*Journal of Wine Economics* (2025), **20**, 371–384
doi:10.1017/jwe.2025.10091

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# Comparison of language models for wine sentiment analysis

Chenyu Yang and Jing Cao

Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, USA
**Corresponding author:** Jing Cao Email: jcao@smu.edu

**Abstract**

This study presents a comparative evaluation of sentiment analysis models applied to a large corpus of expert wine reviews from Wine Spectator, with the goal of classifying reviews into binary sentiment categories based on expert ratings. We assess six models: logistic regression, XGBoost, LSTM, BERT, the interpretable Attention-based Multiple Instance Classification (AMIC) model, and the generative language model LLAMA 3.1, highlighting their differences in accuracy, interpretability, and computational efficiency. While LLAMA 3.1 achieves the highest accuracy, its marginal improvement over AMIC and BERT comes at a significantly higher computational cost. Notably, AMIC matches the performance of pretrained large language models while offering superior interpretability, making it particularly effective for domain-specific tasks such as wine sentiment analysis. Through qualitative analysis of sentiment-bearing words, we demonstrate AMIC's ability to uncover nuanced, context-dependent language patterns unique to wine reviews. These findings challenge the assumption of generative models' universal superiority and underscore the importance of aligning model selection with domain-specific requirements, especially in applications where transparency and linguistic nuance are critical.

**Keywords:** interpretability; language model; sentiment analysis; wine review

**JEL classifications** C45; C80; D83

## I. Introduction

Sentiment analysis (SA) is a foundational task in natural language processing (NLP) that seeks to identify and quantify subjective opinions expressed in text. Its applications span a wide range of domains, from consumer product reviews and political discourse to financial forecasting and public health monitoring (Sharma et al., 2025). In recent years, the proliferation of user-generated content and the advancement of deep learning techniques have significantly expanded the scope and sophistication of SA methodologies (Sánchez-Rada and Iglesias, 2019).

© The Author(s), 2025. Published by Cambridge University Press on behalf of American Association of Wine Economists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

One particularly rich and nuanced domain for SA is wine reviews. These reviews, often written by experts and enthusiasts, blend sensory descriptions with evaluative language, making them ideal for studying how sentiment is conveyed through specialized vocabulary. Unlike general product reviews, wine reviews frequently employ metaphor, domain-specific jargon, and subtle tonal shifts, posing unique challenges for traditional SA models (Yang and Cao, 2025).

Early SA approaches relied heavily on rule-based systems and statistical models. Rule-based SA methods typically depend on sentiment lexicons, such as predefined dictionaries that assign sentiment scores to individual words, to determine the overall sentiment of a text. However, these approaches often struggle to generalize effectively, as they lack the ability to capture the nuances of language, contextual meaning, and domain-specific knowledge (Berka, 2020; Levallois, 2025). Statistical SA methods employ a rigorous probability-based approach to modeling text data, such as logistic regression (Tyagi and Sharma, 2018) and Naive Bayes (Das and Chen, 2007). They often use bag-of-words or term frequency–inverse document frequency (TF-IDF) representations (Sezerer and Tekir, 2021; Spärck Jones, 1972). While interpretable and computationally efficient, these models struggled to capture contextual nuances and domain-specific sentiment expressions. The advent of neural network models, particularly those incorporating recurrent architectures like the long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) and attention-based transformers like the bidirectional encoder representations from transformer (BERT) model (Devlin et al., 2019), marked a significant leap in performance by enabling models to learn contextual embeddings and long-range dependencies in text.

However, the rise of large generative language models (LLMs), such as LLAMA 3.1 (Grattafiori et al., 2024), has introduced a new paradigm in NLP. These models are advanced artificial intelligence systems designed to understand and generate human-like text. They are pretrained on massive corpora and capable of zero-shot and few-shot learning, promise versatility and state-of-the-art performance across a wide array of tasks (Ahmed and Devanbu, 2022). Yet, their application to SA, especially in specialized domains like wine reviews, remains underexplored. While LLMs can generate fluent and plausible outputs, their interpretability, domain adaptability, and computational efficiency are often lacking.

In this paper, we critically examine the performance of LLMs in the context of wine review SA. Building on our prior work that introduced the interpretable AMIC model (Yang and Cao, 2025), we compare the effectiveness of LLAMA 3.1 against a suite of traditional and neural network models, including logistic regression, XGBoost, LSTM, BERT, and AMIC. Our findings challenge the assumption that generative models are universally superior, highlighting the trade-offs between accuracy, interpretability, and resource demands.

