

Sentiment Classification of Saudi Stock Market Tweets: Lexicon-Based Weak Supervision, Classical Machine Learning, and Zero-Shot Open-Source Large Language Models

Fahad Kamal Alsheref

Fabdu1wahed@kku.edu.sa

King Khalid, University <https://orcid.org/0000-0001-5894-9583>

Sura Awad Aldahwi

King Khalid University

Research Article

Keywords: Arabic sentiment analysis, financial social media, Saudi Exchange (Tadawul), large language models, weak supervision, TF-IDF, text classification

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11394067/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Social media platforms have become a rich, real-time source of investor opinion, yet Arabic-language financial text remains under-studied, and much applied work validates classifiers against labels produced by the very rules they later learn to imitate. This study examines sentiment classification of 1,000 Arabic tweets about companies listed on the Saudi Exchange (Tadawul), issued by 41 accounts. Using a domain lexicon of 180 financial terms, tweets were assigned positive, negative or neutral labels, and four classical models (Multinomial Naïve Bayes, Logistic Regression, Random Forest and a linear Support Vector Machine) were trained on TF-IDF features. Under stratified five-fold cross-validation the linear SVM reproduced the lexicon labels with an accuracy of 0.861 and a macro-F1 of 0.819, whereas Naïve Bayes reached only 0.446 macro-F1. We then added a zero-shot open-source large language model (LLM), Qwen2.5-7B-Instruct, run locally with 4-bit quantisation and no task-specific training. Its agreement with the lexicon labels was low (accuracy 0.376, macro-F1 0.401, Cohen's $\kappa = 0.19$), driven by a strong tendency to label lexicon-neutral tweets, such as dividend declarations and earnings announcements, as positive. A diagnostic analysis showed that this disagreement reflects, to a large extent, weaknesses in the lexicon labels: the lexicon terms were not subjected to the same orthographic normalisation as the tweets, so only 42.4% of tweets matched any term; after correcting this, the match rate rose to 67.8%, 30% of labels changed, and LLM–lexicon agreement rose to 57.5% ($\kappa = 0.31$). We conclude that high accuracy against rule-derived labels demonstrates label reproduction rather than sentiment understanding, that zero-shot LLMs offer a label-free, domain-aware alternative whose outputs differ systematically from lexicon rules, and that a human-annotated reference set is indispensable for establishing which is correct. We release the pipeline, prompt and an annotation protocol to support such validation.

1. Introduction

Investor sentiment, the collective mood of market participants, is increasingly regarded as a determinant of price formation that complements fundamentals. Early evidence showed that the tone of financial news predicts short-horizon returns (Tetlock, 2007) and that aggregate mood extracted from microblog posts carries information about index movements (Bollen et al., 2011). The proliferation of financial discussion on platforms such as X (formerly Twitter) has since made large-scale, near-real-time measurement of investor opinion feasible, and a substantial literature now surveys the techniques and applications of financial sentiment analysis (Du et al., 2024).

The Saudi Exchange (Tadawul) is the largest equity market in the Middle East and has attracted growing participation under the Vision 2030 reforms. Discussion of Saudi equities on social media is predominantly in Arabic, a morphologically rich language with dialectal variation, optional diacritics, orthographic ambiguity and informal spelling, all of which complicate automated analysis (Abu Farha & Magdy, 2021). Financial text adds a further layer of difficulty: the same word may carry opposite valence depending on context (for instance, “correction” in technical analysis), and many posts are factual announcements rather than expressions of opinion.

Practical sentiment pipelines for under-resourced domains frequently follow a common recipe: build a keyword lexicon, use it to label a corpus, and train a supervised classifier on the resulting labels. High cross-validated accuracy is then reported as evidence of success. This recipe is vulnerable to a well-known problem in weak supervision: a flexible classifier trained on labels generated by a deterministic rule will learn to approximate the rule, so its accuracy measures fidelity to the labelling function rather than to the underlying construct (Ratner et al., 2017). Label noise of this kind is also known to distort both training and evaluation (Northcutt et al., 2021).

