Sentiment Analysis of Financial News Headlines

Yilong Wang*
Technische Universität Berlin
Germany
yilong.wang@campus.tu-berlin.de

Linda Li*
Technische Universität Berlin
Germany
linda.li@campus.tu-berlin.de

Figure 1: Word cloud of the most frequent content words across all sentences in the Financial PhraseBank corpus.

1 Introduction

Financial sentiment analysis differs from general opinion mining in a fundamental way: sentiment is rarely conveyed through explicit evaluative words, but instead emerges from numerical changes, comparisons, and domain-specific reporting conventions. We study these challenges using the *Financial PhraseBank* corpus [1], which contains short financial news sentences labelled as positive, neutral, or negative. The corpus is strongly imbalanced toward the neutral class and rich in financial terms that overlap across classes, making simple keyword-based classification difficult.

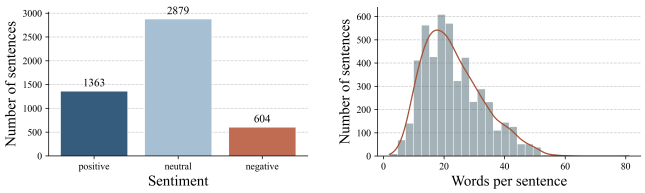
We first analyze class distribution, sentence length, vocabulary, and frequent n-grams, and use these observations to design a conservative preprocessing strategy that preserves financial entities, numerical placeholders, and directional expressions. We then compare Naive Bayes and a feed-forward neural network using Bag-of-Words (BoW) and TF-IDF features, examine PMI-based word similarities, and evaluate a fine-tuned DistilBERT model. Our experiments show that contextual pre-training is the decisive factor: DistilBERT surpasses the best classical model by 13% in Macro F1, while polar-versus-neutral confusion remains the core difficulty across all approaches.

2 Dataset and Exploratory Analysis

We use the *Financial PhraseBank* corpus [1], following the subset specified for this project. The dataset contains 4,846 financial news sentences labelled as positive, neutral, or negative.

2.1 Dataset Characteristics

The corpus is imbalanced and dominated by short to medium-length sentences. Figure 2 shows that neutral examples dominate the corpus, while negative ones are comparatively rare. This makes



(a) Class distribution **(b) Sentence length distribution**

Figure 2: Dataset statistics for Financial PhraseBank.

accuracy alone less informative, so we use macro-averaged metrics in later experiments. Sentence lengths are right-skewed, and longer texts may contain multiple facts or comparisons that make sentiment harder to identify.

The vocabulary is highly domain-specific and sentiment is often expressed indirectly. Figure 1 shows that the most frequent content words are mainly financial reporting terms such as company references, currencies, sales, profit, and numerical units rather than explicit opinion words. This increases lexical overlap across classes and makes sentiment classification more challenging.

2.2 Frequent Lexical Patterns

Frequent words and n-grams reflect recurring reporting templates rather than clear class-specific cues. Common bigrams such as *net sales*, *operating profit*, *eur million*, and *corresponding period* show that many sentences follow standard financial reporting patterns. These phrases are informative about topic and style, but they often appear across multiple sentiment classes. As a result, raw Bag-of-Words features may have limited discriminative power, which motivates the use of weighted representations such as TF-IDF and contextual models in later tasks.

2.3 Implications for Classification

Three challenges follow from these observations: ① class imbalance makes accuracy misleading, so we use Macro F1 as the primary

*❄ Equal contribution — both authors are co-first authors.

NLP Course 2026, TU Berlin
2026.

Table 1: Example of the preprocessing output.

Original	According to the company's updated strategy, Basware targets long-term net sales growth in the range of 20%-40%.
full	according to the company updated strategy basware targets long term net sales growth in the range of <num> <percent> <num> <percent>
filter	according company updated strategy basware targets long term net sales growth range <num> <percent> <num> <percent>

metric; ② sentiment depends on numerical direction and multi-word phrasing rather than isolated keywords, which structurally limits token-independent features such as BoW; ③ high lexical overlap across classes makes feature weighting a critical design choice. These observations motivate the preprocessing strategy and model selection in the following sections.