Our study is inspired by previous studies on sentiment classification using a large corpus of 141,409 wine reviews sourced from Wine Spectator, spanning the years 2005 to 2016 (Katumullage et al., 2022; Yang et al., 2022). Each year, the magazine's editors conducted blind tastings of over 15,000 wines, providing detailed tasting notes, numeric ratings on a 100-point scale, and recommendations. Most wines received scores between 80 and 100. Specifically, Katumullage et al. (2022) applied three neural network models – CNN, BiLSTM, and BERT – to classify wine sentiment, labeling

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

reviews with scores of 90 or above as *positive* and those below as *negative*. Their results demonstrated that these deep learning models using wine reviews significantly outperformed logistic regression models that relied solely on numeric variables like price and vintage.

By focusing on wine reviews, a domain where sentiment is intricately tied to context and vocabulary, we aim to provide a nuanced evaluation of model performance and suitability. Our results underscore the continued relevance of task-specific, interpretable models and call into question the uncritical adoption of LLMs for sentiment classification tasks. The remainder of this article is structured as follows. Section 2 provides an overview of the six SA models evaluated in this study. Section 3 presents a comparative analysis of these models on wine review sentiment classification, focusing on accuracy, interpretability, and computational efficiency. Finally, Section 5 concludes with a discussion of the implications and future directions.

## II. Methodology

To evaluate the effectiveness of various SA models in the context of wine reviews, we conducted a comparative study using the curated dataset from Wine Spectator. Specifically, we compared six modeling approaches that span traditional statistical methods, deep learning architectures, and large-scale generative language models. These models differ in terms of architecture, interpretability, computational cost, and reliance on pretraining. Below, we describe each method in detail.

### *a. Logistic regression with TF-IDF*

Logistic regression is a widely used linear classification algorithm that models the probability of a binary outcome based on a linear combination of input features. In the SA context, it serves as a transparent and computationally efficient baseline for classifying text data. To apply logistic regression to wine reviews, we first transformed the raw textual data into a structured numerical format using the TF-IDF representation (Sezerer and Tekir, 2021; Spärck Jones, 1972). TF-IDF is a statistical measure that reflects how important a word is to a document in a collection. It balances two components:

- Term Frequency (TF): the number of times a word appears in a document, capturing its local importance.
- Inverse Document Frequency (IDF): a measure that down-weights words that appear frequently across many documents, reducing the influence of common but uninformative terms (e.g., "the," "and," and "wine").

We implemented this transformation using the *TfidfVectorizer* from the *scikit-learn* library in Python. To improve model generalization and reduce noise, we configured the vectorizer with:

- min_df = 5: to exclude words that appear in fewer than five documents.
- max_features = 10,000: to cap the vocabulary size and limit dimensionality.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

The resulting TF-IDF matrix is a high-dimensional, sparse representation of the corpus, where each row corresponds to a wine review and each column to a unique term in the vocabulary. This matrix was then split into training and testing sets using a 90/10 ratio.

The logistic regression model was trained on the TF-IDF features to learn the relationship between word usage patterns and the binary sentiment labels (*positive* vs. *negative*). The model outputs a probability score for each review, where the threshold is set at 0.5 to assign a final class label. Despite its simplicity, this approach offers several advantages. The first advantage is the interpretability where the learned coefficients directly indicate the contribution of each word to the sentiment prediction. The second is speed which means its training and inference are fast, even on large datasets. The third is baseline performance, which provides a strong benchmark against more complex models. While logistic regression lacks the ability to capture word order or contextual dependencies, its transparency and ease of implementation make it a valuable tool in early-stage model development and comparative studies.

### *b. XGBoost (Extreme Gradient Boosting)*

Extreme Gradient Boosting (XGBoost) is a powerful ensemble learning algorithm based on gradient-boosted decision trees (Chen and Guestrin, 2016). It is particularly well-suited for structured and high-dimensional data, offering a balance between predictive performance and computational efficiency. In the context of wine review SA, XGBoost provides a robust alternative to both linear models and deep learning approaches, especially when paired with effective feature engineering.