Recent instruction-tuned large language models (LLMs) offer an alternative that requires neither a hand-built lexicon nor labelled data. They can be prompted to classify sentiment directly, and open-weight models can be executed locally on consumer hardware through quantisation (Dettmers et al., 2023), which matters for reproducibility, cost and data governance. Their reliability for Arabic financial microblogs, however, has received little empirical attention.

This paper therefore asks three research questions:

RQ1. How well do classical TF-IDF-based classifiers reproduce lexicon-derived sentiment labels for Saudi stock tweets?

RQ2. How does a zero-shot, open-source LLM compare with those labels, and where do the two disagree?

RQ3. How sensitive are the lexicon labels to preprocessing choices, and what does this imply for the interpretation of reported accuracies?

The study makes three contributions. First, it provides a transparent end-to-end comparison of lexicon, classical machine-learning and LLM approaches on an Arabic Saudi-market corpus. Second, it documents a concrete, quantified instance of label-construction error (a normalisation mismatch between lexicon and text) and traces its consequences. Third, it provides a reproducible local-LLM protocol and an annotation protocol for the human validation set that the evidence shows to be necessary.

2. Literature Review

2.1 Social media sentiment and financial markets

The empirical link between textual tone and asset prices is well established. Tetlock (2007) showed that pessimism in a widely read newspaper column predicts downward pressure on market prices followed by reversion, and Loughran and McDonald (2011) demonstrated that general-purpose sentiment dictionaries misclassify many words in financial documents, motivating domain-specific lexicons. Bollen et al. (2011) extended this line to microblogging by showing that public mood measured from Twitter feeds was associated with subsequent changes in the Dow Jones Industrial Average. Mehta et al. (2021)

combined social media sentiment with deep learning for stock-market prediction, reporting that sentiment features improved forecasts relative to price-only inputs. The broader survey by Du et al. (2024) organises financial sentiment analysis into lexicon-based, machine-learning, deep-learning and pre-trained language-model approaches and highlights data scarcity, domain shift and label quality as persistent obstacles.

2.2 Sentiment-augmented price prediction with deep learning

A large share of applied research couples sentiment with sequence models for price or trend prediction. Long Short-Term Memory networks (Hochreiter & Schmidhuber, 1997) are widely used because of their ability to model temporal dependence, and several studies fuse numerical indicators with sentiment signals. Wu et al. (2022) proposed S_I_LSTM, which integrates multiple data sources, technical indicators and investor sentiment, and reported improved prediction accuracy over price-only baselines. Behura et al. (2023) examined multi-layered sequential LSTM architectures for price prediction. Gu et al. (2024) combined FinBERT-derived news sentiment with an LSTM to forecast prices, in line with the finding that finance-specific encoders extract more useful sentiment than general ones (Araci, 2019). Collectively, this literature treats sentiment as an input feature; it generally presumes, rather than examines, that the sentiment labels themselves are valid.

2.3 Classical and transformer-based text classification

For short-text classification, linear models over TF-IDF features remain strong baselines. Support Vector Machines were shown to be particularly effective for high-dimensional sparse text representations (Joachims, 1998), Random Forests offer robust non-linear classification (Breiman, 2001), and Naïve Bayes and Logistic Regression provide fast, interpretable references. Transformer encoders pre-trained with masked language modelling (Devlin et al., 2019) substantially raised the state of the art, and Arabic-specific models, including AraBERT (Antoun et al., 2020), ArabicBERT (Safaya et al., 2020) and ARBERT/MARBERT (Abdul-Mageed et al., 2021), have been benchmarked on Arabic sentiment and sarcasm detection (Abu Farha & Magdy, 2021). These models typically require labelled training data; their quality is therefore bounded by that of the labels.

2.4 Large language models for sentiment analysis

Instruction-tuned LLMs can perform classification from a natural-language description of the task without parameter updates (Brown et al., 2020). Zhang et al. (2023) evaluated LLMs across a range of sentiment tasks and found that they perform competitively in simple zero-shot polarity classification while being less reliable on fine-grained or structured tasks and sensitive to prompt design. Open-weight families such as Qwen2.5 (Yang et al., 2024), which include multilingual coverage of Arabic, together with 4-bit quantisation (Dettmers et al., 2023), make it feasible to run such models on a single consumer GPU. To our knowledge, systematic evaluation of zero-shot open LLMs on Saudi-market Arabic tweets, relative to both lexicon rules and supervised baselines, has not been reported.