3 Text Preprocessing

3.1 Preprocessing Strategy

We use a domain-aware preprocessing pipeline. Following § 2, the goal is to reduce surface variation while preserving financial cues relevant to sentiment. All sentences are lowercased; non-informative punctuation and possessive endings are removed; compact financial expressions are split; and numerical values are normalized to placeholders such as <num>, <percent>, and <negnum>. We retain frequent markers such as *eur*, *mn*, and *mln*, since numerical and reporting expressions often carry sentiment in this domain. Table 1 provides an example of the preprocessing output.

3.2 Full and Filtered Input

We construct two variants to balance context preservation and sparsity reduction. The *full* variant keeps all normalized tokens, whereas the *filter* variant additionally removes standard English stop words while preserving negation and directional terms such as *not*, *more*, *less*, *up*, and *down*. We do not apply stemming or lemmatization, because aggressive normalization may distort domain-specific expressions. The resulting representations remain compact while preserving recurring financial patterns such as *eur* *<num>* *mn* and *<num>* *<percent>*.

4 Sentiment Classification

4.1 Experimental Setup

The corpus is partitioned into training and test sets via stratified random sampling, yielding 3,876 sentences (80 %) for training and 970 (20 %) for evaluation. We evaluate five configurations formed by combining models, feature representations[1], and input preprocessing as described below.

4.1.1 Models. **Naive Bayes (NB)** assigns each document to the class with the highest posterior probability under the assumption of conditionally independent terms, using MultinomialNB (α=0.5). No explicit imbalance correction is needed, as the class priors are estimated directly from the training distribution. **Feed-Forward Network (FFN)** is a two-layer MLP comprising hidden layers of 256 and 128 units with ReLU activations, trained using the Adam optimiser ($\mathrm{lr} = 10^{-3}$) with early stopping on validation loss.

4.1.2 Feature Representations. **Bag-of-Words (BoW)** represents each document as a vector of raw term counts, treating the text as an unordered multiset of tokens. **TF-IDF** scales each term count by the inverse document frequency, down-weighting tokens that occur frequently across the corpus and amplifying rarer, more discriminative terms.

4.1.3 Input Preprocessing. Each model–representation combination is evaluated on both the *full* and *filtered* text variants defined in § 3, yielding five configurations in total.

4.2 Results

Table 2 reports four metrics for each configuration: **Accuracy**, **Macro Precision (P)**, **Macro Recall (R)**, and **Macro F1**. We focus on Macro F1 as the primary metric because it weights all three classes equally and is therefore more informative than accuracy under the class imbalance present in this dataset. FFN with TF-IDF on *full* text is the strongest overall model, outperforming the best NB configuration by 7% in Macro F1.

Table 2: Test-set results for all five configurations. Best value per column in bold.

Model	Features	Input	Acc.	P	R	F1
NB	BoW	*full*	0.709	0.645	0.657	0.644
	TF-IDF	*full*	0.726	**0.750**	0.580	0.620
	TF-IDF	*filtered*	0.718	0.735	0.575	0.613
FFN	TF-IDF	*full*	**0.755**	0.719	**0.710**	**0.714**
	TF-IDF	*filtered*	0.744	0.703	0.701	0.702

BoW outperforms TF-IDF on Macro F1 within the NB family, despite lower accuracy. TF-IDF leads in accuracy (Table 2), but at the cost of recall: it down-weights high-frequency terms such as *increased* and *declined* that serve as reliable polarity cues in financial text, pushing predictions toward the neutral class.

Stop-word removal consistently hurts Macro F1 across both model families. Macro F1 falls by roughly 1% for NB and close to 2% for FFN compared to the respective *full* variants, even though key directional words are retained. This suggests that the surrounding function words provide context the models need to identify sentiment-bearing phrases.

4.3 Error Analysis

Figure 3 and Table 3 reveal consistent error patterns across all five configurations.

Polar-Neutral discrimination is the dominant failure mode. Positive-to-Neutral and Neutral-to-Positive errors account for the largest share of misclassifications across all models. Sentences announcing acquisitions or revenue milestones carry positive sentiment for human raters, yet adopt the detached register of financial headlines; within a single short sentence they are lexically indistinguishable from genuinely neutral text.

