Classification of Forward-Looking Statements

Managers often discuss projections and future plans to provide timely and relevant information to investors (Huang et al. 2022). Prior studies have introduced word lists (Li 2010; Muslu et al. 2015; Bozanic et al. 2018) and used machine learning algorithms (Brown et al. 2023) to identify FLS. We fine-tune FinBERT to classify FLS and then compare its performance with other machine learning algorithms.

Our sample includes 3,600 sentences from the MD&A section of 10-Ks of Russell 3000 firms from 1994 to 2019. We use stratified random sampling to ensure a balanced sample: 75% (2,700 sentences) contains at least one forward-looking keyword and the remaining 25% (900 sentences) does not have such keyword. We use the union set of forward-looking keywords identified by Li (2010), Muslu et al. (2015), and Bozanic et al. (2018). We ensure a balanced representation from each industry by selecting 10% of the sample (270 sentences with FLS keywords and 90 without keywords) from each two-digit GICS sector.

We manually classify each sentence into one of the three categories: *Specific FLS, Non-Specific FLS,* and *Not-FLS* (Brown et al. 2023). We label a sentence as *Specific FLS* if it is about the future of the company, as *Non-Specific FLS* if it is future-oriented but could be said of any company (e.g., cautionary language or risk disclosure), and as *Not-FLS* if it is not about the future. Our final sample includes 583 *Specific FLS*, 1,061 *Non-Specific FLS,* and 1,056 *Not-FLS* from the 2,700 sentences with a forward-looking keyword and 12 *Specific FLS*, 35 *Non-Specific FLS,* and 853 *Not-FLS* from the 900 sentences without such keywords.

Our manual labeling results suggest that using the keyword approach results in substantial errors, especially false positives where sentences having a keyword do not discuss a company's future. Specifically, the keyword search is only 39.9% accurate ((583 + 853) / 3,600)

in separating *Specific FLS* from other sentences (including *Non-Specific FLS* and *Not-FLS*). We combine *Non-Specific FLS* with *Not-FLS* because researchers and investors are usually more interested in *Specific FLS* and regard *Non-Specific FLS* as boilerplate. We list examples of both false positive and false negative errors in Appendix.

We tabulate the performance of the machine learning algorithms using the full and reduced training samples in Table 1, panels A and B, respectively. We observe that LLMs outperform other NLP algorithms, with accuracy rates ranging from 68.6% to 80.3%. FinBERT further outperforms BERT (85.3% and 82.2%, respectively). Also, FinBERT's advantage over other algorithms becomes larger (except for NB) when we use smaller training samples (plotted in Figure 1). In sum, our results reinforce our conclusion from the sentiment classification that domain-adapted LLMs can outperform other machine learning algorithms in NLP tasks in financial texts, especially with smaller training samples.

References

Bozanic, Z., D. T. Roulstone, and A. Van Buskirk. 2018. Management Earnings Forecasts and Other Forward-Looking Statements. *Journal of Accounting and Economics* 65 (1): 1–20.

Brown, S. V., L. A. Hinson, and J. W. Tucker. 2023. Financial Statement Adequacy and Firms' MD&A Disclosures. *Contemporary Accounting Research*, *forthcoming*.

Huang, A. H., J. Shen, and A. Y. Zang. 2022. The Unintended Benefit of the Risk Factor Mandate of 2005. *Review of Accounting Studies* 27 (4): 1319–1355.

Li, F. 2010. The Information Content of Forward-Looking Statements in Corporate Filings-a Naïve Bayesian Machine Learning Approach. *Journal of Accounting Research* 48 (5): 1049–1102.

Muslu, V., S. Radhakrishnan, K. R. Subramanyam, and D. Lim. 2015. Forward-Looking MD&A Disclosures and the Information Environment. *Management Science* 61 (5): 931–948.

Appendix: Examples of Forward-looking Statement (FLS) Sentences Mislabeled by the Keyword Method

In this appendix, we list selected sample sentences mislabeled using FLS keywords along with their labels from FinBERT and researchers.

Sentences with FLS keywords (keywords in bold) labeled by researchers and FinBERT as *Not-FLS*

*1) Forward-looking statements give our current expectations or **forecasts** of **future** events.*
*2) The Company **believes** its performance in both areas is excellent.*
*3) The fair value of the option grants **is estimated** on the date of grant using the Black-Scholes option pricing model with assumptions on dividend yield, risk-free interest rate, expected volatilities, expected forfeiture rate and expected lives of the options.*
*4) Through this process, we independently validate net revenues, identify and resolve potential fair value or trade booking issues on a timely basis **and seek** to ensure that risks are being properly categorized and quantified.*

Sentences without FLS keywords labeled by researchers and FinBERT as *Specific FLS*

1) A significant amount of such work is planned for Maritech during 2010.
2) The swaps mature over the next two years.
3) Our total obligation related to condensate purchases expires in 2021.
4) The Company's 12.5% Subordinated Debentures ("Debentures") mature on April 14, 2004 and require principal payments of $1,943,000 on October 14, 2000, and of $2,307,000, $2,125,000, and $2,125,000, respectively on April 14 of 2002, 2003 and 2004.

Sentences without FLS keywords labeled by researchers and FinBERT as *Non-Specific FLS*

1) If the fair value of the reporting unit exceeds its carrying value, there is no potential impairment, and the second step is not performed.
2) Disclosure of a contingency is required if there is at least a reasonable possibility that a loss has been incurred.
3) If the carrying amount of the reporting unit's goodwill exceeds the implied fair value of that goodwill, an impairment loss is recognized in an amount equal to that excess.
4) When products are shipped with terms that require transfer of title upon delivery at a customer's location, revenues are recognized on date of delivery.

