Financial Sentiment Classification: Comparing Classical Models and FinBERT

Elias Michael
230020425
MSc Data Science
elias.michael@city.ac.uk

1 Problem Statement and Motivation

Financial markets are highly sensitive to sentiment expressed in news headlines, investor commentary, and analyst reports. Events such as corporate earnings announcements, regulatory changes, or technological breakthroughs can lead to immediate fluctuations in asset prices. However, given the vast and growing volume of financial text data, manual sentiment analysis is impractical for analysts, investors, and regulatory agencies who seek real-time insights.

Natural Language Processing (NLP) presents a scalable solution by automating sentiment extraction from financial texts. Financial Sentiment Analysis (FSA), a subfield of NLP, classifies financial communications—including press releases, earnings reports, and social media commentary—into categories such as positive, neutral, or negative. The growing reliance on algorithmic trading and data-driven decision-making underscores the critical role of FSA.

Financial texts, however, introduce domain-specific challenges. Terms such as “liability” or “bullish” may hold context-dependent sentiment meanings that general-purpose NLP models often misinterpret. For instance, phrases like “riding a bull” or “bearish outlook” require contextual financial knowledge. These challenges motivate the use of domain-specific models like FinBERT.

This project addresses the question: *How do traditional machine learning models compare with domain-specific transformer models like FinBERT for financial sentiment classification?* Using the Financial PhraseBank dataset, we build and evaluate sentiment classification pipelines with Logistic Regression, Linear SVM, and FinBERT. Our goal is to assess whether FinBERT’s financial domain adaptation offers measurable improvements in accuracy, especially on context-heavy or ambiguous texts.

2 Research Hypothesis

We hypothesize that: **FinBERT, a domain-specific transformer model pretrained on financial corpora, will significantly outperform traditional machine learning models (Logistic Regression and Linear SVM) in financial sentiment classification.**

This hypothesis is motivated by recent advancements in transformer-based models, which excel at capturing deep contextual representations in text. FinBERT, specifically fine-tuned on financial documents, is expected to recognize financial jargon, numerical references, and implicit sentiment more effectively than feature-based classical models.

Traditional models, while efficient and interpretable, rely on sparse lexical features (e.g., TF-IDF) and may struggle with nuanced expressions or variable sentence structures. However, these models offer valuable baselines and retain practical advantages such as lower inference costs and greater explainability.

Our experiments are designed to critically test this hypothesis across balanced metrics (macro-averaged precision, recall, F1), particularly focusing on minority sentiment classes and edge-case examples.

3 Related Work and Background

The field of Financial Sentiment Analysis (FSA) intersects NLP, finance, and behavioral economics. Early FSA systems used domain-specific lexicons such as (Loughran and McDonald, 2011), which addressed misclassifications caused by general-purpose sentiment tools. (Malo et al., 2014) later introduced the Financial PhraseBank, a benchmark dataset with manually labeled financial sentences, which this study also employs.

Recent work emphasizes the need for contextual understanding in FSA. (Du et al., 2024) highlight the challenges posed by specialized financial vocabulary, sarcasm, and ambiguity. Their survey emphasizes the superiority of domain-adapted models over generic NLP tools for financial tasks.

(Karanikola et al., 2023) provide a comparative study between classical models (Logistic Regression, SVM) and deep learning approaches (BERT, RoBERTa, FinBERT), showing that transformer models, particularly FinBERT, consistently outperform baselines when sufficient domain alignment and preprocessing are applied. However, they also stress that preprocessing decisions and dataset balance critically affect outcomes.

Targeted sentiment analysis is another evolving subfield. (Pan et al., 2023) argue that sentiment often depends on the mentioned entity's context—for example, rising oil prices may benefit energy firms but harm airlines. They advocate for models that consider entity-level polarity.

FinBERT, introduced by (Araci, 2019), builds on BERT (Devlin et al., 2019) but is fine-tuned using financial texts such as SEC filings and earnings reports. FinBERT achieves superior results on financial sentiment datasets due to its ability to learn subtle financial semantics, such as optimistic projections masked by neutral tone.

Our project builds on this foundation by comparing FinBERT with classical baselines under controlled preprocessing and evaluation settings. While FinBERT offers state-of-the-art performance, its computational costs and opaque interpretability remain important trade-offs we explore.