[1] NB is evaluated with both BoW and TF-IDF; the FFN uses TF-IDF only, as its learned weights subsume the role of the count-based BoW signal.

Figure 3: Confusion matrices for all five configurations. The dotted line separates NB (a–c) from FFN (d–e).

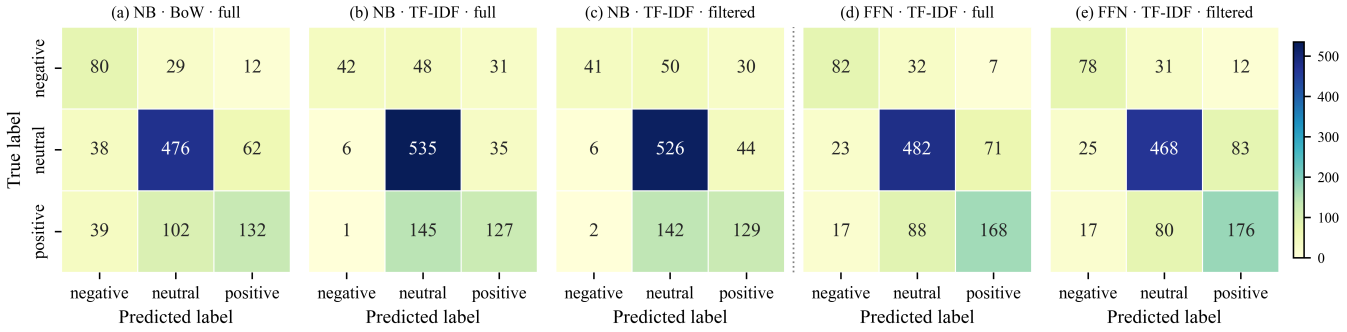
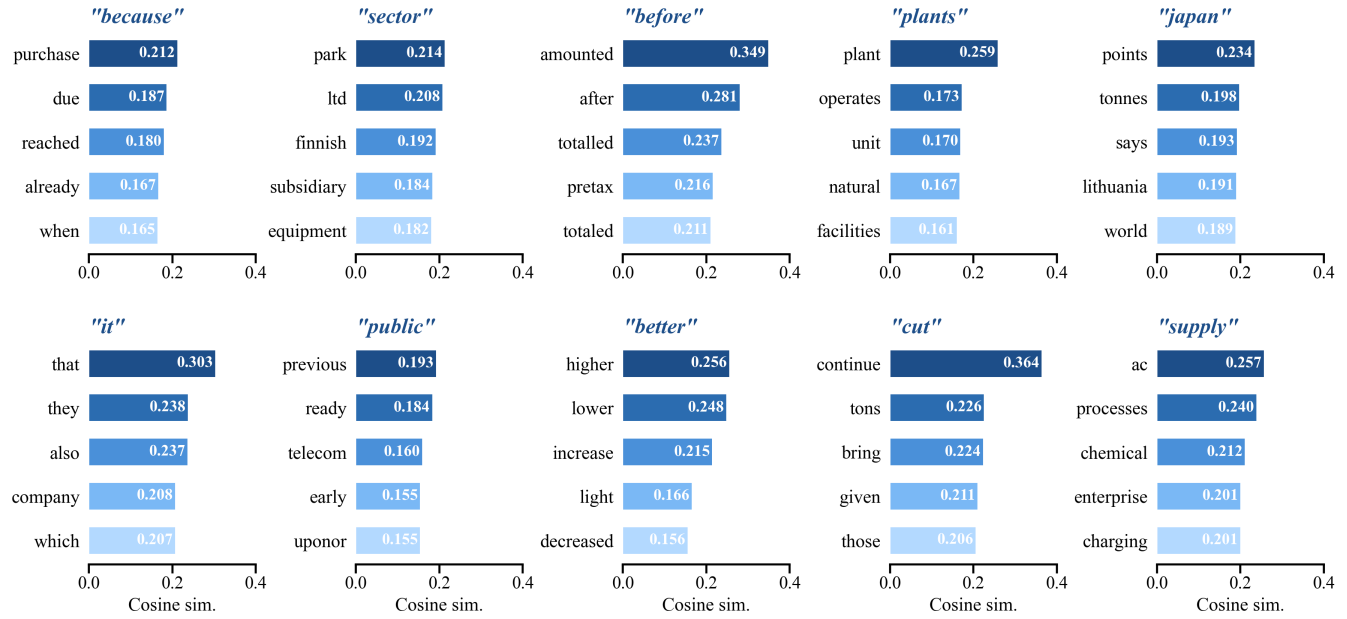


