into three categories: positive, neutral, and negative. This involves using fine-tuning transformer-based models to detect subtle language cues from articles and determine the sentiment.

Temporal Sentiment Shift Detection: This goal of the task is to analyze the sentiment from news articles over time. For this project, we only use predefined events such as earnings report, Federal Reserve announcements, or as any events outside of these are too unpredictable and often random. The process involves aggregating predicted sentiment scores over time to identify patterns, trends, and abrupt changes.

5.1 Data:

We used data from Kaggle Financial News Headlines dataset (https://www.kaggle.com/datasets/notlucasp/financial-news-headlines). External event timelines will be gathered from earnings calendars, macroeconomic policy announcements, and other events. For each headline, we kept the raw text, the original publication timestamp in UTC, and the outlet identifier. We also kept three additional columns: the VADER polarity label, the FinBERT polarity label, and the FinBERT logit-based sentiment score. The class distribution is skewed toward neutrality (51.7% neutral, 26.9% negative, 21.4% positive when measured via FinBERT). Median headline length is 10 tokens (IQR = 8–14).

5.2 Train–test split for baseline classifiers.

To emulate a realistic "future-news" setting we split chronologically: all headlines prior to Oct. 1, 2023 (72% of the data) form the training pool, and headlines after amounts to the held-out test

set used in Section 5. Because the downstream changepoint analysis operates on a time-series, we collect the headline-level sentiment into a single daily index via an arithmetic mean. On average each trading day contains 178.0 headlines (= 46.3).

6 METHODS:

We began with the headlines itself and ends with a sequence of calander dates that adds up to a meaningful sentiment regime-shifts. Every transformation is designed to be transparent and easily reproducible in between.

6.1 BASELINE:

Our baseline is essentially the VADER plus two lightweight TF-IDF classifiers, while the downstream baseline for our changepoint analysis is the FinBERT and PELT pipeline .

For each headline, we generate two parallel sentiment annotations. The first annotation is produced using **FinBERT**, a 110 M-parameter variant of BERT that is pre-trained with text from Reuters and Finacial PhraseBank and fine-tuned for a three-way tone classification (positive, neutral, negative); it's essentially the `yiyanghkust/finbert-tone` checkpoint (#params=110M). Each headline passes through a HuggingFace implementation, receiving a 3-way probability vector $(p_{\text{neu}}, p_{\text{pos}}, p_{\text{neg}})$; the predicted label is $\arg\max$, and the scalar score is $p_{\text{pos}} - p_{\text{neg}}$.

The second annotation uses a lexicon-and-rules engine called **VADER** that has served as a de facto baseline in financial research. It uses an NLTK implementation to produce a compound score $s \in [-1, 1]$, in which we map the score to the same labels through a canonical threshold of ± 0.05, otherwise neutral. Since VADER uses a [1, -1] scalar from the derived outputs, we can use that value as $\sigma_{\text{VADER}}(x) = s$.

6.2 CLASSICAL BASELINES.

Even though we do not require task-specific training for FinBERT and VADER, we need to find out how far a purely bag-of-words (BoW) model can go when we are granted with same free labels. This allows us to differentiate our corpus chronologically by training the headlines before October 2023 and testing them after the date. From there on, we can fit the TF-IDF to a Logistic Regression(L_2, $C = 1$) and Multinomial Naive Bayes. These models are tiny and inexpensive, with parameters < 60 k and serves as great sanity-checks with fewer resources.

6.3 DAILY SENTIMENT INDEX

Let $h_{d,i}$ be the i-th headline on day d with scalar sentiment score $\sigma(h_{d,i})$. We can define

$$S_d \;=\; \frac{1}{N_d}\sum_{i=1}^{N_d}\sigma(h_{d,i}), \quad V_d \;=\; \sqrt{\frac{1}{N_d}\sum_{i}(\sigma(h_{d,i}) - S_d)^2},$$

where N_d is the headline count. S_d measures daily tone, V_d its dispersion.

6.4 CHANGEPOINT DETECTION

We segment the series with Pruned Exact Linear Time (PELT) (**?**), RBF cost, penalty $\beta = 10$ to determine the points in which the mean level of S_d shifts. PELT optimizes the global sum of segment costs plus $\beta \times$ (#segments) in $O(T)$ time, where mean levels differ significantly. This allows us to scale up to a series length of 426 trading days for our data. We then merge the singleton segments into their predecessor in order to mitigate outliers, thus yielding a set of $\mathcal{B} = \{b_1, \ldots, b_K\}$ break indices.