Accomplishments

- **Task 1: Perform exploratory data analysis (EDA)** – *Completed*
  Explored the structure of the Financial PhraseBank dataset using descriptive statistics and visualizations (e.g., class distribution, sentence length boxplots). Identified a strong class imbalance and syntactic differences across sentiment classes, guiding evaluation metric choices.

- **Task 2: Design and test multiple preprocessing configurations** – *Completed*
  Developed a modular pipeline combining basic cleaning, lemmatization, and stopword removal. Four configurations were evaluated for their impact on classical model performance. Lemmatization-only (Variant 3) consistently outperformed others, preserving key domain-specific terms.

- **Task 3: Investigate class imbalance handling using SMOTE** – *Completed and discarded*
  SMOTE was tested alongside preprocessing configurations but led to reduced generalization performance and noisier class boundaries. As a result, it was excluded from the final pipeline in favor of macro-averaged metrics.

- **Task 4: Extract features using multiple vectorization techniques** – *Completed*
  Implemented Bag-of-Words, TF-IDF, and Word2Vec. TF-IDF with unigram and bigram features yielded the best performance for classical models, while Word2Vec underperformed due to the dataset's compact size and brevity.

- **Task 5: Train and evaluate classical ML models (Logistic Regression, Linear SVM)** – *Completed*
  Used 5×5 nested cross-validation and hyperparameter tuning to evaluate classical models across all preprocessing and feature configurations. Linear SVM with TF-IDF and lemmatization achieved the strongest baseline.

- **Task 6: Extract and analyze feature importance from Logistic Regression** – *Completed*
  Interpreted model weights from the TF-IDF features to identify the most predictive terms per sentiment class, enhancing transparency and explainability of classical models.

- **Task 7: Fine-tune the FinBERT transformer model** – *Completed*
  Fine-tuned FinBERT using Hugging Face's Trainer API with minimal preprocessing. Achieved state-of-the-art performance (macro F1 = 0.9697), significantly outperforming classical models on all sentiment classes.

- **Task 8: Evaluate models on held-out stratified test set** – *Completed*
  Applied final models to a 20% test set reserved for generalization assessment. Reported macro-averaged precision, recall, and F1 to account for label imbalance.

- **Task 9: Conduct qualitative error analysis** – *Completed*
  Reviewed misclassified samples from each model. Found that FinBERT handled subtle semantic cues better, while classical models

tended to misclassify implicit or comparative sentiment.

- **Task 10: Compile final report and visualizations** – *Completed*
  Structured the report according to academic standards, including detailed tables, confusion matrices, F1-score breakdowns, and illustrative error examples for transparency and reproducibility.

4 Approach and Methodology

This study adopts a two-pronged methodological framework to evaluate sentiment classification in financial headlines: (1) classical machine learning models trained on engineered features, and (2) a fine-tuned transformer-based model (FinBERT) designed for financial text understanding. The goal was to systematically compare their effectiveness under a unified and controlled pipeline.

4.1 Preprocessing Strategy

We implemented a modular NLP preprocessing pipeline using Python's `re`, `string`, and `spaCy` libraries. Four preprocessing variants were evaluated:

- **V1:** Basic cleaning (lowercasing, punctuation and digit removal)
- **V2:** V1 + Stopword Removal (using a financial-augmented spaCy list)
- **V3:** V1 + Lemmatization
- **V4:** V1 + Stopword Removal + Lemmatization

Variant V3 (lemmatization only) consistently yielded the best performance for classical models by preserving syntactic cues while reducing feature sparsity. In contrast, V2 and V4 showed performance drops, suggesting that removing function words may discard important comparative and negation cues common in financial discourse.

For FinBERT, minimal preprocessing was applied to maintain compatibility with its pretrained tokenizer: excess whitespace and control characters were cleaned, but stopwords and punctuation were retained.

4.2 Feature Extraction

To convert text into machine-readable format, three widely-used vectorizers were tested:

- **Bag-of-Words (BoW):** Used unigrams and bigrams (`ngram_range=(1,2)`) via `CountVectorizer`. While BoW offered interpretability, its disregard for word frequency and semantic weighting limited its effectiveness.
- **TF-IDF:** Implemented using `TfidfVectorizer` with sublinear term frequency scaling, smoothed inverse document frequency, and L2 normalization. This yielded the best results across all classical models, especially with V3 preprocessing.
- **Word2Vec:** Utilized pre-trained word embeddings from Gensim's Google News vectors. Sentence representations were constructed by averaging token vectors. Despite its semantic richness, Word2Vec underperformed on short sentences and small datasets, consistent with prior findings.