2.5 Label quality, weak supervision and validity

Weak supervision programmatically generates training labels from heuristics such as keyword rules (Ratner et al., 2017). The resulting labels are noisy and systematically biased toward what the heuristics can detect. When the same heuristic also defines the evaluation target, accuracy conflates agreement with the rule and correctness. In communication research, van Atteveldt et al. (2021) compared manual annotation, crowd-coding, dictionary and machine-learning approaches to sentiment and concluded that dictionaries often have poor validity unless validated against human coding, with manual annotation the appropriate gold standard. Confident-learning methods further show that even small rates of label error can alter model rankings (Northcutt et al., 2021). These considerations motivate the validity-oriented design of the present study.

2.6 Research gap

In sum, (i) Arabic financial microblog sentiment for Tadawul remains sparsely studied; (ii) lexicon-and-classifier pipelines are common but rarely audited for label validity; and (iii) the behaviour of zero-shot open LLMs relative to such pipelines is unexamined. The present study addresses these gaps empirically.

3. Methodology

3.1 Data

The corpus consists of 1,000 Arabic tweets about Saudi-listed companies, authored by 41 accounts that include financial news outlets (e.g., Argaam, 19.8% of tweets), market-analysis accounts and individual traders. Each record contains the tweet text, account, date, retweet count and like count. The median tweet length after cleaning is 15 words (mean 19.3). A reference table of 250 listed companies, with their Tadawul codes, ISINs and short names, was used for company linkage. Dates in the source file are recorded in several inconsistent formats and were not used in the analysis. Only public posts were used, and account handles are not reported in the results.

3.2 Preprocessing

Emojis, URLs, user mentions and the hash symbol were removed; punctuation and symbols were stripped; elongated characters were collapsed; and Arabic letters were normalised by unifying alef variants (ا → آ, إ, إ), converting ta marbuta to ha (ة → ه) and alef maksura to ya (ي → ي). Cleaned text was used for the TF-IDF classifiers; the emoji-stripped but otherwise original text was supplied to the LLM so that it could exploit punctuation, digits and hashtags.

3.3 Lexicon-based label construction

Following the original project, sentiment labels were generated from three Arabic term lists specific to the stock-market context: 60 positive terms (e.g., ربح, نمو, استثمار), 60 negative terms (e.g., خسارة, انهيار, تقلص)

الاصناف) and 60 neutral terms (e.g., `الاصناف` ,`الاصناف` ,`الاصناف`). A tweet receives the class with the highest count of matching terms; ties, including the zero-match case, are resolved by the VADER compound score (Hutto & Gilbert, 2014), which, because its vocabulary is English, returns a zero score for Arabic text and therefore assigns the neutral class. We refer to this procedure as the original lexicon labelling (L1). It yielded 229 positive (22.9%), 144 negative (14.4%) and 627 neutral (62.7%) tweets. Because the neutral class is also the default for non-matching tweets, it combines genuinely neutral tweets with tweets about which the lexicon has no information.

3.4 Classical machine-learning models (RQ1)

Cleaned tweets were vectorised with TF-IDF (unigrams; default settings of scikit-learn; Pedregosa et al., 2011) and classified with Multinomial Naïve Bayes, Logistic Regression (maximum 1,000 iterations), Random Forest (100 trees) and a linear-kernel SVM. Vectorisation was fitted inside each training fold to avoid leakage. Performance was estimated with stratified five-fold cross-validation (random seed 42) and is reported as mean $\pm$ standard deviation of accuracy, macro-F1 and weighted-F1. Macro-F1 is emphasised because of class imbalance.

3.5 Zero-shot LLM classification (RQ2)

We used Qwen2.5-7B-Instruct (Yang et al., 2024), an openly released instruction-tuned model with multilingual capability, loaded in 4-bit NF4 quantisation (Dettmers et al., 2023) on a single 8 GB consumer GPU (NVIDIA RTX 4070 Laptop). No fine-tuning, few-shot examples or external APIs were used. A system prompt described the role (a financial analyst for the Saudi market), the target (sentiment toward the market or the mentioned company's share price) and a three-option answer key: A = positive, B = negative, C = neutral. Rather than sampling free text, we read the next-token probabilities of the three option tokens, renormalised them and selected the maximum. This yields deterministic output, avoids parsing errors, and provides class probabilities. The full prompt is given in Appendix A.

