The role of large language models in generating sentiment signals: A comparative and effectiveness-based study

Domenica Fioredistella Iezzi[a], Roberto Monte[a]
[a] Department of Enterprise Engineering Mario Lucertini, University of Rome Tor Vergata, Italy
[b] Department of Civil Engineering and Computer Science Engineering, University of Rome Tor Vergata, Italy

1. Introduction

Large language models (LLMs) have transformed natural language processing (NLP), achieving state-of-the-art performance across many linguistic tasks, including sentiment analysis. Unlike traditional machine-learning approaches that depend on manual feature engineering and task-specific datasets, LLMs are pretrained on vast textual corpora and capture nuanced semantic and contextual relationships across domains. Examples include GPT-based models such as ChatGPT (OpenAI, 2023), Meta's LLaMA (Touvron et al., 2023), and transformer-based encoders such as BERT (Devlin et al., 2019) and its tweet-oriented variant CardiffNLP (Barbieri et al., 2020). Their broad pretraining and contextual understanding make LLMs particularly well suited to extract sentiment signals from informal, noisy, and unstructured user-generated content on social media.

In this study, we consider the monthly sales volume time series for the Toyota Camry in the U.S. from June 1, 2009, to December 31, 2022, and for the FIAT 500L from August 1, 2012, to December 31, 2018. We also collected corresponding Twitter (now X) time series, gathering approximately 163,000 English tweets about the Toyota Camry (1,000 per month) and 20,387 tweets about the FIAT 500L.

We aim to evaluate whether the sentiment generated by ChatGPT, LLaMA, and BERT improves sales forecasting performance and compare their relative predictive contribution (Iezzi and Monte, 2025). Beyond forecasting accuracy, we analyze how sentiment signals differ across brands, time periods, and contextual conditions, assessing each model's ability to capture sentiment dynamics in real-world, domain-specific data. These findings might contribute to understanding the comparative strengths and limitations of generative and transformer-based models in sentiment-oriented forecasting and their implications for social media-based market analysis.

The rest of this paper is structured as follows: Section 2 offers an overview of LLMs and their significance to our study. Section 3 presents the key empirical findings related to the impact of sentiment proxies on sales forecasts accuracy. Finally, Section 4 concludes the paper and explores potential directions for future research and development.

2. Large language models

The framework for LLMs encompasses both the technical infrastructure and methodological principles that govern their development, deployment, and evaluation. At its foundation lies the Transformer architecture (Vaswani et al., 2017), which introduced self-attention as a scalable alternative to recurrent and convolutional networks. Within this framework, models are typically pre-trained on massive corpora, capturing linguistic and world knowledge, and subsequently adapted to downstream tasks via fine-tuning or prompting.

A distinction in this framework can be drawn between encoder-based models, such as BERT and CardiffNLP, and decoder-based or generative LLMs, such as GPT-3 and GPT-4 (Radford et al., 2018; Brown et al., 2020; OpenAI, 2023), LLaMA, and Mistral (Jiang et al., 2023). BERT and CardiffNLP, designed as a bidirectional encoder, are optimized for representation learning. By encoding text into meaningful vectors, they excel at tasks such as classification, sentiment analysis, and question answering when paired with task-specific layers. However, they are not inherently generative and cannot produce extended free-form text.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Domenica Fioredistella Iezzi, Roberto Monte, *The role of large language models in generating sentiment signals: A comparative and effectiveness-based study*, pp. 231-236, © 2026 Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1062-1.41

In contrast, ChatGPT, based on the GPT family, employs autoregressive decoding to generate coherent, contextually appropriate text and is further refined through reinforcement learning with human feedback (RLHF), which enhances alignment with human expectations. LLaMA emphasizes efficiency and accessibility by providing competitive performance with smaller parameter counts compared to proprietary LLMs. Mistral, a more recent contribution, advances this framework by incorporating efficient training strategies, including mixture-of-expert techniques, to achieve high performance with reduced computational cost.

From a software perspective, frameworks such as Hugging face transformers (Wolf et al., 2020), PyTorch Lightning, and TensorFlow provide standardized APIs and pre-trained checkpoints, enabling reproducibility and lowering the barrier to entry. Yet, beyond the technical layer, the framework also includes ethical and methodological components: evaluation protocols, bias assessment, and responsible AI guidelines (Bender et al., 2021; Bommasani et al., 2021). This broader perspective highlights that the framework is not limited to algorithms but represents a socio-technical system.

Sentiment analysis, often referred to as opinion mining, has long been a central task in natural language processing (NLP). Traditional approaches relied on lexicon-based methods or machine learning classifiers trained on handcrafted features, which performed adequately on short and well-structured texts but struggled with linguistic nuances such as sarcasm, irony, and domain-specific vocabulary (Liu, 2012). The introduction of deep learning methods, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), improved performance by enabling the automatic extraction of features from raw text. However, these models still suffered from limitations in capturing long-range dependencies and contextual meaning.