4.3 Modeling and Training

We evaluated three model families:

- **Logistic Regression (LR):** A simple and interpretable linear classifier. Used L2 regularization and trained on BoW, TF-IDF, and Word2Vec representations.
- **Linear Support Vector Machine (SVM):** Selected for its robustness in high-dimensional spaces and strong performance on sparse data. Hyperparameters were tuned using 5×5 nested cross-validation.
- **FinBERT:** The `ProsusAI/finbert` model from Hugging Face was fine-tuned using the Trainer API. Training involved a 3-epoch schedule with early stopping and weighted loss to mitigate class imbalance. FinBERT's attention mechanism and financial-domain pretraining enabled it to capture sentiment nuances missed by classical models.

4.4 Evaluation Protocol

All models were evaluated on a stratified 80/20 train-test split using macro-averaged precision, recall, and F1-score, ensuring fair assessment across the imbalanced sentiment classes. Nested cross-validation was used during model selection for classical models to prevent overfitting.

4.5 Libraries and Tools

The implementation leveraged:

- **Preprocessing:** `spaCy`, `re`, `nltk`
- **Classical models:** `scikit-learn`, `gensim`
- **FinBERT:** `transformers`, `datasets`, `Hugging Face Trainer`
- **Visualization:** `matplotlib`, `seaborn`, `missingno`

4.6 Challenges and Adjustments

- **SMOTE Oversampling:** Initially explored for class balancing, but discarded after degraded performance on minority sentiment classes.
- **Stopword removal:** Although commonly used, its exclusion harmed syntactic understanding in financial text, especially with TF-IDF.
- **FinBERT tuning:** Required minimal preprocessing and extensive GPU memory, limiting batch size and training iterations.

Overall, the methodology ensured consistency in preprocessing and evaluation across all models while respecting architectural requirements. This design enabled a fair and reproducible comparison between interpretable classical models and state-of-the-art transformer-based approaches.

5 Dataset and Initial Analysis

This study uses the Financial PhraseBank dataset (Malo et al., 2014), a widely adopted benchmark in financial sentiment analysis. We specifically selected the "All Agree" subset, where all annotators concurred on the sentiment label, ensuring high label reliability and reducing ambiguity common in subjective text corpora.

5.1 Dataset Overview

After deduplication, the final dataset consists of 2,259 labeled sentence-sentiment pairs. Each sentence is a brief excerpt from financial news articles, earnings releases, or press statements. The labels fall into three classes: *positive*, *neutral*, or *negative*, reflecting sentiment polarity.

The dataset is highly imbalanced, with the neutral class dominating the label distribution. A chi-squared goodness-of-fit test confirmed a statistically significant deviation from a uniform distribution (χ^2 = 845.52, p $<$ 0.0001), motivating the use of macro-averaged evaluation metrics in all experiments.

Class	**Count**	**Proportion**
Positive	570	25.2%
Neutral	1389	61.5%
Negative	300	13.3%
Total	2259	100%

Table 1: Label distribution in the dataset

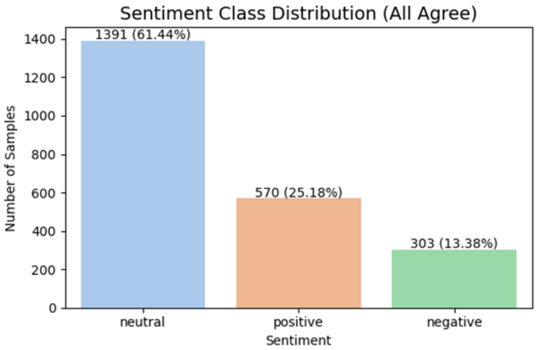

Figure 1: Sentiment class distribution in the Financial PhraseBank (All Agree subset). The dataset is heavily imbalanced with over 60% of instances labeled as neutral.

5.2 Sentence Characteristics

Sentences range from 2 to 81 words, with a mean of 22 words and approximately 122 characters. This brevity reflects the "headline-style" communication typical in financial reporting.