In our implementation, we used the *XGBClassifier* class from the *xgboost.sklearn* module in Python. The input text data was first transformed into numerical feature vectors using the TF-IDF representation, as described in the logistic regression section. This transformation captures the relative importance of words in the corpus while preserving sparsity, which is ideal for tree-based models. The XGBoost classifier was configured with the following key settings:

- Objective: binary:logistic, which models the probability of a binary outcome using logistic regression at the leaf nodes.
- Evaluation Metric: logloss, a standard loss function for binary classification that penalizes incorrect predictions based on their confidence.

To ensure optimal performance, we conducted a comprehensive grid search over a range of hyperparameters using 5-fold cross-validation. We applied early stopping during training by monitoring the model's performance on a held-out validation set. If the validation loss did not improve for 10 consecutive rounds, training was halted to prevent overfitting and reduce unnecessary computation. Once the best hyperparameter configuration was identified, the final model was retrained on the full training set and evaluated on the test set. XGBoost's ability to model non-linear relationships and feature interactions, along with its resilience to multicollinearity, made it a strong candidate in our comparative framework.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

Although XGBoost is less interpretable than linear models like logistic regression, as it relies on an ensemble of decision trees, it remains more transparent than deep neural networks. Feature importance scores can be extracted to provide insights into which terms most strongly influence sentiment predictions. Overall, XGBoost can achieve competitive accuracy while maintaining relatively low computational cost, making it a practical and effective choice for text classification tasks in specialized domains like wine reviews.

### c. LSTM (Long short-term memory)

LSTM networks are a specialized type of Recurrent Neural Networks (RNNs) designed to model sequential data with long-range dependencies. Traditional RNNs often struggle with vanishing gradients and limited memory capacity, which hinders their ability to learn from long sequences (Alpay et al., 2016). LSTMs address these issues by incorporating gated memory cells that regulate the flow of information, enabling the network to retain or discard data as needed over extended sequences. This makes LSTMs particularly effective for NLP tasks, where understanding context and word order is essential.

For our sentiment classification task, each wine review was first tokenized into a sequence of word indices using a standard tokenizer with a fixed vocabulary size of 10,000. These sequences were then padded to a uniform length to facilitate efficient batch processing. The tokenized input was mapped to dense vector representations via an embedding layer, which was initialized with pre-trained *GloVe* (Global Vectors for Word Representation) embeddings (Pennington et al., 2014) using 100-dimensional vectors. Words not found in the pre-trained vocabulary were initialized with random vectors to be learned during training. The resulting embedded sequences were passed through a single LSTM layer with 100 hidden units. To reduce the risk of overfitting, this layer was configured with a dropout rate of 0.5. The final hidden state of the LSTM, representing the encoded information from the entire sequence, was then fed into a fully connected feedforward layer with a sigmoid activation function to perform binary classification.

The model was implemented using the Keras API with a TensorFlow backend. Training was conducted using the *Adam* optimizer with a learning rate of 0.001, and binary cross-entropy was used as the loss function. To prevent overfitting, early stopping was applied based on validation loss, with a patience of 3 epochs. This architecture enables the LSTM to effectively capture the temporal dynamics and contextual relationships within the text, providing robustness to both syntactic and semantic variations in the wine reviews.

### d. BERT (Bidirectional encoder representations from transformers)

BERT is a transformer-based language model that has significantly advanced performance across a wide range of NLP tasks. Unlike unidirectional models, which process text in a single direction (either left-to-right or right-to-left), BERT reads input sequences bidirectionally, allowing it to learn deep contextual representations of words based on both their preceding and following context. This bidirectional nature enables

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

BERT to better understand the nuances of language. In our study, BERT is one of two models that utilize transfer learning, having been pre-trained on large-scale corpora using two self-supervised objectives: masked language modeling and next sentence prediction (He et al., 2022). These objectives help BERT develop a general-purpose understanding of language structure and semantics.

To adapt BERT for our binary sentiment classification task, we employed a fine-tuning approach using the pre-trained bert-base-uncased model from the *Hugging Face Transformers* library in Python. Each wine review was first tokenized using BERT's WordPiece tokenizer, with a maximum sequence length of 512 tokens. As required by the model architecture, special tokens [CLS] (classification token) and [SEP] (separator token) were added to the input sequence. Fine-tuning was performed by appending a fully connected dense layer on top of BERT's final hidden state corresponding to the [CLS] token, which is designed to capture a holistic representation of the entire input sequence. This dense layer projected the [CLS] embedding to a single scalar output, followed by a sigmoid activation function to produce a probability score. A threshold of 0.5 was applied to convert this score into a binary sentiment prediction, in line with standard classification practices.