3.6 Evaluation (RQ2–RQ3)

Agreement between the LLM and the lexicon labels was measured with accuracy, per-class precision, recall and F1, macro-F1 (with a 95% percentile bootstrap interval from 1,000 resamples of tweets) and Cohen's κ (Cohen, 1960). Because the lexicon is itself an imperfect reference, these figures quantify agreement, not correctness. To address RQ3 we constructed a second labelling (L2) with identical lexicon terms and decision rule, but with the same orthographic normalisation applied to the lexicon terms as to the tweets (so that, for example, the term `الاصناف` is matched against its normalised form `الاصناف`), and measured changes in match coverage, label distribution and LLM agreement. Finally, a random sample of LLM–lexicon disagreements was inspected qualitatively.

4. Results

4.1 Classical models reproduce lexicon labels (RQ1)

Table 1
Five-fold cross-validated performance of TF-IDF classifiers against lexicon labels (L1); mean $\pm$ SD, n = 1,000.

Model	Accuracy	Macro-F1	Weighted-F1
Multinomial Naïve Bayes	0.684 $\pm$ 0.007	0.446 $\pm$ 0.033	0.605 $\pm$ 0.017
Logistic Regression	0.812 $\pm$ 0.028	0.719 $\pm$ 0.047	0.794 $\pm$ 0.032
Random Forest	0.835 $\pm$ 0.030	0.767 $\pm$ 0.042	0.826 $\pm$ 0.031
Linear SVM	0.861 $\pm$ 0.023	0.819 $\pm$ 0.027	0.857 $\pm$ 0.024

The linear SVM performed best on every metric (accuracy 0.861; macro-F1 0.819), followed by Random Forest and Logistic Regression; Naïve Bayes lagged far behind in macro-F1 (0.446) because it collapsed minority classes into the majority neutral class (Table 1). This ranking mirrors the 70/30 hold-out result of the original project (SVM accuracy 0.88), indicating that the finding is not an artefact of a single split. Importantly, these values describe the ability of models to recover the lexicon’s decisions from word features. Because the lexicon labels are a deterministic function of the presence of specific words, a high-capacity linear model on TF-IDF features is well placed to reconstruct them.

4.2 Zero-shot LLM versus lexicon labels (RQ2)

Table 2
Agreement of zero-shot Qwen2.5-7B-Instruct with lexicon labels (L1).

Class	Precision	Recall	F1	Lexicon n	LLM n
Positive	0.300	0.904	0.450	229	691
Negative	0.486	0.868	0.623	144	257
Neutral	0.846	0.070	0.130	627	52

Lexicon label taken as reference. Overall: accuracy 0.376; macro-F1 0.401 (95% bootstrap CI 0.374–0.427); weighted-F1 0.274; Cohen’s κ = 0.192; n = 1,000.

Agreement was low (accuracy 0.376; κ = 0.19, “slight” agreement). The LLM assigned 69.1% of tweets to the positive class compared with 22.9% under the lexicon, and only 5.2% to neutral compared with 62.7% (Table 2; Fig. 1). Of the 627 lexicon-neutral tweets, 467 (74.5%) were labelled positive and 116 (18.5%) negative by the LLM, and only 44 (7.0%) were retained as neutral. In contrast, the LLM agreed well on the polar classes in the sense of recall: it recovered 90.4% of lexicon-positive and 86.8% of lexicon-negative tweets, with only 2 of 144 lexicon-negative tweets labelled neutral.

Inspection of a random sample of disagreements showed a consistent pattern. The most frequent group comprised factual corporate announcements that the lexicon labelled neutral solely because no term matched, but which most readers would interpret as favourable: cash-dividend declarations (e.g., a

distribution of SAR 1.35 per share), reported profit growth (e.g., profits of SAR 78.2 million, up 20%), and analyst notes stating that earnings exceeded expectations. The LLM labelled these positive with probability close to 1.0. A second group consisted of posts with explicit sentiment vocabulary that was missed by the lexicon, such as a trader noting a “positive reaction” of a stock at a target level. These observations are qualitative and drawn from a small sample; they were not produced by independent human annotators and should be read as hypotheses to be tested against a gold standard. They also point to a potential LLM bias: treating any corporate news about distributions or profit as positive may overstate sentiment when such events were already expected by the market.

