CIS 5300: Milestone 2

Team Members: Nick Cirillo, Suha Memon, Kalen Truong, Bruce Zhang

1. Evaluation Measures

To assess the performance of our financial sentiment analysis model, a 3-class classification task (Negative: 0, Neutral: 1, Positive: 2), we use standard multi-class metrics. Given the potential for class imbalance in financial datasets, we prioritize metrics that account for performance across all sentiment categories.

1.1 Primary Metric: Macro-Averaged F1-Score

We selected **Macro-averaged F1-score** as our primary metric. By calculating the unweighted mean of per-class F1 scores, this metric treats all sentiment classes equally, ensuring that the model's performance on minority classes is not overshadowed by the majority class.

$Macro\ F1\ = \frac{1}{C}\sum_{c=1}^{C} F1_c$, where $C = 3$ is the number of classes.

1.2 Secondary Metrics

For completeness and detailed error analysis, we also report the following:

- Accuracy: The overall proportion of correct predictions.
  $Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(y_i = \widehat{y}_i)$, where n is total samples, y_i is the true label, and $\widehat{y}_i$ is the prediction.
- Precision and Recall (Per-Class): Calculated for each class c.
  $Precision_c = \frac{TP_c}{TP_c+FP_c}, \quad Recall_c = \frac{TP_c}{TP_c+FN_c}$
- F1-Score (Per-Class): The harmonic mean of precision and recall.
  $F1_c = 2 \cdot \frac{Precision_c \cdot Recall_c}{Precision_c+Recall_c}$
- Weighted-Averaged F1: The average F1 weighted by the support (number of true instances, n_c) of each class.
  $Weighted\ F1\ = \sum_{c=1}^{C} \frac{n_c}{n} \cdot F1_c$

---

2. Baselines and Performance

We established two baselines to frame our model's performance: a simple majority-class baseline to set a minimum floor, and a strong pre-trained baseline to set a competitive target. Both were evaluated on a held-out test set of **585 examples**.

2.1 Simple Baseline: Majority Class Predictor

Methodology: This baseline ignores the input text entirely. It identifies the most frequent class in the training set (which is **Neutral**, comprising 54.29% of training data) and predicts "Neutral" for every instance in the test set.

Performance:

As expected, the model achieves an accuracy equal to the proportion of neutral examples in the test set. However, it fails completely to identify negative or positive sentiments. Overall, the performance is poor.

- **Accuracy:** 48.55%
- **Macro-F1:** 0.2179
- **Weighted-F1:** 0.3173

Class	Precision	Recall	F1-Score	Support
Negative	0.0000	0.0000	0.0000	100
Neutral	0.4855	1.0000	0.6536	284
Positive	0.0000	0.0000	0.0000	201

2.2 Strong Baseline: FinBERT

Methodology: We used **FinBERT** (ProsusAI), a BERT model pre-trained specifically on financial text. Unlike the simple baseline, FinBERT analyzes the semantic content of the headlines to predict sentiment.

Performance:

FinBERT demonstrates a significant improvement over the simple baseline (+25.8% accuracy), validating the necessity of semantic analysis. It performs strongest on Positive-sentiment news but shows some confusion between Negative and Neutral classes.

- **Accuracy:** 74.36%
- **Macro-F1:** 0.7295
- **Weighted-F1:** 0.7508

Class	Precision	Recall	F1-Score	Support
Negative	0.5200	0.7800	0.6240	100
Neutral	0.7753	0.7289	0.7514	284
Positive	0.8929	0.7463	0.8130	201

Confusion Matrix (FinBERT):

The confusion matrix below highlights that the primary source of error is the misclassification of Neutral (61 instances) and Positive (11 instances) headlines as Negative, resulting in lower precision (0.52) for the Negative class.

		Predicted		
		Neg	Neu	Pos
Actual	Neg	78	20	2
	Neu	61	207	16
	Pos	11	40	150

2.3 Conclusion

The gap between the Simple Baseline (Macro-F1: 0.22) and the Strong Baseline (Macro-F1: 0.73) defines the "performance corridor" for our project. Our goal is to develop a model that significantly outperforms the Majority Class Predictor and approaches or exceeds the performance of the FinBERT baseline.

References

[1] Powers, D. M. (2011). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." *Journal of Machine Learning Technologies*, 2(1), 37-63.

[2] Sokolova, M., & Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks." *Information Processing & Management*, 45(4), 427-437.

[3] Grandini, M., Bagli, E., & Visani, G. (2020). "Metrics for Multi-Class Classification: an Overview." *arXiv preprint arXiv:2008.05756*.

[4] Stehman, S. V. (1997). "Selecting and interpreting measures of thematic classification accuracy." *Remote Sensing of Environment*, 62(1), 77-89.

[5] Rosenthal, S., Farra, N., & Nakov, P. (2017). "SemEval-2017 Task 4: Sentiment Analysis in Twitter." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502-518.

[6] Liu, B. (2012). "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.

[7] Araci, D. (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." *arXiv preprint arXiv:1908.10063*.