Rillor

METHODS NOTE

Dictionaries, FinBERT or language models for financial sentiment

How dictionaries, FinBERT and prompted language models compare in published financial sentiment tests, and what to record so each label can be reproduced after prompts and models change.



7 November 2026 Public communications and markets rillor.com/insights/financial-sentiment-dictionaries-finbert-llms

Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Dictionaries trail. In every head-to-head study reviewed, a word-count dictionary scores below the leading model, from a recall of 12.5% on positive Financial PhraseBank sentences to 50.1% accuracy against the sign of later returns [5, 17].
- Fine-tuned encoders such as FinBERT and RoBERTa score about the same as or above a language model prompted zero-shot on labels they were trained for: a weighted F1 of 0.717 against 0.587 on FOMC stance, and an accuracy of 0.86 against 0.78 and 0.83 for two hosted models on the PhraseBank [11, 14]. Prompted models lead when the label departs from the encoder's training [12, 15].
- The label definition decides the result. Benchmarks define sentiment as an influence on the stock price, an indication of future policy tightening or easing, an effect on one currency pair or the sign of later returns [5, 11, 15, 17].
- Hosted model labels vary with prompt wording, repeated runs and model updates. In general tasks, a one-word change in the instruction kept Krippendorff's alpha at or below 0.6, and runs at temperature 0 with a fixed seed varied in accuracy by up to 15% [21, 24]. No source reviewed measures this on financial sentiment.
- Every method needs a version. Dated hosted snapshots used in 2023 studies were shut down in 2024 and 2026 [26]. On 12 October 2026 the public FinBERT checkpoint gained a CC BY-NC-SA 4.0 license and a copy of its weights in a new file format, which the commit states is identical [8]. The Loughran-McDonald lists have added and removed words [2].

Background

Gentzkow, Kelly and Taddy describe text as data in three steps: represent the text as numbers, map those numbers to predicted values of an unknown outcome, then use the predictions in a descriptive or causal analysis [4]. Sentiment classification is the second step [4]. Its output, a label such as positive, hawkish or dovish, feeds a later descriptive or causal analysis, such as a study of stock prices [4]. The survey notes a large finance literature that uses internet and news text to predict stock prices [4].

Because the label feeds later analysis, its errors travel. The same authors stress validating the fitted model against the quantity of interest, through manual audits in which people code a subset by hand, and note that twenty or thirty documents are often enough to show whether a model behaves as intended [4].

Three families of method are in use. Finance-specific dictionaries count words from fixed lists, following Loughran and McDonald (2011) [3]. Fine-tuned encoders, such as Araci's FinBERT (2019), train a pretrained language model on labeled financial sentences [7]. Prompted language models, tested on financial text from 2023 onward, classify text from an instruction with few or no labeled examples [11, 13, 14, 15]. This note sets out what the comparisons show, then treats each label as a record that has to be reproducible.

Method

What was examined. Published tests of the three families on English financial text: each benchmark's label definition, the scores reported for each method, the human agreement behind each benchmark, studies of consistency in language model annotation, and the version records of the dictionaries, checkpoints and hosted models involved.

Sources. Thirty sources, each opened on 24 October 2026: the Loughran-McDonald dictionary page, documentation and founding paper [1, 2, 3], a survey of text as data [4], the Financial PhraseBank paper and card [5, 6], the two FinBERT models and the public checkpoint's card and history [7, 8, 9, 10], head-to-head studies on financial text [11, 12, 13, 14, 15, 16, 17, 18, 19], annotation and reliability studies on general text [20, 21, 22, 23, 24], drift and versioning records [25, 26, 27, 28, 29] and a study of look-ahead bias [30].

How. Each figure is taken from its source's tables or abstract, with its dataset, label, metric, model version and run date where the source gives them. Hosted models are described as hosted commercial models, with snapshot dates where recorded. Where sources disagree, both figures are given. The 2025 dictionary counts in Findings 2 and 8 were made for this note from the CSV linked on the dictionary page, downloaded on 24 October 2026, by counting words with a positive year flag (current) and a negative year flag (removed) in each category. The authors do not publish these counts.

