

Market Mood & Moves: Sentiment-Driven Stock Prediction

Mentors: Meet and Sarthak

Sanjiban Paul (24B2154)

WiDS 2025-26

1 Behavioral Finance and Market Psychology

Financial markets are ever-changing and adaptive systems. The traditional approach assumes the presence of rational agents and efficient information flow, but empirical results prove that the behavior of investors follows patterns that are not entirely rational. This indicates that human behavior, overconfidence, loss aversion, as well as speculative bubbles introduce conditions of temporary inefficiency that could be capitalized on through the implementation of sentiment analysis tools; which is the objective of this project.

1.1 Quantifying Market Sentiment

Market sentiment is inferred from text data, which produces aggregate numbers as market sentiment measures. Let S_t denote the market sentiment on day t , and X_t denote traditional financial covariates such as volume or volatility. The expected return can be modeled as:

$$R_{t+1} = \alpha + \beta S_t + \gamma X_t + \varepsilon_t$$

where ε_t represents noise not captured by the model. Proper temporal alignment ensures that only information available before the market close of day t is used to predict R_{t+1} , avoiding look-ahead bias.

1.2 Behavioral Effects on Asset Prices

Investor psychology affects trading decisions:

- **Herding:** Investors follow the majority, amplifying trends.
- **Overreaction:** Prices temporarily overshoot due to news sentiment.
- **Confirmation Bias:** Traders overweight information that supports existing beliefs.
- **Loss Aversion:** Negative news has a larger impact than positive news of equal magnitude.

By quantifying these effects through sentiment analysis, trading systems can forecast short-term deviations from expected values.

2 Textual Data and NLP Techniques

2.1 Text Preprocessing

Financial text is unstructured, noisy, and domain-specific. Text preprocessing transforms text to numerical forms that can be processed by models. The usual preprocessing techniques include:

- **Tokenization:** Splitting sentences into words or subword units.
- **Stop-word removal:** Elimination of frequent words with less semantic meaning.
- **Lemmatization:** Reducing a word to its basic form, e.g., “declining” \rightarrow “decline”.

These steps preserve semantic meaning while reducing dimensionality.

2.2 Lexicon-Based and Model-Based Sentiment

Lexicon-based approaches give scores to words individually. For example, “profit” is positive, “loss” is negative. However, financial terms often carry domain-specific meanings:

- “Liability” is a neutral word in accounting, though negative in general English.
- “Cost” might be neutral in operations context.

Context-aware representations, for instance, transformers, are able to manage this limitation.

2.3 Aggregating Sentiment Scores

Once sentiment is extracted from individual documents, aggregation strategies include:

$$S_t = \frac{1}{N_t} \sum_{i=1}^{N_t} s_{t,i}$$

where N_t is the number of documents on day t , and $s_{t,i}$ is the sentiment score of document i . This produces a daily sentiment indicator aligned with market returns.

3 Word Representations and Embeddings

3.1 Static Embeddings

Traditional embeddings such as Word2Vec map each word w to a fixed vector $\mathbf{v}_w \in \mathbb{R}^d$. This representation is independent of the context, which can cause issues since a word can have multiple meanings: the word “bank”, for example, conflates financial and geographical meanings.

3.2 Contextual Embeddings

Transformer models encode context dynamically:

$$\mathbf{v}_i = f(w_i \mid w_1, w_2, \dots, w_n; \theta)$$

A word’s representation relies on surrounding words, aiding in meaning and context; while also resolving an ambiguity. Subword tokenization (WordPiece) increases the flexibility in the treatment of out-of-vocabulary words by breaking them down, as in the case of ‘Cryptoleverage’.

3.3 Positional Embeddings

Since transformers process tokens in parallel, positional embeddings E_{position} are added to encode sequential order, ensuring that the model distinguishes “Cat chases mouse” from “Mouse chases cat”.

4 Transformer Architecture

4.1 Encoder Block and Self-Attention

Each Transformer layer applies self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q (query), K (key), and V (value) matrices are learned representations of tokens. This mechanism allows the attention of every token to every other one, capturing dependencies across long sequences.

4.2 GELU Activation

The Gaussian Error Linear Unit is used instead of ReLU:

$$\text{GELU}(x) = x\Phi(x)$$

where $\Phi(x)$ is the cumulative distribution function of a standard Gaussian. GELU introduces smooth stochasticity that improves optimization in deep networks.

4.3 Pre-training Objectives

Masked Language Modeling (MLM) Randomly mask 15% of tokens and predict them. Using the 80-10-10 strategy:

- 80% replace with [MASK]
- 10% replace with a random word
- 10% leave unchanged

Next Sentence Prediction (NSP) Predict whether sentence B follows sentence A. This helps model coherence and contextual relationships between sentences.

5 Domain Adaptation and FinBERT

5.1 Domain Shift

Financial text distributions vary greatly from generic English text distributions. Generic models can err in classifying “loss” as “share,” or “share” as “loss.” Domain adaptation aligns model parameters θ with financial data:

$$\theta_{\text{fin}} = \arg \min_{\theta} \mathbb{E}_{x \sim P_{\text{finance}}} [\mathcal{L}_{\text{MLM}}(x; \theta)]$$

5.2 Training Pipeline

1. General Pre-training on Wikipedia (BERT)
2. Further Pre-training on financial corpus (TRC2-Financial)
3. Supervised Fine-tuning on sentiment-labeled datasets (Financial PhraseBank)

It helps in improving accuracy and F1 value significantly compared to word embeddings or lexicon-based solutions.

5.3 Challenges

- **Catastrophic Forgetting:** Learning financial language may override general knowledge.
- **512-Token Limit:** Long documents are segmented into 512-token chunks.
- **Slanted Triangular Learning Rates:** Learning rates are increased and then decreased for efficient fine-tuning.

6 Quantitative Evaluation

6.1 Performance Metrics

- Compound Annual Growth Rate (CAGR)

$$\text{CAGR} = \left(\frac{P_T}{P_0} \right)^{1/T} - 1$$

- Maximum Drawdown (MDD)

$$\text{MDD} = \min_t \frac{P_t - \max_{s \leq t} P_s}{\max_{s \leq t} P_s}$$

- Sharpe Ratio

$$\text{Sharpe} = \frac{\mathbb{E}[R_p - R_f]}{\sqrt{\text{Var}(R_p - R_f)}}$$

6.2 Backtesting and Robust Evaluation

Backtesting using walk-forward analysis and out-of-sample data ensures that strategies are not overfit and generalize to new and unknown market conditions.

7 Conclusion

Using the principles of behavioral finance, text-based sentiment analysis, and transformer-based context embeddings, FinBERT, among other models, offers an empirically sound solution approach in terms of short-term market predictions.

8 References

- GeeksforGeeks website
- Wall Street Quants YT Channel
- <https://www.bavest.co/en/post/market-psychology-and-sentiment>