The model was trained using the *AdamW* optimizer with a learning rate of $2 \times 10^{-5}$, and binary cross-entropy was used as the loss function. Training was conducted for up to 4 epochs, with early stopping based on validation loss to prevent overfitting. A batch size of 32 was used during fine-tuning, and gradient clipping was applied to enhance training stability and prevent exploding gradients. By fine-tuning all parameters of the pre-trained BERT model on our domain-specific dataset, we enabled it to adapt its general language understanding to the specific task of wine review sentiment classification, resulting in high classification accuracy.

### e. AMIC (Attention-based multiple instance classification)

The Attention-based Multiple Instance Classification (AMIC) model is a custom-designed neural architecture that aims to deliver both high accuracy and interpretability in SA. Unlike many deep learning models that act as a black-box, AMIC is built to show which words in a text contribute to its sentiment classification and how.

The core idea behind AMIC is that not all words in a review carry equal weight when it comes to sentiment. Some words—like *elegant, harsh*, or *overripe*—clearly express positive or negative opinions, while others—such as *the* or *and*—are neutral. AMIC is designed to identify these sentiment-bearing words and assign them context-aware sentiment scores. The model processes each review at the word level using pre-trained 300-dimensional *GloVe* embeddings (Pennington et al., 2014) to capture semantic meaning. It consists of two modeling blocks:

1. Sentiment Word Identification: This block determines which words in the review are likely to carry sentiment. It uses a self-attention mechanism (Cheng et al., 2016) to evaluate each word in context and assigns a probability indicating whether the word contributes to the overall sentiment.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

2. Sentiment Scoring: This block estimates how positive or negative each identified sentiment word is. These scores are not fixed but depend on the word's context within the review.

The final sentiment score for a review is computed by aggregating the sentiment scores of the identified sentiment words. Specifically, the model computes a weighted sum of word-level sentiment scores, where weights are determined by the sentiment word indicators. This aggregate score is passed through a sigmoid function to produce a probability of the review being classified as *positive*. The model is trained using binary cross-entropy loss. Training is performed using gradient descent with backpropagation, and early stopping is applied based on validation loss.

AMIC's architecture enables it to produce interpretable, context-aware sentiment predictions. Unlike black-box models, AMIC explicitly identifies which words influence the sentiment classification and quantifies their contributions. This transparency makes it particularly suitable for domains like wine reviews, where understanding the linguistic basis of sentiment is as important as predictive accuracy.

### f. LLAMA 3.1 (fine-tuned)

LLAMA 3.1 is a large-scale, decoder-only transformer model with 8 billion parameters, originally developed for autoregressive text generation tasks. To assess its applicability to sentiment classification, we first evaluated the model in both zero-shot and few-shot prompting settings without any fine-tuning. In the zero-shot setting, we posed direct classification prompts such as:

> *Here is a wine review: "Aromas of blackberry and leather lead into a structured palate of ripe plum and subtle spice." Is this review positive or negative?*

Despite the clarity of the prompt, LLAMA 3.1 consistently predicted positive sentiment across nearly all test cases. This tendency may be attributed to the formal and often neutral tone of wine reviews, which can obscure sentiment cues and bias the model toward positive interpretations.

We then explored few-shot prompting by including a small number of labeled examples (typically 3–5) within the prompt to guide the model. However, even with carefully selected positive and negative review examples, LLAMA 3.1's predictions remained largely unchanged from the zero-shot setting, indicating limited adaptability in this context.

Given these limitations, we proceeded to fine-tune LLAMA 3.1 for our binary sentiment classification task. We adopted a *text-to-text* formulation, where each input consists of a wine review followed by a prompt (e.g., "*Is this review positive or negative?*") and the target output is either *positive* or *negative*. To efficiently adapt the model while minimizing computational overhead, we employed Low-Rank Adaptation (LoRA; Hu et al., 2021), a parameter-efficient technique that introduces trainable low-rank matrices into selected layers of the model while keeping the majority of parameters frozen. This approach significantly reduces memory and computational requirements, enabling efficient adaptation of large models like LLAMA 3.1 to domain-specific tasks.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