Criteria. Each method was read against four questions: what label it was tested on, how far it agrees with human labels, how consistent its labels are across runs and prompts, and whether the same labels can be produced again later.

Findings

1. Define the label before choosing a method

Dataset or study	Unit	What the label means	Who labeled
Financial PhraseBank [5, 6]	News sentence	Whether the news "may have positive, negative or neutral influence on the stock price", from an investor's view	16 annotators, 5 to 8 per sentence
Trillion Dollar Words [11]	FOMC minutes, press conference or speech sentence	Dovish "indicates future monetary policy easing", hawkish "would indicate a future monetary policy tightening"	Two annotators with a guide
Hansen and Kazinnik [12]	FOMC statement sentence, 2010 to 2020	Five steps, by whether the economy is seen as growing too slowly or too quickly	Three reviewers, averaged
Fatouros et al. [15]	Headline tied to a currency pair	Potential short-term impact on that pair	The authors
Kirtac and Germano [17]	News article on a U.S. stock	Sign of the three-day excess return after publication	No human labels

A positive-negative label does not carry over to monetary policy. The FOMC authors note that a sentence containing "increase" can be either dovish or hawkish, and that the Fed often pairs a stance with a moderating counterweight, so they split sentences at words such as "but" and "however" before labeling [11]. Mapped from positive and negative to dovish and hawkish, Araci's FinBERT reached an F1 score of 0.425 on their data, and the authors judged that it does not perform well [11].

A label tied to an instrument differs again. Fatouros et al. cite the headline "CAD is one of the better places to sell the US dollar on a pullback", which a general tool might call positive, correctly for the Canadian dollar and misleadingly for USD/CAD [15]. Kirtac and Germano's accuracy measures agreement with the market's later move [17]. The PhraseBank labels record perceived sentiment and contain no market outcome data [18]. The unit matters too: Hansen and Kazinnik classify each sentence outside the context of its statement [12], while Kirtac and Germano label whole articles [17].

2. Dictionaries: why finance needs its own word lists

Loughran and McDonald studied 50,115 firm-year 10-Ks from 1994 to 2008 and found that 73.8% of the negative word counts from the Harvard list came from words typically not negative in finance, such as tax, cost, capital, board, liability and vice [3]. Others, such as mine, cancer and crude, mark an industry more than a negative event [3]. They built a negative list of 2,337 words and five further lists, noting that misclassified words add noise that attenuates regression coefficients [3]. They handle negation only for positive words, because negative news is often framed with positive words, as in "did not benefit" [3]. They note that words such as increase and decrease depend on whether revenues or costs follow [3].

The current Master Dictionary page, updated March 2026, offers the file Loughran-McDonald_MasterDictionary_1993-2025 with seven sentiment categories: negative, positive, uncertainty, litigious, strong modal, weak modal and constraining [1]. Each sentiment word carries the year it was added, and a negative year marks its removal [1]. The lists are free for academic research, with commercial licenses on request [1]. The authors use explicit inflections in place of stemming and add new tokens each year, such as fracking, bitcoin and covid [2]. Counted for this note, the 2025 file holds 86,553 words, with 2,345 current negative, 347 positive, 297 uncertainty, 903 litigious, 19 strong modal, 27 weak modal and 184 constraining words [1].

On PhraseBank sentences with full agreement, a Loughran-McDonald word count recalled 12.5% of positive and 16.2% of negative sentences, with F1 scores of 0.204 and 0.223 [5]. The authors add that a better lexicon alone boosts performance considerably, as these lists did over a general dictionary [5]. On five-step FOMC stance, the Loughran-McDonald dictionary reached an accuracy of 0.28 and a Cohen's kappa of 0.07, against 0.35 and 0.08 for the Henry financial dictionary [12]. Dictionary methods rarely labeled sentences dovish [12]. Hansen and Kazinnik point to "transitory", which Fed communication used to describe high inflation as temporary, a reading a dictionary may score as positive, neutral or not at all [12].