A boxplot analysis of sentence lengths by sentiment class (Figure 2) revealed that:

- Neutral sentences tend to be shorter and more factual.
- Positive and negative sentences are typically more descriptive or evaluative, with stronger sentiment expressions.

5.3 Lexical and Syntactic Properties

Unigram and bigram frequency analysis revealed frequent financial terms such as "net sales," "operating profit," and "quarter earnings." These lexical

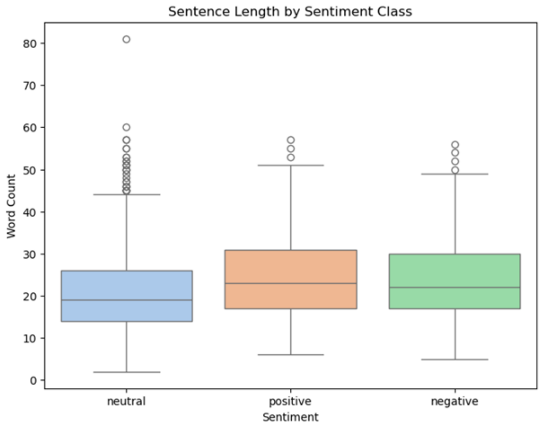

Figure 2: Boxplot of sentence lengths (in word count) by sentiment class. Neutral sentences are typically shorter and more factual; positive and negative sentences tend to be more descriptive.

cues are crucial for sentiment interpretation in financial contexts.

Part-of-speech tagging with spaCy showed a dominance of:

- **Nouns**: company names, financial metrics, sectors
- **Verbs**: directional indicators (e.g., “increased,” “declined,” “forecasted”)
- **Numerical tokens**: expressing percentage change or performance metrics

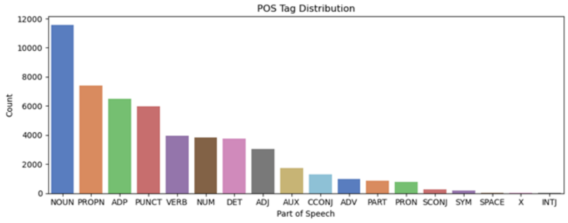

Figure 3: Distribution of part-of-speech tags across the dataset. Financial texts are heavily noun- and verb-driven, with notable use of numerals and proper nouns.

5.4 Data Integrity

No missing values were found in either the sentence or label columns. All entries were UTF-8 encoded, and labels were manually reviewed for consistency. This clean structure enabled reproducible modeling without the need for aggressive filtering or imputation.

6 Dataset Preprocessing

To prepare the text for model training, we designed a preprocessing pipeline that preserved domain-specific semantics while reducing noise and sparsity. The preprocessing strategy was tailored differently for classical machine learning models and transformer-based models like FinBERT.

6.1 Preprocessing for Classical Models

We implemented four preprocessing configurations to test how different levels of linguistic simplification affected classical model performance:

- **V1 – Basic Cleaning:** Lowercased text, removed punctuation and digits using regular expressions.
- **V2 – Basic Cleaning + Stopword Removal:** Removed common stopwords using an augmented version of spaCy’s default list with financial terms retained.
- **V3 – Basic Cleaning + Lemmatization:** Applied spaCy’s lemmatizer to reduce tokens to base forms, while retaining function words.
- **V4 – Basic Cleaning + Stopword Removal + Lemmatization:** Combined both techniques.

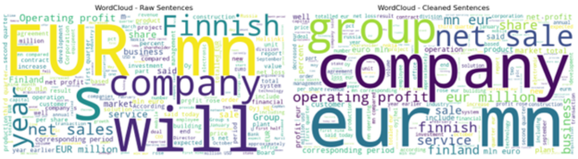

Figure 4: Word clouds of raw (left) vs. cleaned (right) sentences. Preprocessing reduces noise and standardizes terminology while preserving key financial terms.

These variants were evaluated across three vectorizers (TF-IDF, BoW, and Word2Vec) and two classifiers (Logistic Regression and Linear SVM). Variant V3 (lemmatization only) yielded the best macro-averaged F1 scores across most setups. Notably, removing stopwords in V2 and V4 often degraded performance—likely because function words (e.g., “not,” “if,” “but”) carry contrastive or negating sentiment in financial headlines.