As outlined above, the advent of Transformer-based models provided contextual embeddings that significantly advanced sentiment classification. In contrast, generative LLMs extend the capabilities to zero-shot and few-shot sentiment analysis, allowing classification without task-specific fine-tuning. This feature reduces the dependency on annotated datasets.

Open-source alternatives such as LLaMA and Mistral further democratize sentiment analysis, making it attractive for researchers and institutions with limited resources. Mistral, not exploited in this paper, introduces mixture-of-experts architectures to optimize efficiency and inference speed, which is particularly relevant for real-time sentiment analysis of social media streams. Meanwhile, ChatGPT, a conversational deployment of GPT models, illustrates how LLMs can be used interactively for qualitative sentiment exploration, generating not only labels (positive, negative, neutral) but also explanations and contextual justifications.

Beyond classification accuracy, LLMs enable a richer perspective on sentiment analysis. They can capture fine-grained emotions, stance, and pragmatic aspects of communication that go beyond polarity scores (Zhang et al., 2022). At the same time, challenges remain: LLMs risk amplifying biases present in their training data, raising concerns about fairness in applications such as political opinion mining or customer feedback analysis (Bender et al., 2021). Therefore, ethical use remains a critical issue for the sustainable adoption of these technologies. Moreover, the computational and environmental costs of deploying large-scale sentiment models are nontrivial, motivating research on efficient fine-tuning methods such as LoRA and quantization (Hu et al., 2022). Table 1, based on Toyota Tweets corpus, constitutes an example of the different approaches to sentiment classification across above-mentioned tools.

In the present study, we employed multiple LLMs, specifically CardiffNLP, LLaMA, and ChatGPT-4. These models were selected to provide a comparative perspective across different architectures and training paradigms, thereby allowing us to evaluate their relative performance in capturing sentiment signals and their subsequent impact on forecasting accuracy.

Table 1. Comparative sentiment scoring of Toyota Camry-related tweets using ChatGPT-4, LLaMA, CardiffNLP

TWEET	ChatGPT-4	LLaMA	CardiffNLP
My Toyota has been incredibly reliable — smooth ride, great fuel efficiency, and zero issues so far. Couldn't be happier!	0.95	0.90	0.98
The Toyota gets me where I need to go. It's practical and comfortable, nothing particularly exciting but it does the job.	0.10	−0.05	−0.35
Really disappointed with my Toyota — constant issues and poor customer support. Definitely not what I expected.	−0.91	−0.98	−0.93

3. Impact of sentiment proxies on sales forecast accuracy

Models using ChatGPT-4, along with LLaMA and CardiffNLP, significantly reduce the root mean square error (RMSE) and mean absolute error (MAE) compared to benchmark specifications. This suggests that LLMs are not just introducing noise; they are effectively capturing key indicators of consumer sentiment and market attention.

These results indicate a substantial improvement over earlier sentiment indicators, which offered only modest gains in forecasting performance. Nevertheless, applying LLaMA to Italian tweets reveals several systematic anomalies that undermine the reliability of its sentiment scoring. For instance, the model often over-assigns negative sentiment, even in cases where the content is purely informational or neutral.

A notable example is the multiple retweets of a factual announcement regarding Fiat's decision to use the Monfalcone terminal for 500L production, which are consistently assigned negative scores (ranging from -0.25 to -0.83). This occurs despite the absence of evaluative language, indicating a bias towards interpreting industrial or economic news as detrimental.

Furthermore, the model often confuses problem-related terminology with sentiment intensity. For instance, tweets mentioning "problemi in Serbia" (problems in Serbia) consistently receive highly negative scores (-1), even when the message explicitly states that production is not at risk. In this case, the model fails to grasp negation and contextual nuances, exaggerating the negative sentiment solely based on the term "problemi" (problems). Additionally, there is evidence of semantic drift in unrelated contexts. For example, a reference to the Fiat CEO switching to electronic cigarettes is inaccurately rated as negative (-0.75), despite not expressing any evaluative stance toward the brand or product. This shows a challenge in differentiating sentiment related to the entity (Fiat cars) from general background statements.

The model also exhibits inconsistencies when handling reviews or evaluative content. A clearly positive phrase, such as "La prova su strada della Fiat 500L di Quattroruote" (Quattroruote's road test of the Fiat 500L), receives only a moderate positive score (+0.5), which underestimates the positive connotation typically associated with road tests in automotive reviews. These anomalies indicate that the model struggles with Italian texts due to limited linguistic adaptation and sensitivity to the domain.

This yields spurious negativity, inadequate handling of negation, and a poor recognition of contextual positivity. Such limitations make the model less robust than specialized multilingual sentiment classifiers, like CardiffNLP, particularly when applied to short, domain-specific social media texts.