Gentzkow, Kelly and Taddy write that dictionary methods suit cases where prior information is strong and reliable and the data are weak, and that auditing against hand coding is especially important for them [4]. A dictionary is fixed and transparent, so its version can be stated exactly.

3. Fine-tuned encoders: FinBERT and FOMC stance models

Araci argues that general-purpose models handle the specialized language of finance poorly and that word counts miss deeper meaning [7]. FinBERT is BERT trained further on financial text and fine-tuned on the PhraseBank. It reached an accuracy of 0.86 and an F1 of 0.84 on all sentences, and 0.97 and 0.95 on full agreement, against 0.71 for two earlier methods [7]. Plain BERT reached 0.85 in the same setup [7]. With 250 training examples, FinBERT reached as much as 80% accuracy [7]. A sentence reporting a smaller loss than in an earlier quarter, labeled positive by people, was predicted negative because the model "fails to do the math" [7].

The public checkpoint released with Araci's work outputs softmax scores for three labels, mapped 0 positive, 1 negative and 2 neutral [8]. Its card states that the weights are licensed under CC BY-NC-SA 4.0, which restricts commercial use, and that its publisher provides no support or updates [8].

Yang, Uy and Huang released a different FinBERT, pretrained on 4.9 billion tokens of 10-K and 10-Q filings, earnings call transcripts and analyst reports [9]. It reached 0.872 on the PhraseBank and 0.887 on 10,000 hand-labeled analyst sentences, against 0.835 and 0.850 for uncased BERT [9]. Huang, Wang and Yang report that their FinBERT substantially outperforms the Loughran-McDonald dictionary and five other machine learning methods on analyst sentences and catches sentiment that others label neutral. Its advantage is especially marked when training samples are small [10].

For FOMC stance, Shah, Paturi and Chava labeled 1,070 sentences of minutes, 315 of press conferences and 994 of speeches, with 90.68% agreement between annotators on the combined data after splitting [11]. RoBERTa-large reached a weighted F1 of 0.717, against 0.633 for a fine-tuned FinBERT base model, and its stance measure was tested against Treasury and stock markets and macroeconomic indicators [11]. Data, models and code are released under CC BY-NC 4.0 [11]. In Shah and Chava's FOMC test, the reported labeling time was 3.85 milliseconds for RoBERTa-base and 3.91 seconds for the hosted model in 2023, about 1,000 times higher by the authors' account [13].

4. Prompted language models as annotators

Gilardi, Alizadeh and Kubli tested a hosted commercial model on 6,183 tweets and news articles for relevance, stance, topic and frame detection [20]. Its zero-shot accuracy exceeded crowd workers' by about 25 percentage points on average, against trained annotators' labels [20]. Average agreement between two coders was about 56% for crowd workers, 79% for trained annotators, and 91% and 97% between two model runs at temperatures 1 and 0.2 [20]. The cost was under $0.003 per annotation in 2023 [20]. The texts concern content moderation and politics.

Pangakis, Wolken and Fasching replicated 27 annotation tasks across 11 datasets: median accuracy 0.850, median F1 0.707, with precision or recall below 0.5 on nine tasks [22]. They conclude that every automated annotation with a language model must be validated against human labels [22].

On financial text, FinBen's 36 datasets across 24 tasks and 15 models show language models doing well on extraction and textual analysis and struggling with advanced reasoning, generation and forecasting [16]. An instruction-tuned finance model led its sentiment tasks, with limited ability to generalize [16]. Li et al. found few-shot prompting added 1% to 4% over zero-shot, and a fine-tuned BERT scored 9% above the five-shot hosted model on headline classification [14]. Shah and Chava report that two open 12-billion-parameter models failed to follow the instruction on 6.05% and 42.14% of FOMC sentences, while the hosted model always complied [13]. Vamvourellis and Mehta found that explicit reasoning prompts lowered performance at every agreement level, and their "most accurate and human-aligned combination" was a hosted snapshot of 13 May 2024 with no reasoning prompt [18].