6.2 Tokenization and Vectorization

Text data was transformed into numerical form using three different techniques:

- **Bag-of-Words:** Counted word and bigram occurrences. Simple and interpretable, but lacks frequency weighting.
- **TF-IDF:** Scaled term frequencies by inverse document frequency with L2 normalization. Performed best across classical models.
- **Word2Vec:** Represented each sentence as the mean of its token embeddings using pretrained Google News vectors. Struggled with short sentences and domain misalignment.

6.3 Class Balancing with SMOTE

To mitigate the dataset's class imbalance, we initially applied SMOTE (Synthetic Minority Oversampling Technique) to the training sets of classical models. While it improved recall for the minority class (negative), it introduced noise and overlap that ultimately hurt macro F1 performance. As a result, SMOTE was excluded from the final pipeline.

6.4 Preprocessing for FinBERT

As FinBERT uses a subword tokenizer pretrained on financial documents, minimal preprocessing was applied:

- Whitespace normalization and basic character cleaning
- No stopword removal, lemmatization, or lowercasing

This ensured compatibility with the pretrained vocabulary and avoided disrupting token-boundary encoding.

6.5 Implementation Notes

The entire preprocessing workflow was encapsulated in modular Python functions, allowing flexible re-use and batch processing across experiments. This ensured consistency and reproducibility throughout the pipeline and enabled rapid ablation testing of individual steps.

7 Baselines

To fairly evaluate the FinBERT transformer model, we first established strong classical baselines using interpretable and computationally efficient models. These baselines provide reference points for measuring the added value of deep contextual language models in financial sentiment classification.

7.1 Baseline Model Selection

We selected two traditional classifiers—**Logistic Regression (LR)** and **Linear Support Vector Machine (SVM)**—based on their widespread use in text classification and strong performance on high-dimensional sparse data. These models served as competitive benchmarks for evaluating FinBERT.

Both models were paired with **TF-IDF vectorization (1- and 2-grams)** and trained on lemmatized input text (preprocessing Variant V3), as this combination consistently outperformed other configurations during initial experimentation.

The final classical pipeline configurations were:

- **Logistic Regression:** TF-IDF (1,2) + Lemmatization (V3) — Macro F1 = 0.8355 ± 0.0252
- **Linear SVM:** TF-IDF (1,2) + Lemmatization (V3) — Macro F1 = 0.8490 ± 0.0189

7.2 Rationale

These baselines were selected to provide:

- **Interpretability:** Both models offer explainable coefficients for feature importance—useful in auditing decisions or understanding model behavior.
- **Computational Efficiency:** Training and inference are fast and CPU-friendly, making these suitable for deployment in resource-constrained environments.
- **Comparability:** Their performance allows us to isolate the contribution of contextual embeddings and domain-specific pretraining in FinBERT.

7.3 Transformer Benchmark

As the main benchmark model, we fine-tuned the domain-specific **FinBERT** model. FinBERT is a BERT-base architecture pretrained on financial corpora (e.g., SEC filings, earnings calls) and has demonstrated strong performance on financial NLP tasks.

Minimal preprocessing was applied to retain alignment with the pretrained tokenizer. The model was fine-tuned using Hugging Face's `Trainer` API, with early stopping on the validation set's macro F1-score. This deep learning model serves as a state-of-the-art reference point for assessing whether classical methods remain competitive in domain-specific sentiment classification.

7.4 Evaluation Strategy

All models were trained and tuned using stratified 5×5 nested cross-validation on the training set to ensure robust estimates. Performance was finally assessed on a held-out test set (20% of the corpus), and macro-averaged metrics were used to account for class imbalance across all comparisons.

8 Results and Evaluation

8.1 Overall Performance

Table 2 summarizes the top 10 performing configurations in terms of macro-averaged F1-score. All classical models were evaluated on a held-out 20% stratified test set after 5×5 nested cross-validation on the training set.

Preproc	Vectorizer	SMOTE	Model	F1
lemm.	TF-IDF	False	SVM	**0.9049**
basic	TF-IDF	False	SVM	0.9023
basic	TF-IDF	True	SVM	0.9001
lemm.	TF-IDF	True	SVM	0.8881
lemm.	BoW	False	SVM	0.8847
basic	BoW	False	LR	0.8847
basic	BoW	False	SVM	0.8791
lemm.	BoW	False	LR	0.8787
basic	TF-IDF	True	LR	0.8762
lemm.	TF-IDF	True	LR	0.8674

Table 2: Top 10 macro-F1 scoring configurations on the test set

8.2 FinBERT Performance

FinBERT achieved the highest overall scores:

- **Accuracy:** 0.9628
- **Macro F1-score: 0.9697**
- **Precision / Recall:** 0.9680 / 0.9720

FinBERT required minimal preprocessing and no feature engineering, benefiting from contextual embeddings and financial-domain pretraining.