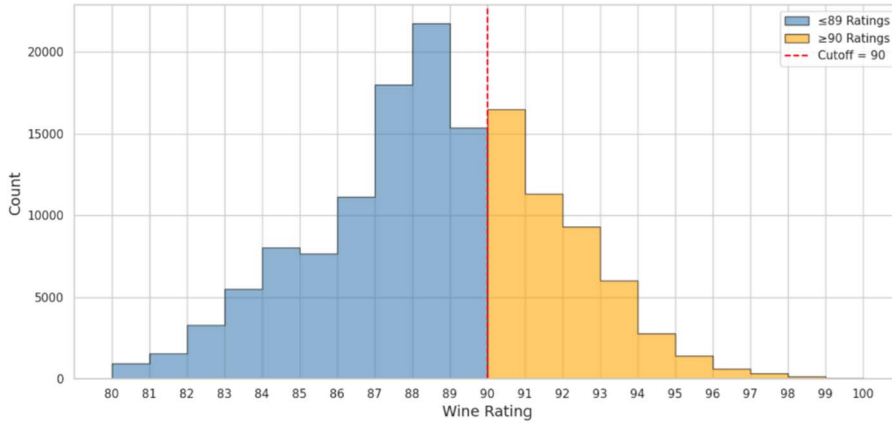



**Figure 1.** Histogram of wine ratings.

**Table 1.** Test accuracy and parameter count of different models

| Model | Accuracy (%) | Sensitivity (%) | Specificity (%) | # Parameters |
|---|---|---|---|---|
| Logistic regression | 87.75 | 79.69 | 91.89 | ~500 K |
| XGBoost | 86.66 | 76.61 | 91.27 | ~1 M |
| LSTM | 88.24 | 80.26 | 92.38 | ~1.5 M |
| BERT (fine-tuned) | 88.92 | 82.26 | 92.37 | ~110 M |
| AMIC | 89.20 | 82.69 | 92.60 | ~2 M |
| LLAMA 3.1 (fine-tuned) | 89.23 | 82.62 | 92.66 | ~8 B |

## III. Application

To evaluate the practical effectiveness of each model, we conducted a comparative analysis using a large-scale dataset of wine reviews sourced from Wine Spectator (Yang et al., 2022). Each review was paired with an expert rating on a 100-point scale. For binary sentiment classification, reviews with scores between 90 and 100 were labeled as *positive*, and those between 80 and 89 as *negative*. The distribution of wine ratings is shown in Figure 1, with fewer than 40% of wines receiving a rating of at least 90.

The classification results are summarized in Table 1. Traditional models, including logistic regression and XGBoost, when paired with TF-IDF feature representations, delivered solid baseline performance with accuracies of 87.75% and 86.66%, respectively. Neural network models demonstrated improved performance. The LSTM model, which captures sequential dependencies, achieved 88.24% accuracy. BERT, a transformer-based model fine-tuned on our dataset, further improved performance to 88.92%, highlighting the benefits of transfer learning and bidirectional context modeling.

The sensitivity of AMIC shows more significant improvement over the benchmark logistic regression (82.69% vs 79.69%) than the overall accuracy metric (89.20% vs 87.5%), while the specificity stays at almost the same level. This difference arises

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press




(a) Top 50 positive sentiment words identified by AMIC.




(b) Top 50 negative sentiment words identified by AMIC.

**Figure 2.** Word clouds generated from AMIC's learned word sentiment scores. Size of a word is proportional to the absolute value of its sentiment score.

because the dataset contains more positively labeled reviews, meaning the number of true positive cases exceeds that of true negatives. Across all three metrics, the AMIC model outperformed both LSTM and BERT. This result is particularly notable because AMIC does not rely on pretraining or massive parameter counts. Instead, it leverages a statistical multiple instance learning structure and the attention algorithm to incorporate the contextual information from wine reviews. What sets AMIC apart is not just its accuracy, but its interpretability. During training, AMIC learns to distinguish between sentiment-relevant and neutral words, assigning each a context-aware sentiment score to each individual word. These scores (presented in Tables A1 and A2 in the appendix) can be visualized and analyzed, offering insights into how the model interprets language.

Figure 2 displays two word clouds generated from AMIC's learned sentiment scores. The left panel, shaped like an upright wine glass, highlights the top 50 words most strongly associated with positive sentiment. In contrast, the right panel, shaped like an inverted wine glass, shows the 50 words most strongly linked to negative sentiment. To enhance readability, morphologically similar words are grouped under a shared root. These visualizations reveal that while many sentiment words align with conventional expectations (e.g., *gorgeous, beautiful, diluted*, and *stale*), others reflect domain-specific usage.