5. Head-to-head results on financial text

Study	Text, label and metric	Dictionary	BERT-family encoder	Language model
Malo et al., 2013 [5]	PhraseBank, full agreement, positive F1	0.204	Not tested	Not tested
Shah, Paturi and Chava, 2023 [11]	FOMC, three-way, weighted F1	0.497 (rule-based)	RoBERTa-large 0.717	0.587 zero-shot, 3 to 4 May 2023
Hansen and Kazinnik, 2023 [12]	FOMC, five steps, accuracy (kappa)	Loughran-McDonald 0.28 (0.07), Henry 0.35 (0.08)	BERT 0.25 (0.03)	0.37 (0.18) zero-shot
Shah and Chava, 2023 [13]	PhraseBank full agreement and FOMC, weighted F1	Not tested	RoBERTa-large 0.976 and 0.698	0.893 and 0.584 zero-shot
Li et al., 2023 [14]	PhraseBank, 50% and full agreement, accuracy	Not tested	FinBERT 0.86 and 0.97	0.78 to 0.86 and 0.90 to 0.97, versions of 7 July 2023
Fatouros et al., 2023 [15]	2,291 forex headlines, accuracy	Not tested	FinBERT 0.561	0.730 to 0.790, four prompts
FinBen, 2024 [16]	PhraseBank and FOMC, F1	Not tested	FinBERT 0.872 accuracy, from an earlier paper	0.78 from earlier papers, and 0.64 and 0.71 on FOMC
Kirtac and Germano, 2024 [17]	965,375 articles, sign of later returns, accuracy	50.1%	FinBERT 72.2% as published, BERT 72.5% fine-tuned on the label	Not prompted. OPT, an open model fine-tuned on the label: 74.4%

Study	Text, label and metric	Dictionary	BERT-family encoder	Language model
Vamvourellis and Mehta, 2025 [18]	PhraseBank, full agreement, macro F1	Not tested	FinBERT 0.962, trained on much of the test set	0.895 and 0.890 with no reasoning prompt (May 2024 and April 2025 snapshots), 0.776 to 0.866 with reasoning prompts or a reasoning model

Language models are prompted hosted commercial models unless the row says otherwise. Kirtac and Germano fine-tuned OPT and prompted no model [17].

Dictionaries trail the leading model in every study that tests one [5, 11, 12, 17]. On five-step FOMC stance, both the Loughran-McDonald and Henry dictionaries scored above BERT and below the prompted model [12].

On labels an encoder was trained for, it scores level with or above a prompted model. The FOMC authors write that the zero-shot model underperformed fine-tuned pretrained language models across all their datasets, though it beat their fine-tuned LSTM models [11]. In Li et al., five-shot prompting brought one hosted model level with FinBERT on accuracy, at 0.86 and 0.97, and slightly above it on F1 at full agreement, 0.97 against 0.95 [14]. Shah and Chava found a wider gap on unpublished data and suggest contamination as a possible reason, and they ran their FOMC tests before that data was released [13].

When the label departs from the encoder's training, prompted models lead. Hansen and Kazinnik's five-step scale has no matching encoder [12]. Fine-tuned on 400 sentences, their hosted model reached an accuracy of 0.61 and a kappa of 0.46 on 100 held-out sentences, against 0.41 and 0.21 for the zero-shot model on the same 100 sentences [12]. On forex headlines, four prompts reached 0.730 to 0.790 against 0.561 for FinBERT [15].

Attribution of the abstract's figures. Fatouros et al.'s abstract reports "approximately 35% enhanced performance in sentiment classification and a 36% higher correlation with market returns" than FinBERT [15]. These come from one 2023 hosted model at a temperature of 0.2, used zero-shot after extensive prompt engineering, on 2,291 headlines for five currency pairs collected over 86 days from January to May 2023 and labeled by the authors [15]. The paper does not show how the 35% is computed, and one prompt mirrors the annotation approach [15]. The authors report a correlation of about 0.49 between true sentiment and daily returns, and state that the alignment does not necessarily imply predictive power over future prices [15].