Figure 4: Top-5 most similar words for each of the 10 query terms, ranked by cosine similarity of PPMI vectors. Bars are coloured darkest-to-lightest from rank 1 to rank 5.



FFN reduces Cross-Polar confusion but not the Polar-Neutral boundary. FFN outperforms the best NB by 7% in Macro F1, with the gain concentrated on cross-polar errors, which are roughly halved. Polar-Neutral confusion persists and slightly worsens, as oversampling shifts the decision boundary toward minority classes.

Class imbalance creates a structural bias toward neutral. With 59.4 % of training examples labelled neutral, all models default to neutral under uncertainty, and higher capacity reduces direct polar confusion without resolving Polar-Neutral ambiguity. The substantially higher Macro F1 in the binary setting confirms that polar-neutral discrimination is the core difficulty of this task.

5 PMI-based Word Similarity

5.1 Method

We represent words as vectors using **pointwise mutual information (PMI)**. From the *full* preprocessed corpus we build a word–word co-occurrence matrix with a symmetric context window of size one: for each token, only its immediate left and right neighbours are counted as context. We keep words that occur at least ten times, which gives a vocabulary of 1,280 words over 154,706 co-occurrence events.

From these counts we compute the **positive PMI (PPMI)** of each word–context pair (w, c),

$$\mathrm{PPMI}(w, c) = \max\left(0,\ \log_2 \frac{P(w,c)}{P(w)\,P(c)}\right),$$

where $P(w, c)$ is the joint co-occurrence probability and $P(w)$, $P(c)$ the marginals, estimated from the counts. The resulting matrix is sparse: only 2.4 % of its cells are non-zero. Each word is then represented by its PPMI row vector, and neighbours are ranked by **cosine similarity**:

$$\mathrm{sim}(u, v) = \frac{\mathbf{v}_u \cdot \mathbf{v}_v}{\|\mathbf{v}_u\|\,\|\mathbf{v}_v\|},$$

where $\mathbf{v}_u$, $\mathbf{v}_v$ are the PPMI vectors of two words u and v.

Table 3: Misclassification counts for all five configurations and their average. Column labels correspond to the panels in Figure 3: (a) NB+BoW+*full*, (b) NB+TF-IDF+*full*, (c) NB+TF-IDF+*filtered*, (d) FFN+TF-IDF+*full*, (e) FFN+TF-IDF+*filtered*.

True → Predicted	(a)	(b)	(c)	(d)	(e)	Avg.
positive → neutral	102	145	142	88	80	111.4
neutral → positive	62	35	44	71	83	59.0
negative → neutral	29	48	50	32	31	38.0
neutral → negative	38	6	6	23	25	19.6
negative → positive	12	31	30	7	12	18.4
positive → negative	39	1	2	17	17	15.2

Table 4: DistilBERT results versus the best classical model.

Model	Acc.	P	R	F1
FFN · TF-IDF	0.755	0.719	0.710	0.714
DistilBERT (3-class)	0.816	0.785	0.834	0.805
DistilBERT (binary)	**0.944**	**0.927**	**0.948**	**0.936**

5.2 Results

PPMI captures local, mostly syntactic relations. Figure 4 shows the top-5 neighbours of each query word. These rankings are driven by shared local context rather than deep meaning. For example, *better* is closest to *higher*, *lower*, *increase*, and *decreased*: these words are grouped together not because they are synonyms, but because they appear in the same comparison contexts in financial reports. The other query words follow the same logic, pairing with terms that occupy the same syntactic slots.