For instance, the word *stained* in wine reviews often conveys a sense of depth and complexity in flavor, qualities typically associated with high-end wines. A review might describe "… the long charcoal *stained* finish has a nice tug of roasted bay leaf and truffle, and this shows terrific range …," using *stained* to evoke richness and layered character. Similarly, the word *carpet* is used metaphorically to describe a luxurious, velvety texture, as in "…sail across a *carpet* of superfine tannins lingering on the spicy finish…,"

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

suggesting a smooth and refined mouthfeel. Note that those words are not typically associated with positive sentiment in everyday language, yet they convey a positive connotation in the domain of wine reviews.

AMIC's analysis of positively weighted sentiment words reveals a consistent association between high-quality wines and descriptors that imply sophistication, nuance, and structural complexity. In contrast, words that might carry positive connotations in everyday language, such as *quick* or *breezy*, are often viewed negatively in the context of wine reviews. These terms suggest simplicity or a lack of development, as seen in phrases like "…light and *quick* with lemon pulp and jicama notes…" or "…tender with modest green apple and green melon notes featuring an open *breezy* finish… ." Similarly, descriptors like *straightforward* and *easygoing* are interpreted as indicators of limited depth or refinement, as in "…this *straightforward* red shows light cherry herbal and vanilla flavors…" or "…light and *easygoing* with pretty pear and green melon flavors… ." These examples underscore AMIC's ability to uncover domain-specific sentiment patterns that diverge from general language norms, which is an essential capability for accurate SA in specialized fields like wine evaluation.

The fine-tuned LLAMA 3.1 model achieved the highest classification accuracy at 89.23%, narrowly surpassing AMIC and BERT. While this result underscores the impressive capabilities of LLMs, the performance gain was marginal, just a fraction of a percentage point above AMIC. This modest improvement came at a substantial cost: LLAMA 3.1's 8-billion-parameter architecture demands significantly more computational power, memory, and fine-tuning effort compared to smaller, task-specific models. Furthermore, LLAMA 3.1 does not provide word-level sentiment score for interpretability.

Despite its scale, LLAMA's advantage in this domain-specific task was not transformative. In fact, AMIC, with only 2 million parameters and no reliance on pretraining, delivered nearly equivalent accuracy while offering superior interpretability and efficiency. This outcome challenges the assumption that LLMs are inherently better for all NLP tasks, particularly in specialized domains like wine reviews where domain-specific language and subtle sentiment cues play a critical role.

## IV. Discussion

This study sets out to evaluate a range of language models for SA in the domain of wine reviews, a setting that presents unique linguistic challenges due to its use of metaphor, domain-specific vocabulary, and subtle evaluative cues. While the fine-tuned LLAMA 3.1 model achieved the highest classification accuracy, its marginal advantage over smaller, more efficient models such as AMIC and BERT invites a more nuanced interpretation of what "best" means in applied NLP tasks.

### *a. Performance in context*

LLAMA 3.1's top accuracy of 89.23% was only slightly higher than AMIC's 89.20% and BERT's 88.92%. This narrow margin is particularly striking given LLAMA's massive scale—8 billion parameters compared to AMIC's 2 million—and the significant computational resources required for its fine-tuning and deployment. In contrast, AMIC

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

achieved near state-of-the-art performance without pretraining, using a lightweight architecture specifically designed for interpretability and domain adaptability.

This result is not merely a technical footnote, it reflects a deeper insight into the nature of sentiment in wine reviews. Unlike general product reviews, wine descriptions often rely on layered, poetic language. Words like *stained, carpet*, or *breezy* carry sentiment that is highly context-dependent and often counterintuitive. For example, *stained* might evoke richness and complexity in a wine's finish, while *breezy* could imply a lack of depth or structure. AMIC's architecture, which explicitly identifies and scores sentiment-bearing words, is well-suited to capturing these nuances.

### b. Interpretability and domain insight

One of AMIC's most valuable contributions is its interpretability. By assigning sentiment scores at the word level and aggregating them into a document-level prediction, AMIC allows users to see not only what the model predicts, but why. This is especially important in the wine domain, where understanding the linguistic basis of sentiment can inform marketing, product development, and consumer education.