4.3 Sensitivity of lexicon labels to preprocessing (RQ3)

The disagreement pattern in Section 4.2 prompted an audit of the labelling procedure. In the original pipeline, tweets were normalised before matching (e.g., “XXXXXXXX” becomes “XXXXXXXX” and “XXXXXXXX” becomes “XXXXXXXX”), whereas the lexicon terms were not; any lexicon term containing a hamza-bearing alef, ta marbuta or alef maksura could therefore never match. Under L1, only 424 of 1,000 tweets (42.4%) matched any lexicon term. After applying identical normalisation to the lexicon (L2), 678 tweets (67.8%) matched (Table 3).

Table 3
Effect of normalising the lexicon on coverage, label distribution and LLM agreement.

Quantity	L1 (original)	L2 (normalised lexicon)
Tweets with ≥ 1 lexicon match	424 (42.4%)	678 (67.8%)
Positive / Negative / Neutral	229 / 144 / 627	491 / 104 / 405
Agreement between L1 and L2 labels	70.0%	—
LLM accuracy against labels	0.376	0.575
LLM macro-F1 against labels	0.401	0.480
LLM Cohen’s κ against labels	0.192	0.307

Thirty per cent of labels changed. The positive class more than doubled (229 → 491) and the neutral class fell from 62.7% to 40.5%, while LLM agreement improved from 37.6% to 57.5% accuracy (κ from 0.19 to 0.31). Hence a large part of the apparent LLM–lexicon disagreement is attributable to lexicon-label defects rather than LLM error. Residual disagreement (κ = 0.31, “fair”) remains substantial, consistent with the LLM’s positive-class bias and the coarse nature of keyword counting. The classical results in Section 4.1 should be read in the same light: they were obtained on L1, and a model that scores 0.86 against a labelling that ignores 58% of tweets has learnt, for the most part, to detect the absence of matched terms.

5. Discussion

The central finding is methodological. Supervised accuracy against rule-derived labels is an unreliable indicator of sentiment-analysis quality: the SVM's 0.86 accuracy and the LLM's 0.38 accuracy against the same labels cannot be interpreted as a ranking of which system better understands investor sentiment, because the reference is neither independent nor validated. This is consistent with warnings that dictionary-based measures require validation against human coding (van Atteveldt et al., 2021) and with the theory of weak supervision (Ratner et al., 2017). The L1/L2 comparison supplies concrete evidence: a single preprocessing inconsistency changed nearly one-third of labels.

For practitioners, the LLM results suggest both promise and caution. A zero-shot 7B open model recovered the polar lexicon classes with high recall and surfaced a category of items (dividends and earnings news) that the lexicon systematically left neutral, without any training data and on commodity hardware. At the same time, its heavy positive skew and near-absence of neutral predictions indicate that unadjusted zero-shot outputs would overstate market optimism. Possible remedies, to be tested, include calibrating decision thresholds on a small labelled sample, adding few-shot exemplars of neutral announcements, separating event type from valence, and ensembling prompts or models.

For Arabic NLP in finance, the results emphasise that morphological and orthographic normalisation must be applied consistently to every resource in the pipeline, including lexicons. Lexicon expansion with morphological variants and dialectal forms, and the use of Arabic-specific financial resources, are further natural improvements.

5.1 Limitations

- No human-annotated reference. This is the principal limitation. All accuracies are agreements with imperfect lexicon labels; no claim is made about the absolute correctness of any method. A stratified annotation sheet of 210 tweets and an evaluation protocol accompany this paper to enable this validation.
- Sample size and scope. 1,000 tweets from 41 accounts, dominated by news outlets (e.g., 19.8% from a single outlet), limit generalisability; retail-investor opinion and sarcasm may be under-represented.
- Single LLM and prompt. Only one model, one prompt and no few-shot condition were evaluated; results may differ for larger models, Arabic-specialised models or alternative prompts.
- Dependence of the classical results on L1. Cross-validation scores were computed on labels shown to be defective; they are reported to characterise the original pipeline rather than as recommended performance estimates.
- No market linkage. The study evaluates sentiment classification, not the predictive relationship between sentiment and returns, which requires aligned price data and appropriate econometric design.
- Fine-tuned transformers excluded. The original project fine-tuned an Arabic BERT without a held-out evaluation; it is therefore not reported here.