6 Pre-trained Language Models

6.1 Experimantal Setup

We fine-tune DistilBERT [2],[2] a 66 M-parameter model with six transformer layers, in both the 3-class and the binary settings. The model is trained on the raw sentences with its own WordPiece tokenizer; we apply no additional preprocessing, since the subword vocabulary already handles rare and compound financial terms. We use AdamW with learning rate 2×10^{-5}, weight decay 0.01, batch size 32, and at most five epochs.

6.2 Results and Discussion

DistilBERT clearly outperforms the classical models. Table 4 compares it with the best classical model from §4 (FFN with TF-IDF). On the 3-class task DistilBERT reaches 0.805 Macro F1, 9.1 points above FFN, with the largest gain on the *negative* class, whose recall rises from 0.71 to 0.88. This reflects the benefit of contextual modelling: financial sentiment often depends on multi-word patterns that BoW and TF-IDF cannot represent, and pre-training adds general-language knowledge unavailable from this small corpus.

The neutral class remains the main source of difficulty. The binary task is much easier (0.936 Macro F1) than the 3-class task, confirming that *neutral* drives most errors. The confusion matrices (Figure 5) keep the same structure as the classical models but with less off-diagonal mass: the dominant error is still *neutral*-to-*positive*, while *positive*-to-*neutral* and *negative*-to-*neutral* clearly reduced.

Figure 5: Confusion matrices of DistilBERT on the 3-class and binary sentiment classification tasks.

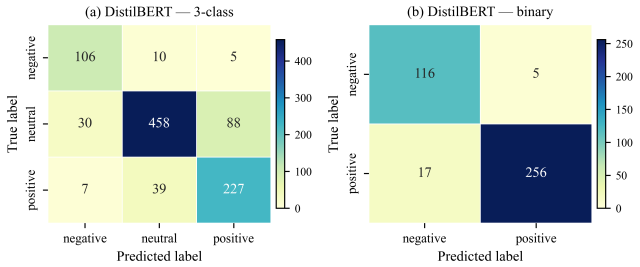
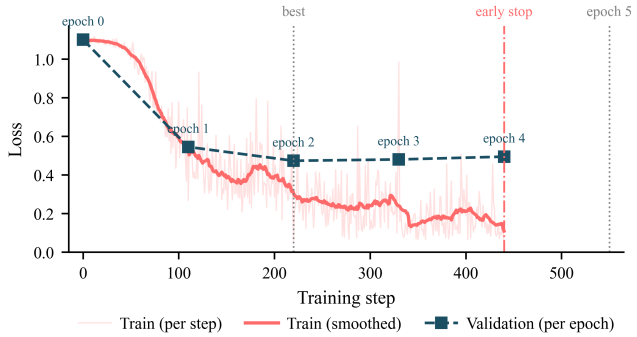


Figure 6: Training and validation loss of DistilBERT on the 3-class task. Validation is lowest at epoch 2 (*best*) and rises afterwards, triggering *early stop* at epoch 4.



Training is stable, but the 3-class model overfits. As Figure 6 shows, the 3-class training loss falls steadily while the per-epoch validation loss bottoms out at epoch 2 and then rises, so early stopping restores the epoch-2 checkpoint. This is expected with only about 3,500 training sentences; the binary model shows no such divergence and trains for all five epochs.

7 Conclusion

We compared classical and neural approaches to financial sentiment analysis on the *Financial PhraseBank* corpus across five tasks. Within the classical family, BoW achieves higher Macro F1 than TF-IDF under Naive Bayes, because TF-IDF down-weights frequent polarity cues such as *increased* and *declined*; stop-word removal consistently hurts both model families for the same reason. DistilBERT surpasses the best classical model in Macro F1 by 13%, demonstrating that contextual pre-training is the decisive factor in this domain. Nevertheless, all models share the same failure mode: polar-versus-neutral confusion accounts for the majority of errors regardless of architecture, and the binary Macro F1 of 0.936 confirms that distinguishing polar from neutral sentiment is the core difficulty of this benchmark.

References

[1] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65 (2014).

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[2] https://huggingface.co/distilbert/distilbert-base-uncased