Kirtac and Germano measure something else. Their figures are agreement with the sign of later returns, with OPT and BERT fine-tuned on that label and FinBERT and the dictionary used as published [17]. They do not compare with human-label benchmarks.

Cross-paper comparisons are weak. FinBen's FOMC scores come from its own 496-sentence test set, averaged over three runs, while its hosted PhraseBank figures come from earlier papers [16]. No source reviewed tests a dictionary, a fine-tuned encoder and a current hosted model on the same human-labeled financial set on the same day.

6. Testing against human labels and benchmarks

Sixteen people with financial market knowledge, three researchers and 13 master's students at one business school, annotated the PhraseBank, with 5 to 8 per sentence and a majority vote for the reference label [5]. In an agreement study on 150 sentences, average pairwise agreement was 74.9% [5]. Annotators separated positive from negative 98.7% of the time and agreed on positive against neutral 75.2% of the time, which the authors trace to the difficulty of telling "commonly used company glitter" from actual positive statements [5]. The dataset card notes that all annotators came from one institution [6].

Its size depends on the source: 4,840 sentences with more than 50% agreement and 2,259 with full agreement in the paper [5], 4,845 in Araci [7], and 4,846 and 2,264 in the card's configurations, with no train or test split provided [6]. Each study draws its own test set.

Scores rise with agreement. The FinBERT checkpoint's macro F1 went from 0.722 on sentences with 50% to 65% agreement to 0.962 at full agreement, and a May 2024 hosted snapshot's from 0.561 to 0.895 [18]. For an April 2025 snapshot, neutral sentences labeled positive made up 14.4% of all sentences [18]. FOMC labels are hard for people too: on 264 of Hansen and Kazinnik's 500 sentences, a reviewer differed by more than one step [12].

The FinBERT checkpoint was trained on a substantial portion of the PhraseBank, so Vamvourellis and Mehta treat its score only as a reference point [18]. Licenses also differ by source. The PhraseBank card states CC BY-NC-SA 3.0, while FinBen lists CC BY-SA 3.0 and says its compiled datasets are shared under the MIT license [6, 16].

Pangakis, Wolken and Fasching propose a test: write a codebook, have at least two subject matter experts and the model label the same random sample of 250 to 1,250 texts, and judge the model on held-out data [22]. Different codebooks for people and model open a gap between the two sets of instructions [22].

7. Reliability across runs, prompts and settings

Reiss classified 234 website texts with ten instructions, two temperatures and ten repetitions, 46,800 inputs, using a hosted model as of 1 April 2023 [21]. Against an alpha of 0.8 as the reliability threshold, consistency between temperature settings was 0.75 for single runs and 0.91 after taking the majority of ten [21]. Two instructions differing only in "classify" or "rate" never exceeded 0.6, and across 45 instruction pairs alpha ranged from 0.24 to 0.7 [21]. A codebook that works for human coders does not necessarily work for the model [21].

Atil et al. ran five models at temperature 0, top-p 1 and a fixed seed, ten times on eight tasks [24]. Accuracy varied by up to 15%, the gap between the possible high and low results reached 70%, and no model delivered repeatable accuracy on all tasks [24]. They suggest the effect may come from efficient batching on shared compute and expect it to persist [24]. Sclar et al. found differences of up to 76 accuracy points from prompt formatting alone in one open model, and a median spread of 6.4 points across 320 formats on a hosted model, and recommend reporting a range across formats [23].

Pooling helps. Pangakis et al. labeled each text at least seven times at temperature 0.6, and accuracy was 19.4 percentage points higher on texts whose runs agreed fully, which was 85.1% of them [22]. Run agreement can flag texts for human review. None of these studies used financial sentiment text [20, 21, 23, 24], so their effect sizes on it are unmeasured.