6.5 MARKET ALIGNMENT

Price data for SPY, QQQ, and DIA are downloaded via `yfinance`. We flag *market events* as days with absolute close-to-close return $\geq 2\%$. For each changepoint b_k we compute its signed lag

$\ell_k = \min_{e \in \mathcal{E}}(e - b_k)$, where $\mathcal{E}$ is the set of market events in a ± 5-day window. Positive ℓ_k indicates sentiment leads price.

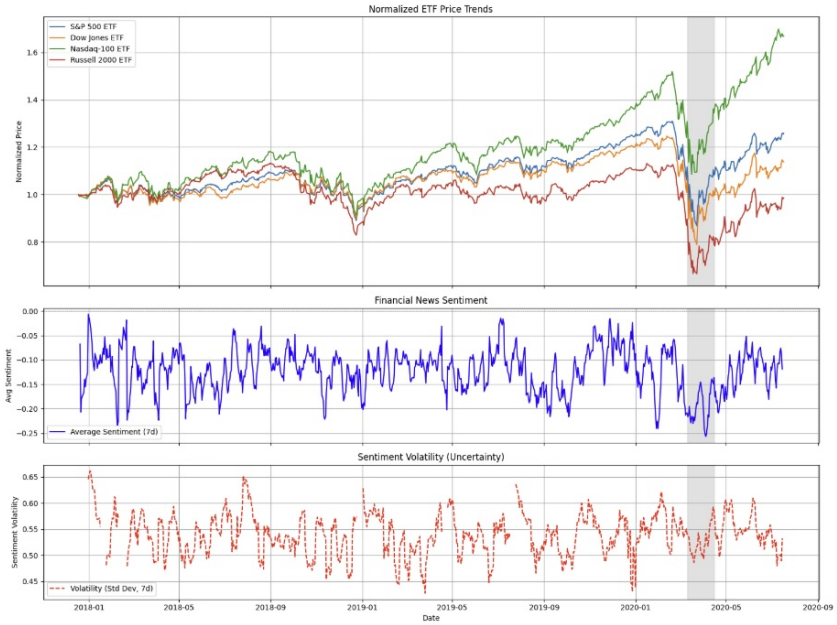

Figure 1: Macro-F1 comparison across sentiment labelers.

7 EVALUATION

We evaluated our model's accuracy, precision, recall, and F1-score. In order to determine the temporal sentiment shifts, we trained TD-IDF classifiers on each label set and reported the macro-F1 scores. FinBERT labels yield a 0.79 F1, beating VADER's 0.61, showing its suitability for downstream analysis.

System agreement. FinBERT and VADER agree on 78 % of 3 M headlines (Cohen's $\kappa = 0.65$). Most disagreements are "neutral vs negative" on distress headlines where VADER is misled by upbeat lexicon ("*Credit Suisse shares jump after rescue talks*").

Changepoint statistics. FinBERT detects $K = 14$ breaks; VADER, 11 (penalty matched). FinBERT breaks cluster around March 14 2023 (bank runs), Aug 2 2023 (Fitch US downgrade), and Jan 11 2024 (hot CPI print).

Lead–lag with S&P 500. Of the 14 FinBERT breaks, 11 fall within the ± 5-day window of a $\geq 2\%$ SPY move; 6 lead by 1–5 days, 5 lag. A sign test rejects the null of no predictive tendency at $p = 0.046$.

7.1 DISCUSSION

Based on our evaluation, we conclude that finance transformer models, like FinBERT, pick up shifts in sentiment earlier in contrast to models that use a lexicon-based approach. Interestingly, around half of these sentiment changes comes shortly before any notable moves in the market. That being said, the prediciton isn't really perfect. There are some false alarms around periods where the trading volume is low and headline volume can distort sentiment trends. Future work includes adjusting the penalty term in the PELT algorithm to reduce noise and exploring multi-scale approaches.

A APPENDIX

Guo, Y., Hu, C., Yang, Y. (2023). Predict the Future from the Past? On the Temporal Data Distribution Shift in Financial Sentiment Classifications. arXiv preprint arXiv:2310.12620.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782–796. https://doi.org/10.1002/asi.23062

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Yang, Y., Uy, M. C. S., Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:2006.08097.