Figure 1 illustrates the temporal dynamics of negative sentiment in Italian tweets about the Fiat 500L and its relationship with sales volume from August 2012 to 2022. The first two panels compare the sentiment classification results obtained using CardiffNLP and ChatGPT-4, while the third panel presents the monthly sales volumes. Both sentiment classifiers show generally consistent trends, capturing peaks of negativity that coincide with significant moments in the Fiat 500L's production and reception in the Italian market.

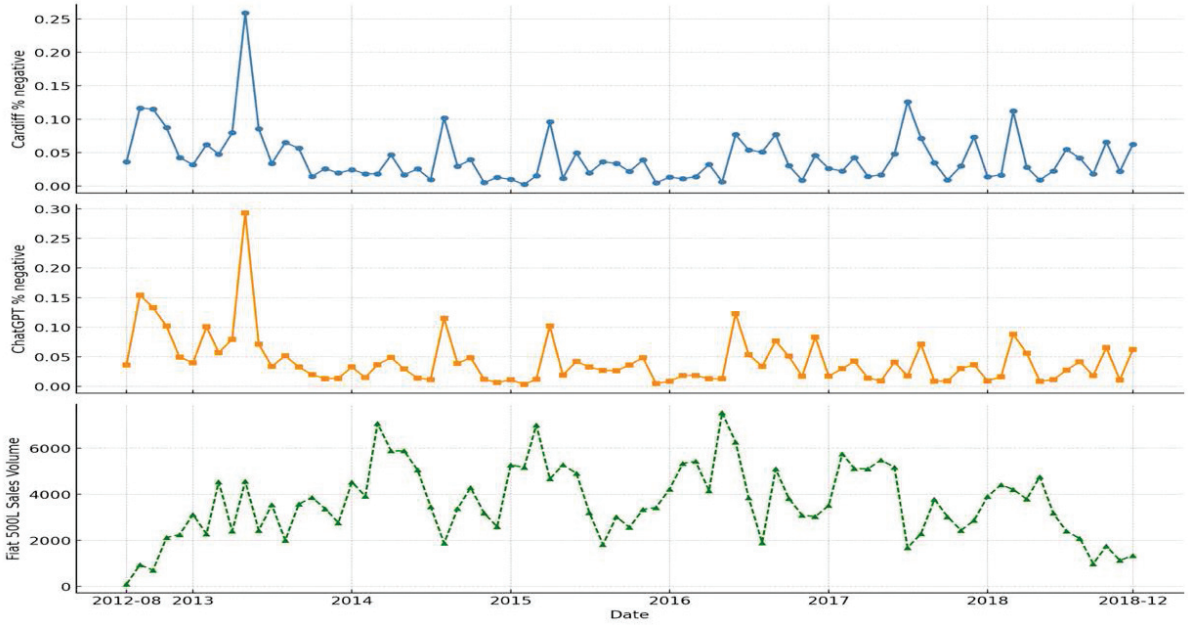


Figure 1. Negative sentiment in Fiat 500L tweets (CardiffNLP vs. ChatGPT-4) and monthly sales volume, 2012–2022

However, ChatGPT-4 tends to produce more variability and sharper spikes in sentiment, whereas CardiffNLP results in smoother series. This difference may be attributed to variations in model architecture and calibration when applied to domain-specific datasets.

The correlation between the negative sentiment series generated by ChatGPT-4 and CardiffNLP is 0.89, indicating a strong alignment over time. When sentiment trends are compared with sales volumes, an inverse relationship emerges periods of increased negativity in online discussions often correspond with declines in Fiat 500L sales. Although correlation does not imply causation, the parallelism between these dynamics suggests that consumer sentiment extracted from social media provides valuable insights into market performance.

Table 2. Forecasting accuracy of ETS and MARSS models with and without sentiment variable for the FIAT 500L

Models	logLik	AIC	BIC	AICC	RMSE	MAE	MAPE	SMAPE%	MASE
ETS-AAA (y)	-606.879	1247.759	1285.738	1250.759	867.621	706.553	48.772	18.469	0.698
MARSS (y) No sent.	-895.996	1815.991	1851.117	1818.487	748.246	568.791	39.116	16.495	0.562
MARSS (y) ChatGPT-4	-549.541	1123.082	1149.891	1128.653	724.569	584.011	40.155	14.365	0.577
MARSS (z) ChatGPT-4	-549.541	1123.082	1149.891	1128.653	0.029	0.021	85.173	35.240	0.559
MARSS (y) CardiffNLP	-508.202	1040.405	1067.214	1045.976	707.621	542.767	33.094	12.879	0.536
MARSS (z) CardiffNLP	-508.202	1040.405	1067.214	1045.976	0.024	0.021	82.671	29.984	0.663

Tables 2 and 3 present the forecasting performance of ETS and MARSS models, both with and without the inclusion of sentiment variables derived from various sources (ChatGPT-4, LLaMA, CardiffNLP). The baseline ETS models (ETS-AAA, ETS-ANA) present the poorest performance, exhibiting significantly higher RMSE, MAE, and MASE values compared to the MARSS specifications. Similarly, the MARSS model that does not include sentiment variables shows worse predictive accuracy, eventually highlighting the added value of incorporating sentiment features. When sentiment is included, notable differences emerge among the various sources.