8. Version drift in word lists and models

Chen, Zaharia and Zou compared March and June 2023 versions of two hosted services. One service's accuracy at identifying prime numbers fell from 84.0% to 51.1%, which they say makes results challenging, if not impossible, to reproduce [25]. They note that earlier drift studies covered simpler machine learning services on tasks such as sentiment analysis, while their own study covers generative services [25].

The vendor's documentation lists the smaller chat model's March and June 2023 snapshots as shut down on 13 September 2024, and the larger model's March 2023 snapshot as shut down on 26 March 2026 [26]. Lopez-Lira and Tang recorded their snapshot of 14 March 2023 [19], one of those shut down, so their labels cannot be regenerated from it [26]. Several 2023 studies name an undated model alias, some with call dates [11, 15, 20, 21], while Li et al. record versions as of 7 July 2023 [14] and Vamvourellis and Mehta record dated snapshots [18]. Spirling argues that owners of closed models can alter them at any time and that versions become obsolete within weeks or months [27]. Palmer, Smith and Spirling add that maintainers change open models too, by re-estimating weights or ceasing maintenance [28].

The public FinBERT checkpoint had no commits between 23 May 2023 and 12 October 2026, and its card of 23 May 2023 states no license [8]. On 12 October 2026, four commits added license terms and disclaimers, reordered the card, added a weight file in a new format, converted from the existing one with outputs the commit states are identical, and added a copyright notice [8]. The Transformers library loads a model at a revision that can be a branch, a tag or a commit id [29].

List	2011 paper [3]	Version used by Hansen and Kazinnik [12]	2025 file, this note's count [1]
Negative	2,337	2,356	2,345
Positive	353	354	347

The documentation's change log records 2012 removals such as casualty, refinance and surrender, 26 negative words added in 2014 such as cyberattack, and 2020 removals of ten negative words such as closed, critical and late and seven positive words such as benefit, effective and great [2]. In the 2025 file, the 2020 removals carry a negative year flag and the 2012 removals carry none [1, 2]. The flags therefore do not rebuild every earlier list, and a study that used an earlier list needs that file.

Leakage is a related risk. Glasserman and Lin show that a model whose training overlaps the test period may know the returns that followed an article, or be distracted by what it knows about the company, and in their in-sample test distraction outweighed look-ahead bias [30]. Lopez-Lira and Tang chose a sample period after their snapshot's training data ended [19].

9. What to store with every label

Field	Why it matters	Sources
Label definition, unit and codebook	Benchmarks define sentiment differently, and people and model should share one codebook	[5, 11, 12, 15, 17, 22]
Label map	0 is positive in the FinBERT checkpoint and dovish in the FOMC dataset	[8, 11]
Method and exact version	Dictionary file and date, checkpoint and commit id, or hosted snapshot	[1, 8, 18, 19, 29]
Date of the run	Aliases change and snapshots are retired	[14, 25, 26]
Prompt, settings and runs	Wording, temperature, format and run count all move results	[15, 21, 22, 23, 24]
Parse failures	Some models return no usable label	[13]
Human benchmark	Annotators, agreement and the subset used	[5, 12, 18]
Training overlap and cutoff	Training on the test set or the period studied inflates results	[13, 18, 19, 30]
License of each resource	Several state non-commercial terms	[1, 6, 8, 11]

Implications

For researchers choosing a method. Start from the label. If a labeled set exists for the exact definition, the evidence favors a fine-tuned encoder, which scores level with or above a zero-shot prompted model [11, 13, 14] and labels in milliseconds where a hosted model takes seconds [13]. If the label is new, such as five-step stance or impact on one instrument, a prompted model can serve as a first annotator, checked against expert labels on a few hundred texts [12, 15, 22]. A dictionary suits work where a fixed, transparent version matters more than context, with an audit against hand coding [4].

For teams producing labels. Treat each label as a dated record. Pin open checkpoints to a commit id [29]. Record hosted snapshots by their dated identifiers and keep the outputs, since a retired snapshot cannot produce them again [19, 26]. Run each text more than once and send disagreements to a person [21, 22]. Report results across prompt formats [23], and record the model's training cutoff against the text's date [30].