8.3 Confusion Matrices

Figure 5 shows the confusion matrices for all three final models: Logistic Regression, Linear SVM, and FinBERT. FinBERT demonstrates superior ability to disambiguate between positive and negative sentiment while maintaining near-perfect classification of neutral cases.

8.4 Impact of Preprocessing and SMOTE

We observed the following:

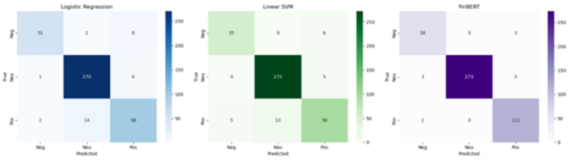

Figure 5: Confusion matrices for Logistic Regression, Linear SVM, and FinBERT on the test set. FinBERT achieves the most accurate class separation.

- Lemmatization-only preprocessing (V3) consistently outperformed other configurations.
- Stopword removal decreased performance due to loss of syntactic context.
- SMOTE improved recall for the minority class but introduced noise, reducing F1 and ultimately being excluded.

8.5 Evaluation Metric Choice

Macro-averaged metrics were prioritized to account for the class imbalance (see Figure 1). Accuracy alone would exaggerate performance due to the overrepresentation of neutral labels.

8.6 Qualitative Error Analysis

To better understand model behavior, we analyzed examples misclassified by all models, and those uniquely misclassified by FinBERT or classical models.

Misclassified by all models:

- *"The group reported a net loss of EUR 3.2 million compared to a profit of EUR 2.4 million last year."* **True:** Negative, **Predicted:** Neutral *Challenge: Requires trend comparison and numerical reasoning.*
- *"Costs were cut across divisions without affecting revenue."* **True:** Positive, **Predicted:** Neutral *Challenge: Sentiment implied indirectly through causal phrasing.*

Correctly classified by FinBERT only:

- *"Operating profit reached a record high in Q2."* — **Positive**
- *"Reported a smaller-than-expected loss."* — **Positive**
- *"The board anticipates weaker demand in the coming quarter."* — **Negative**

FinBERT captures subtle financial modifiers like “record high” and “smaller-than-expected,” which classical models cannot encode via TF-IDF.

Correctly classified by classical models only:

- *"The company will announce its strategic outlook next week."* **True:** Neutral, **FinBERT Prediction:** Positive

- *"Q1 results were in line with market expectations."* **True:** Neutral, **FinBERT Prediction:** Positive

These reflect FinBERT’s tendency to overgeneralize based on semantic priors (e.g., "announce," "results"), highlighting trade-offs between contextual modeling and overfitting on small datasets.

9 Lessons Learned and Conclusion

This project explored the comparative effectiveness of classical machine learning models and a transformer-based domain-specific model (FinBERT) for financial sentiment classification. Using the Financial PhraseBank (All Agree subset), we implemented an end-to-end pipeline that included data exploration, preprocessing variants, feature extraction, model training, evaluation, and interpretability analysis.

9.1 Key Findings

- **FinBERT consistently outperformed classical models**, especially in classifying minority sentiment classes, achieving a macro F1-score of 0.9697. Its domain pretraining allowed it to detect subtle sentiment cues like "smaller-than-expected" or "record high" that are often missed by surface-based models.

- **Classical models remained competitive** when coupled with strong preprocessing (lemmatization) and TF-IDF vectorization. While less powerful in context modeling, they offered faster training times, low resource demands, and crucially, greater interpretability.

- **Preprocessing was critical.** Lemmatization improved generalization by reducing sparsity. Stopword removal, however, often harmed performance by discarding useful syntactic and sentiment-carrying words.

- **SMOTE showed limited success.** Although it helped balance classes in training, it introduced noise and generally reduced macro F1-score, confirming that synthetic resampling may not suit compact or semantically rich financial datasets.

- **Model interpretability remains vital**, especially in regulated environments. FinBERT predictions, while accurate, were opaque. Logistic Regression offered valuable transparency via feature weight inspection, helpful for auditing decisions.