The word clouds generated from AMIC's word sentiment scores illustrate this clearly. Positive sentiment words like *gorgeous, velvety*, and *stained* reflect the language of high-quality wine, while negatively weighted terms such as *quick, diluted*, and *easy-going* suggest wines lacking in complexity or refinement. These insights are not only useful for classification; they also offer a lens into how wine quality is communicated through language.

### c. Rethinking model priorities

The findings challenge the assumption that larger models are always better. While LLAMA 3.1 is a powerful generative model, its marginal performance gain in this task does not justify its complexity, especially when compared to AMIC's efficiency and transparency. Moreover, LLAMA's generative nature introduces risks of hallucination and unverifiable reasoning, which are issues that are particularly problematic in domains requiring accountability and trust. In contrast, AMIC's predictions are grounded in observable input features. Its transparent structure ensures that sentiment decisions can be traced back to specific words and phrases, making it a more reliable tool for applications where interpretability is not optional but essential.

### d. Implications for applied NLP

These results underscore the importance of aligning model choice with task requirements. In domains like wine economics, where domain-specific language and interpretability matter as much as raw accuracy, smaller, purpose-built models like AMIC may offer the best balance of performance, usability, and insight. Furthermore, AMIC's success suggests a broader opportunity: the development of interpretable, domain-adapted models that can rival or exceed the performance of large-scale LLMs in specialized tasks. Rather than defaulting to the largest available model, practitioners should consider the full spectrum of trade-offs, including computational cost, transparency, and domain relevance.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

### e. Looking ahead

In this study, the wine ratings are dichotomized into two categories, which inevitably discards some of the richer signal in the data (e.g., a score of 95 conveys more than a score of 91). Rather than comparing models solely on classification performance, we can evaluate them based on their predictive accuracy. Future work could also explore extending AMIC to handle multi-class sentiment, incorporate phrase-level or syntactic features, or adapt it to other domains with rich, specialized language (e.g., art criticism, medical notes, or legal opinions). Additionally, hybrid approaches that combine the interpretability of AMIC with the generalization power of pretrained transformers may offer promising directions.

Ultimately, this study demonstrates that **accuracy and interpretability are not mutually exclusive**. In fact, in many real-world applications, they must go hand in hand. As AI continues to shape decision-making in specialized fields, models like AMIC—transparent, efficient, and domain-aware—will be essential tools for responsible and effective NLP.

**Acknowledgements.** The authors would like to thank Karl Storchmann, the journal editor, and the anonymous reviewer for their constructive feedback and speedy review.

## References

Ahmed, T., and Devanbu, P. (2022). Few-shot training LLMs for project-specific code-summarization. *In 37th IEEE/ACM International Conference on Automated Software Engineering*, Article No.: 177, 1–5. Association for Computing Machinery.

Alpay, T., Heinrich, S., and Wermter, S. (2016). Learning multiple timescales in recurrent neural networks. In Villa A, Masulli P, Pons Rivero A (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2016*. Springer.

Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. *Journal of Intelligent Information Systems*, *55*, 51–66. doi:10.1007/s10844-019-00591-8

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. Association for Computing Machinery.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 551–561. Association for Computational Linguistics.

Das, S. R., and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388. doi:10.1287/mnsc.1070.0704

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–4186. Institute for Operations Research and the Management Sciences.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., and Van Der Maaten, L. (2024). The Llama 3 herd of models. *preprint (arXiv:2407.21783)*.

He, J., Zhai, J., Antunes, T., Wang, H., Luo, F., Shi, S., and Li, Q. (2022) FasterMoE: Modeling and optimizing training of large-scale dynamic pre-trained models. *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 120–134. Association for Computing Machinery.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

Katumullage, D., Yang, C., Barth, J., and Cao, J. (2022). Using neural network models for wine review classification. *Journal of Wine Economics*, *17*(1), 27–41. doi:10.1017/jwe.2022.2

Levallois, C. (2025). Umigon-lexicon: Rule-based model for interpretable sentiment analysis and factuality categorization. *Language Resources & Evaluation*, *59*, 913–930. doi:10.1007/s10579-024-09742-y

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Association for Computational Linguistics.

Sánchez-Rada, J. F., and Iglesias, C. A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*, *52*, 344–356. doi:10.1016/j.inffus.2019.05.003

Sezerer, E., and Tekir, S. (2021). A survey on neural word embeddings. *arXiv preprint arXiv*:2110.01804.