For buyers and reviewers of labeled data. Ask for the label definition, the human benchmark behind any accuracy claim, the method version and the date of each label. A score on the PhraseBank's full-agreement subset says little about currency headlines or FOMC stance [15, 18]. Check the license of every dictionary, checkpoint and dataset in the chain [1, 6, 8, 11]. Rillor's public communications and market response data holds selected public statements as issued, with their entities, timestamps and classifications, each classification with its classifier version, and pairs each statement with price observations in a declared window around it.

Limits of this analysis

- No source compares a dictionary, a fine-tuned encoder and a current hosted model on the same human-labeled financial set on the same day, and most head-to-head studies tested 2023 hosted snapshots, several now shut down [26].
- The reliability studies used general text [20, 21, 23, 24], and no source measures hosted model drift on financial sentiment [25].
- Only the abstract of Huang, Wang and Yang was available [10].
- How Fatouros et al. computed "approximately 35%" is not shown, and the license of their headline data was not checked [15].
- Hansen and Kazinnik was read in its version of 5 August 2023 [12]. Cost and timing figures date from 2023 [13, 20].
- The dictionary counts were made for this note from the file as downloaded on 24 October 2026.
- License terms are reported as each resource states them. This note is not legal advice.
- The note covers English text only.

Sources

1. Loughran-McDonald Master Dictionary w/ Sentiment Word Lists. University of Notre Dame, Software Repository for Accounting and Finance, updated March 2026, viewed 24 October 2026. https://sraf.nd.edu/loughranmcdonald-master-dictionary/

2. Tim Loughran and Bill McDonald, Documentation for the Loughran-McDonald Master Dictionary. Software Repository for Accounting and Finance, updated February 2024. https://drive.google.com/file/d/1qMZ4WTDv9bq2iYOxmMP5h_JgdLb46z_h/view

3. Tim Loughran and Bill McDonald, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66(1), 35-65, February 2011. Published article as hosted by the University of Technology Sydney: https://www.uts.edu.au/globalassets/sites/default/files/adg_cons2015_loughran-mcdonald-je-2011.pdf

4. Matthew Gentzkow, Bryan Kelly and Matt Taddy, Text as Data. Journal of Economic Literature 57(3), 535-574, September 2019. https://web.stanford.edu/~gentzkow/research/text-as-data.pdf

5. Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen and Jyrki Wallenius, Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. arXiv:1307.5336, version 2, 23 July 2013. https://arxiv.org/abs/1307.5336

6. financial_phrasebank, dataset card. Hugging Face, last modified 15 December 2025, viewed 24 October 2026. https://huggingface.co/datasets/takala/financial_phrasebank

7. Dogu Araci, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063, 27 August 2019. https://arxiv.org/abs/1908.10063

8. ProsusAI/finbert, model card and commit history. Hugging Face, revision of 12 October 2026, viewed 24 October 2026. https://huggingface.co/ProsusAI/finbert

9. Yi Yang, Mark Christopher Siy Uy and Allen Huang, FinBERT: A Pretrained Language Model for Financial Communications. arXiv:2006.08097, version 2, 9 July 2020. https://arxiv.org/abs/2006.08097

10. Allen H. Huang, Hui Wang and Yi Yang, FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research 40(2), 806-841, 2023. Abstract as posted by the HKUST Department of Accounting, 5 February 2023. http://acct.hkust.edu.hk/about-department/news-highlights/congratulations-prof-allen-huang-paper-contemporary-accounting

11. Agam Shah, Suvan Paturi and Sudheer Chava, Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 6664-6679, July 2023, arXiv:2305.07972. https://arxiv.org/abs/2305.07972