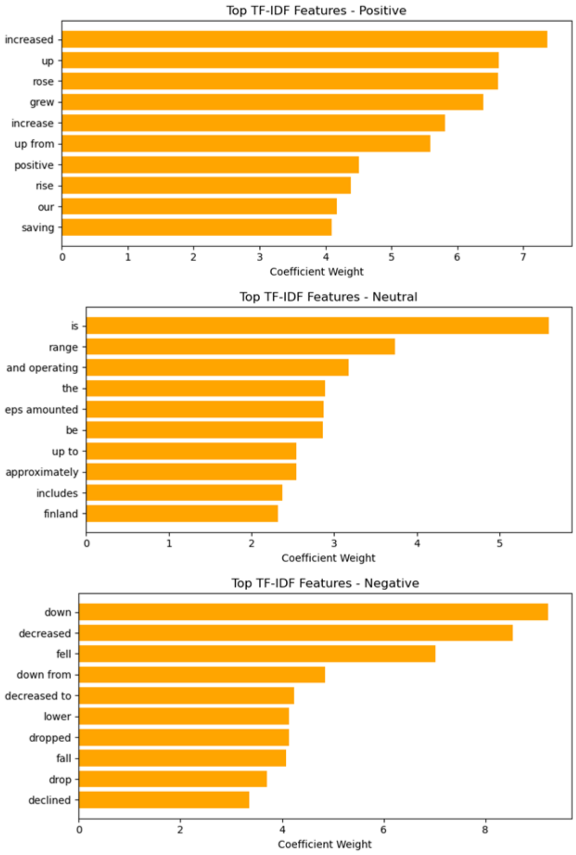

Figure 6: Top weighted TF-IDF features per sentiment class learned by Logistic Regression. Classical models enable transparent feature inspection for positive, neutral, and negative classes.

9.2 Challenges and Limitations

Despite promising results, the study faced several limitations:

- The dataset was relatively small and consisted only of headline-style sentences. This may have favored FinBERT due to its pretraining on similar formats.

- FinBERT's resource demands (GPU usage, slower inference) may limit practical deployment in latency-sensitive or resource-constrained environments.

- Misclassifications revealed that all models struggled with subtle comparative or causally implied sentiment, especially when numerical reasoning or implicit trends were involved.

9.3 Future Work

Building on these insights, we outline several directions for future research:

- **Dataset expansion and diversification:** Incorporate larger and more varied financial text sources (e.g., FiQA (Maia et al., 2018), earnings calls, social media) to capture richer sentiment expressions and investor perspectives.

- **Target-dependent sentiment:** Extend the task to capture entity-specific sentiment (e.g., "oil price rise" positive for ExxonMobil but negative for airlines) using T-BERT or joint sentiment-entity extraction.

- **Explainability in transformers:** Apply LIME, SHAP, or attention visualizations to demystify FinBERT decisions and benchmark them against interpretable baselines.

- **Robustness and multilingual support:** Evaluate model resilience to noise, domain drift, and adversarial examples, and explore multilingual financial models for global use.

- **Deployment considerations:** Explore lighter alternatives like DistilFinBERT or ONNX conversion for real-time inference in trading dashboards or compliance systems.

9.4 Conclusion

In summary, this study confirms that transformer-based models such as FinBERT provide state-of-the-art performance for financial sentiment analysis. However, with thoughtful preprocessing, classical models remain valuable—particularly when explainability, speed, or infrastructure constraints are primary concerns. By highlighting both performance and interpretability trade-offs, this project contributes practical insights to the growing field of financial NLP.

References

Dilan Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rui Du, Weijia Chen, and Sheng Liu. 2024. Challenges and trends in financial sentiment analysis: A comprehensive review. *Information Processing & Management*. Forthcoming.

Vasiliki Karanikola, Eleni Papagiannopoulou, and Andreas Tsakalidis. 2023. A comparative study of transformer-based models for financial sentiment analysis. *Expert Systems with Applications*.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Miguel Maia, Alexandra Balahur, and Caroline Brun. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Pasi Malo, Arjun Kumar Sinha, Panu Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*, 65(4):782–796.

Carlos Pan, Marta Dominguez, and Alejandro Ruiz. 2023. Target-dependent sentiment analysis in multilingual financial news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.