Sharma, N. A., Ali, A. B. M. S., and Kabir, M. A. (2025). A review of sentiment analysis: Tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics*, *19*, 351–388. doi:10.1007/s41060-024-00594-x

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. doi:10.1108/eb026526

Tyagi, A., and Sharma, N. (2018). Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology(UAE)*, *7*(2), 20–23. doi:10.14419/ijet.v7i2.24.11991

Yang, C., Barth, J., Katumullage, D., and Cao, J. (2022). Wine review descriptors as quality predictors: Evidence from language processing techniques. *Journal of Wine Economics*, *17*(1), 64–80. doi:10.1017/jwe.2022.3

Yang, C., and Cao, J. (2025). Interpretable sentiment analysis using the attention-based multiple instance classification model: An application to wine reviews. *Harvard Data Science Review*, *7*(2). doi:10.1162/99608f92.caab9466

## Appendix

**Table A1.** AMIC's list of top 50 positive sentiment words

| | | | | |
|---|---|---|---|---|
| Gorgeous (205.1) | Beautiful (176.9) | Ethereal (171.1) | Beautifully (169.5) | Gloriously (167.7) |
| Gorgeously (162.4) | Thoroughly (157.9) | Drips (150.4) | Beauty (150.1) | Impeccable (149.0) |
| Amazingly (148.4) | Exquisite (146.5) | Strikingly (145.6) | Sumptuous (145.2) | Cognac (144.8) |
| Burgundy (143.8) | Breed (143.5) | Velvet (142.5) | Cascading (142.0) | Haunting (141.9) |
| Seductive (141.2) | Finely (140.3) | Stuffed (140.2) | Soak (140.1) | Lovely (139.8) |
| Soaked (139.1) | Perfectly (138.3) | Deliciously (137.5) | Brilliantly (136.6) | Impeccably (135.8) |
| Wonderful (135.4) | Drip (134.8) | Luxuriant (133.5) | Glistening (132.6) | Silk (131.3) |
| Truffle (131.3) | Charms (131.1) | Brunello (130.6) | Soothing (130.2) | Carpet (130.2) |
| Champagne (130.1) | Perfume (130.0) | Seductively (129.7) | Fabric (129.6) | Unobtrusive (129.2) |
| Sings (128.7) | Swirl (128.4) | Stained (126.8) | Wonderfully (126.8) | Elegance (126.7) |

*Note:* Numbers in parentheses denote raw sentiment scores.

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press

**Table A2.** AMIC's list of bottom 50 negative sentiment words

| | | | | |
|---|---|---|---|---|
| Quick (−234.8) | Generic (−189.1) | Hearted (−185.3) | Simple (−171.3) | Canned (−164.6) |
| Uncomplicated (−162.3) | Diluted (−160.6) | Tinny (−156.1) | Neutral (−154.7) | Straightforward (−152.8) |
| Stale (−150.3) | Cocktail (−149.4) | Easygoing (−148.3) | Fizzy (−147.1) | Picnic (−145.3) |
| Flat (−143.2) | Lovage (−137.2) | Greenish (−135.3) | Unfocused (−131.8) | Breezy (−130.3) |
| Beaujolais (−130.3) | Metallic (−129.8) | Dull (−123.9) | Easy (−122.8) | Tail (−121.0) |
| Modestly (−119.4) | Decent (−118.5) | Fade (−118.4) | Scallion (−118.2) | Modest (−115.9) |
| Cucumber (−115.0) | Cloying (−112.2) | Watermelon (−110.1) | Soft (−107.8) | Parsley (−106.8) |
| Asparagus (−105.9) | Kosher (−105.3) | Muddled (−105.0) | Herbal (−105.0) | Lemonade (−103.8) |
| Detract (−102.7) | Weedy (−102.6) | Blunt (−102.1) | Tired (−101.9) | Muscadet (−101.4) |
| Grass (−101.3) | Grassy (−99.8) | Chilled (−99.1) | Trim (−98.9) | Overripe (−98.7) |

*Note:* Numbers in parentheses denote raw sentiment scores.

**Cite this article:** Yang, C., and Cao, J. (2025). Comparison of language models for wine sentiment analysis. *Journal of Wine Economics*, 20(4), 371–384. https://doi.org/10.1017/jwe.2025.10091

https://doi.org/10.1017/jwe.2025.10091 Published online by Cambridge University Press