6. Conclusion and Future Work

We compared lexicon-labelled classical classifiers with a zero-shot open-source LLM for sentiment on Saudi stock tweets. Classical models, especially a linear SVM, reproduced lexicon labels with high accuracy (0.861), but the zero-shot LLM agreed only weakly (0.376) with the same labels. A diagnostic analysis showed that the labels themselves were affected by a normalisation mismatch that left 57.6% of tweets unmatched; correcting it altered 30% of labels and increased LLM agreement to 57.5%. These findings caution against using rule-derived labels as a validation target and show the value of auditing label construction.

Future work will (i) complete the human annotation of a stratified reference set with at least two annotators and report inter-annotator agreement; (ii) re-evaluate all approaches, including fine-tuned Arabic transformers and larger LLMs with few-shot prompts, against that reference; (iii) calibrate LLM outputs to mitigate positive-class bias; and (iv) link validated company-level sentiment to Tadawul returns and trading volumes using event-study or time-series designs.

Declarations

Data and code availability.

The analysed posts are public; the processing scripts, prompt, lexicon-normalisation audit and annotation sheet are provided with this manuscript. Redistribution of tweet text should respect the platform's terms of service.

Use of AI tools. A large language model (Qwen2.5-7B-Instruct) is a subject of this study, as described in Section 3.5. An AI assistant (Claude, Anthropic) was used to support code development, analysis automation and drafting; the authors are responsible for verifying all content, results and references.

Ethics. The study analysed publicly available posts and reports aggregate results only; account handles are not disclosed.

Conflict of interest and funding. The authors declare no competing interests. Funding: [to be completed].

References

1. Abdul-Mageed M, Elmadany A, Nagoudi EMB (2021) ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of ACL-IJCNLP 2021* (pp. 7088–7105). Association for Computational Linguistics
2. Abu Farha I, Magdy W (2021) Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 21–31). Association for Computational Linguistics

3. Antoun W, Baly F, Hajj H (2020) AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4)* (pp. 9–15). European Language Resources Association
4. Araci D (2019) FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.
5. Behura JP, Pande SD, Ramesh JVN (2023) Stock price prediction using multi-layered sequential LSTM. *EAI Endorsed Trans Scalable Inform Syst* 11(4). Article 4585. <https://doi.org/10.4108/eetsis.4585>
6. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2(1):1–8
7. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
8. Brown TB, Mann B, Ryder N, Subbiah M et al (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst* 33:1877–1901
9. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
10. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023) QLoRA: Efficient finetuning of quantized LLMs. In *Adv Neural Inf Process Syst*, 36
11. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186)
12. Du K, Xing F, Mao R, Cambria E (2024) Financial sentiment analysis: Techniques and applications. *ACM-CSUR* 56(9). Article 220. <https://doi.org/10.1145/3649451>
13. Gu WJ, Zhong YH, Li SZ, Wei CS, Dong LT, Wang ZY, Yan C (2024) Predicting stock prices with FinBERT-LSTM: Integrating news sentiment analysis. In *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing* (pp. 67–72). <https://doi.org/10.1145/3694860.3694870>
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
15. Hutto CJ, Gilbert E (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM 2014* (pp. 216–225)
16. Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98* (pp. 137–142). Springer
17. Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J Finance* 66(1):35–65
18. Mehta P, Pandya S, Kotecha K (2021) Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Comput Sci* 7:e476. <https://doi.org/10.7717/peerj-cs.476>
19. Northcutt CG, Jiang L, Chuang IL (2021) Confident learning: Estimating uncertainty in dataset labels. *J Artif Intell Res* 70:1373–1411
20. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830

21. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C (2017) Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3), 269–282
22. Safaya A, Abdullatif M, Yuret D (2020) KUISAIL at SemEval-2020 Task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 2054–2059)
23. Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. *J Finance* 62(3):1139–1168
24. van Atteveldt W, van der Velden MACG, Boukes M (2021) The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods Measures* 15(2):121–140
25. Wu S, Liu Y, Zou Z, Weng T-H (2022) S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. *Connection Sci* 34(1):44–62
26. Yang A, Yang B, Zhang B et al (2024) Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*
27. Zhang W, Deng Y, Liu B, Pan SJ, Bing L (2023) Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Figures

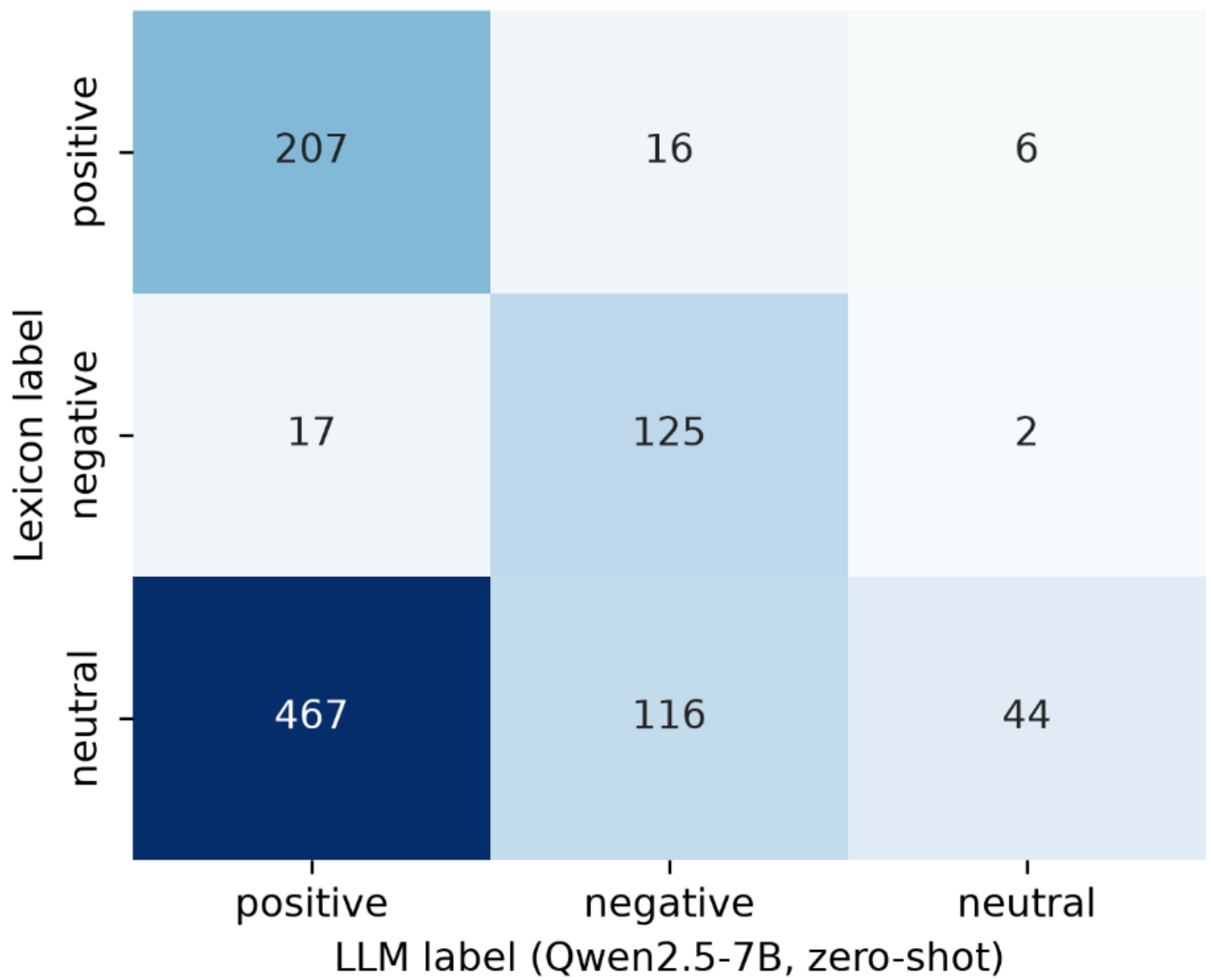


Figure 1

Confusion matrix of lexicon labels (rows) against zero-shot LLM labels (columns).