12. Anne Lundgaard Hansen and Sophia Kazinnik, Can ChatGPT Decipher Fedspeak? Federal Reserve Bank of Richmond, version of 5 August 2023, posted for the Federal Reserve Bank of New York FinTech conference. https://www.newyorkfed.org/medialibrary/media/research/conference/2023/FinTech/400pm_Hansen_Paper_Kazinnik_2023.pdf

13. Agam Shah and Sudheer Chava, Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks. arXiv:2305.16633, 26 May 2023. https://arxiv.org/abs/2305.16633

14. Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu and Sameena Shah, Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. EMNLP 2023 Industry Track, arXiv:2305.05862, version 2, 10 October 2023. https://arxiv.org/abs/2305.05862

15. Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis and Dimosthenis Kyriazis, Transforming Sentiment Analysis in the Financial Domain with ChatGPT. Machine Learning with Applications, 2023, arXiv:2308.07935, 13 August 2023. https://arxiv.org/abs/2308.07935

16. Qianqian Xie and others, FinBen: A Holistic Financial Benchmark for Large Language Models. arXiv:2402.12659, version 2, 19 June 2024. https://arxiv.org/abs/2402.12659

17. Kemal Kirtac and Guido Germano, Sentiment trading with large language models. Finance Research Letters 62, 105227, 2024, arXiv:2412.19245. https://arxiv.org/abs/2412.19245

18. Dimitris Vamvourellis and Dhagash Mehta, Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis. arXiv:2506.04574, 5 June 2025. https://arxiv.org/abs/2506.04574

19. Alejandro Lopez-Lira and Yuehua Tang, Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv:2304.07619, version 6, 28 October 2025. https://arxiv.org/abs/2304.07619

20. Fabrizio Gilardi, Meysam Alizadeh and Maël Kubli, ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Proceedings of the National Academy of Sciences 120(30), e2305016120, 2023, arXiv:2303.15056, version 2, 19 July 2023. https://arxiv.org/abs/2303.15056

21. Michael V. Reiss, Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. arXiv:2304.11085, 17 April 2023. https://arxiv.org/abs/2304.11085

22. Nicholas Pangakis, Samuel Wolken and Neil Fasching, Automated Annotation with Generative AI Requires Validation. arXiv:2306.00176, 31 May 2023. https://arxiv.org/abs/2306.00176

23. Melanie Sclar, Yejin Choi, Yulia Tsvetkov and Alane Suhr, Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. International Conference on Learning Representations (ICLR) 2024, arXiv:2310.11324, version 2, 1 July 2024. https://arxiv.org/abs/2310.11324

24. Berk Atil and others, Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667, version 5, 2 April 2025. https://arxiv.org/abs/2408.04667

25. Lingjiao Chen, Matei Zaharia and James Zou, How is ChatGPT's behavior changing over time? arXiv:2307.09009, version 3, 31 October 2023. https://arxiv.org/abs/2307.09009

26. Deprecations. OpenAI API documentation, viewed 24 October 2026. https://platform.openai.com/docs/deprecations

27. Arthur Spirling, Why open-source generative AI models are an ethical way forward for science. Nature 616, 413, 18 April 2023. https://www.nature.com/articles/d41586-023-01295-4

28. Alexis Palmer, Noah A. Smith and Arthur Spirling, Using proprietary language models in academic research requires explicit justification. Nature Computational Science 4, 2, published 21 December 2023. https://www.nature.com/articles/s43588-023-00585-1

29. Models, from_pretrained revision parameter. Hugging Face Transformers documentation, viewed 24 October 2026. https://huggingface.co/docs/transformers/main_classes/model

30. Paul Glasserman and Caden Lin, Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis. arXiv:2309.17322, 29 September 2023. https://arxiv.org/abs/2309.17322

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Dictionaries, FinBERT or language models for financial sentiment. 7 November 2026. https://rillor.com/insights/financial-sentiment-dictionaries-finbert-llms

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Work with Rillor

Send the sources, assets and time windows your study needs, and we set out the records and baselines that cover them. Contact us at rillor.com/contact?type=communications or contact@rillor.com. More on public communications and markets: rillor.com/markets/forecasting#events