Table 3. Forecasting accuracy of ETS and MARSS models with and without sentiment variable for the Toyota Camry

Models	logLik	AIC	BIC	AICC	RMSE	MAE	MAPE	SMAPE%	MASE
ETS-ANA (y)	-1591.434	3212.869	3257.623	3216.561	9440.055	8610.602	39.914	15.902	1.669
ETS-AAA (y)	-1592.126	3218.252	3268.973	3223.033	10809.330	10241.120	46.714	18.321	1.985
MARSS (y) No sent.	-1567.278	3154.555	3184.319	3156.185	7291.802	6087.463	28.754	11.831	1.180
MARSS (y) ChatGPT-4	-1506.265	3036.531	3072.334	3038.876	3321.115	2816.744	12.627	6.231	0.546
MARSS (z) ChatGPT-4	-1506.265	3036.531	3072.334	3038.876	0.217	0.114	32.600	12.809	0.780
MARSS (y) LLaMA	-1844.953	3713.907	3749.710	3716.253	4957.907	4042.257	19.105	8.465	0.784
MARSS (z) LLaMA	-1844.953	3713.907	3749.710	3716.253	0.161	0.104	50.148	20.349	0.679
MARSS (y) CardiffNLP	-1825.593	3675.186	3710.989	3677.532	3667.565	2735.730	12.197	5.741	0.530
MARSS (z) CardiffNLP	-1825.593	3675.186	3710.989	3677.532	0.148	0.060	87.304	5.741	0.453

Models based on CardiffNLP consistently achieve the highest overall accuracy, with lower RMSE, MAE, MAPE, and SMAPE values, and MASE scores well below 1, indicating superior predictive power compared to the naïve benchmark.

The sentiment models from ChatGPT-4 perform slightly worse than those based on CardiffNLP but still significantly outperform the ETS and MARSS models without sentiment. In contrast, the LLaMA model exhibits somewhat inflated error metrics and inconsistencies across evaluation indicators. When comparing the two MARSS specifications, models based on the direct response variable (y) show greater stability. On the other hand, the alternative specification (z) occasionally yields unrealistically low RMSE and MAE values, along with disproportionately high percentage errors (MAPE, SMAPE), suggesting potential numerical instability or overfitting.

4. Conclusions and future developments

This study provides evidence on the effectiveness of incorporating sentiment proxies generated by LLMs into forecasting frameworks for sales data. By comparing CardiffNLP, LLaMA, and ChatGPT (versions 4 and 5), we highlighted how sentiment signals derived from social media can provide significant improvements in predictive accuracy compared to traditional benchmark models.

Among the tested approaches, CardiffNLP generally achieved the most stable and robust results, while ChatGPT showed substantial improvements yet with higher variability.

However, LLaMA revealed important limitations when applied to Italian tweets, including systematic biases and difficulties in contextual interpretation. These findings suggest that the choice of sentiment model is not neutral but has a direct impact on forecasting performance.

Future research may extend this analysis to additional domains and languages, refining the calibration of sentiment signals through domain-specific adaptation, and integrating multi-modal data sources (e.g., images, audio, video) to capture richer dimensions of consumer perception. Moreover, improving interpretability and ensuring robustness across cultural and linguistic contexts remain critical challenges for the sustainable adoption of sentiment-enhanced forecasting models in both academic and applied settings.

References

Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 120–128). Association for Computational Linguistics.

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://doi.org/10.48550/arXiv.2108.07258.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems,* **33**, pp. 1877–1901.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Iezzi, D.F., Monte, R. (2025). E-word of mouth in sales volume forecasting: Toyota Camry case study. *Big Data Research*, 41: 100542. https://doi.org/10.1016/j.bdr.2025.100542.

Jiang, A., Tay, Y., Liu, Q., Xu, Y., Zheng, B., and others. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*. https://doi.org/10.48550/arXiv.2310.06825.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), pp. 1–167.

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training, *OpenAI*. https://doi.org/10.48550/arXiv.1801.06146.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. Curran Associates. https://doi.org/10.48550/arXiv.1706.03762.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Zhang, T., Rao, Y., Xie, Y., Liu, Y. (2022). Sentiment analysis using pre-trained transformer models. *Journal of Artificial Intelligence Research,* 73*,* pp. 123–145.