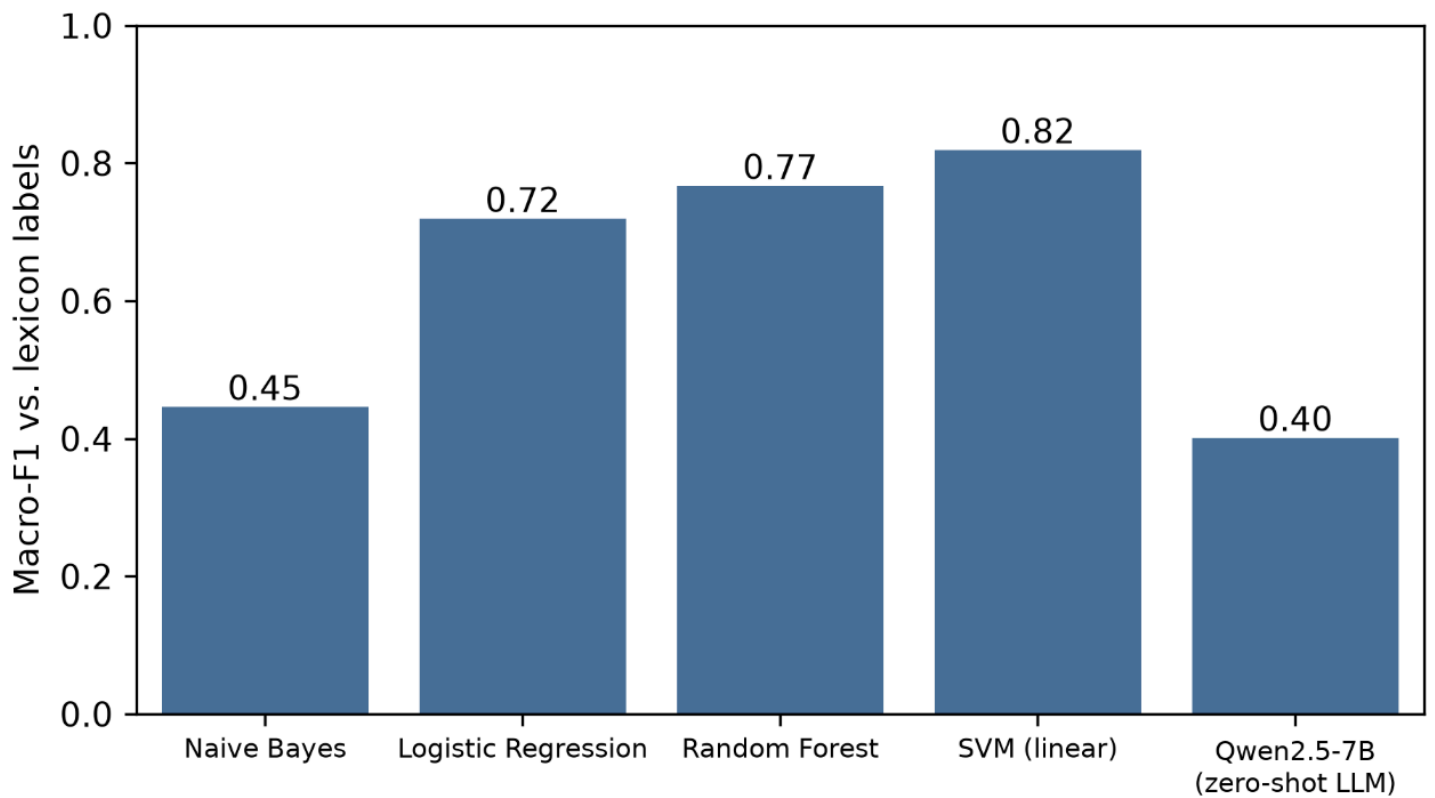


Figure 2

Macro-F1 of each approach against lexicon labels (L1). Classical bars are five-fold CV means; the LLM bar is agreement without any training on these labels.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)