Figure 1

Forward-looking statement (FLS) classification accuracy across different sample sizes

This figure presents the FLS classification accuracy rates of NLP algorithms using different training sample size. We use the full sample (100%) and subsets (80%, 60%, 40%, and 20%). For example, at 100% (20%), we use 2,916 (583) as the training sample and 324 (65) as the validation sample. We use a constant testing sample size of 360 sentences, same as in Table 1, Panel A. Accuracy is the number of correctly classified sentences divided by the total number of sentences in the testing sample.

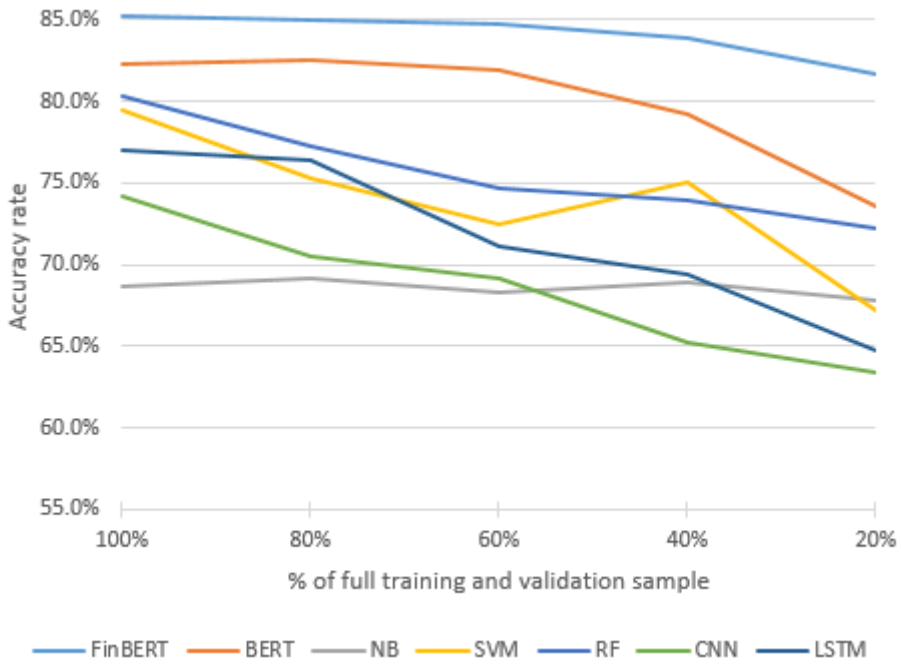

Table 1

Comparison of NLP Algorithms' Performance in Classification of Forward-Looking Statements (FLS)

Panel A: FLS Classification – Full Sample

This panel tabulates the FLS-classification performance of different NLP algorithms in the testing sample. The testing sample contains 360 sentences (59 *Specific FLS*, 110 *Non-Specific FLS*, and 191 *Not-FLS* categories), randomly selected from a sample of 3,600 sentences labeled by researchers. For each NLP algorithm, we report its overall accuracy, precision, recall, F_1 score, and its recall rates in each FLS category. Accuracy is the number of correctly classified sentences divided by the total number of sentences in the testing sample. The overall precision, recall, and F_1 score are the means of the precision, recall, and F_1 score in the three FLS categories. For each FLS category, precision equals the number of sentences that are correctly classified into that category divided by the number of sentences classified into that category by the algorithm; recall equals the number of sentences that are correctly classified into that category divided by the number of sentences classified into that category by researchers; and F_1 score equals the harmonic mean of precision and recall, $F_1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$.

	Overall				*Specific FLS* (59 sentences)	*Non-specific FLS* (110 sentences)	*Not-FLS* (191 sentences)
	Accuracy	*Precision*	*Recall*	$F_1 score$	*Recall*	*Recall*	*Recall*
FinBERT	85.3%	82.1%	81.9%	82.0%	74.6%	79.1%	92.1%
BERT	82.2%	78.2%	77.7%	77.9%	67.8%	73.6%	91.6%
NB	68.6%	64.1%	64.7%	63.8%	45.8%	78.2%	70.2%
SVM	79.4%	77.1%	74.6%	75.7%	64.4%	70.0%	89.5%
RF	80.3%	76.9%	73.8%	74.9%	52.5%	80.0%	89.0%
CNN	74.2%	69.2%	70.2%	69.6%	59.3%	70.0%	81.2%
LSTM	76.9%	72.1%	71.4%	71.7%	54.2%	74.5%	85.3%

Panel B: FLS Classification Accuracy across Different Sample Size

This panel presents the FLS classification accuracy rates of NLP algorithms using different training sample size. We use the full sample (100%) and subsets (80%, 60%, 40%, and 20%). For example, at 100% (20%), we use 2,916 (583) as the training sample and 324 (65) as the validation sample. We use a constant testing sample size of 360 sentences, same as in Table 1, Panel A. Accuracy is the number of correctly classified sentences divided by the total number of sentences in the testing sample.

Training sample	100%	80%	60%	40%	20%	Diff. in Accuracy (100% – 20%)
	Accuracy in the Testing Sample					
FinBERT	85.3%	85.0%	84.7%	83.9%	81.7%	**3.6%**
BERT	82.2%	82.5%	81.9%	79.2%	73.6%	**8.6%**
NB	68.6%	69.2%	68.3%	68.9%	67.8%	**0.8%**
SVM	79.4%	75.3%	72.5%	75.0%	67.2%	**12.2%**
RF	80.3%	77.2%	74.7%	73.9%	72.2%	**8.1%**
CNN	74.2%	70.6%	69.2%	65.3%	63.3%	**10.9%**
LSTM	76.9%	76.4%	71.1%	69.4%	64.7%	**12.2%**