**A Modular Framework for Cost-Efficient Sentiment Analysis with Small LMs**

**Abstract**

Aspect-based sentiment extraction is an essential task in e-commerce analytics, enabling businesses to capture opinions on product attributes such as screen quality, battery life, or delivery experience. Current pipelines rely heavily on large language models (LLMs) such as GPT-4, which achieve high accuracy but incur prohibitive computational and financial costs. This study investigates whether Agents of small language models (SLMs) with 1–3 Billion parameters can deliver comparable performance at significantly reduced cost. We compare a GPT-4o one-shot baseline against a four-agent SLM agents composed of a chunker, feature-finder, sentiment-scorer, and coordinator/cost-logger. Using subsets of Amazon product reviews, we evaluate attribute-level F1, sentiment mean absolute error (MAE), latency, and cost per 100 reviews. Our goal is to provide the first reproducible cost-versus-quality benchmark for micro-agent SLM agents, offering practical guidance to organizations seeking LLM-level accuracy without prohibitive expenses.

**1. Introduction**

Sentiment analysis has emerged as a cornerstone of natural language processing (NLP) and e-commerce analytics, enabling organizations to translate large volumes of unstructured customer feedback into actionable insights. While early systems focused on overall polarity classification (positive, negative, neutral), contemporary applications increasingly require aspect-based sentiment analysis (ABSA), which identifies opinions tied to specific product attributes. For instance, a single review may praise a laptop’s display quality while criticizing its battery life. Extracting such fine-grained sentiment is critical for competitive product design, marketing, and consumer experience optimization.

Recent advances in transformer-based large language models (LLMs) such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and GPT-4 (OpenAI, 2023) have set new performance benchmarks in ABSA. However, these gains come at a substantial cost. LLMs are computationally intensive, exhibit high latency, and require expensive API usage, creating barriers to adoption, particularly for small and mid-sized enterprises. The reliance on monolithic, high-parameter models also raises concerns regarding efficiency, scalability, and sustainability (Schwartz et al., 2020).

To address these limitations, this study investigates whether small language models (SLMs) in the 1–3 Billion parameter range can serve as cost-effective alternatives. Instead of deploying a single large model, the proposed approach decomposes the ABSA pipeline into specialized subtasks such as text chunking, feature extraction, and sentiment scoring. These subtasks are handled by dedicated SLM-based agents within a modular framework. The underlying premise is that specialization and coordination among smaller models may approximate LLM-level performance while significantly reducing computational and financial overhead.

This study is guided by the following research questions:

* **RQ1:** At equal review coverage, can a modular SLM framework achieve attribute-level F1 within five points of a GPT-4o one-shot pipeline?
* **RQ2:** Does task specialization across multiple SLM agents outperform a single SLM of equivalent size?

Based on prior work in modular NLP architectures and efficiency-driven sentiment analysis, the study is grounded in the following hypotheses:

* **H1:** The SLM framework will reduce token costs by at least 80% (≤ 20% of GPT-4o) while maintaining ΔF1 within five points.
* **H2:** Task specialization across multiple SLM agents will yield higher attribute-level F1 and lower mean absolute error (MAE) than a monolithic SLM of equivalent total parameters.

By empirically evaluating this framework on Amazon product review datasets, the research aims to establish the first cost-versus-quality benchmark for modular SLM-based sentiment analysis. The findings are expected to provide guidance for practitioners seeking scalable, accurate, and affordable alternatives to large LLMs in aspect-based sentiment extraction.

**2. Literature Review**

**2.1 Sentiment Analysis and Large Language Models**

Research on sentiment analysis has progressed from lexicon-based and classical machine-learning methods to deep neural architectures that model context more effectively. Transformer-based models, beginning with BERT, established strong baselines across a range of NLP tasks by leveraging bidirectional contextual pre-training (Devlin et al., 2019). Subsequent scale-ups, such as GPT-3, demonstrated competitive few-shot and zero-shot performance on diverse language understanding tasks, including sentiment classification settings, without task-specific fine-tuning (Brown et al., 2020). Although these models provide high accuracy, their computational and financial costs remain substantial, motivating work that evaluates systems on efficiency as well as accuracy (Schwartz et al., 2020).

**2.2 Modular, Distilled, and Ensemble Approaches**

To reduce cost while retaining strong accuracy, several lines of research explore modularity (decomposing a task into coordinated subtasks), model compression/distillation, and ensembling. Multi-task and modular formulations can improve sample efficiency and generalization by structuring learning around complementary subtasks (Ruder, 2017). Knowledge distillation and compression methods (e.g., TinyBERT) show that smaller student models can retain a large fraction of teacher performance with significantly lower inference cost (Jiao et al., 2020). In parallel, ensemble strategies combining multiple competent but diverse models have long been shown to improve robustness and generalization (Dietterich, 2000). Taken together, these strands suggest that coordinating smaller, specialized models can be a viable path toward accuracy-efficiency trade-offs that are difficult to achieve with a single large model.

**2.3 Challenges Specific to Aspect-Level Extraction**

Aspect-based sentiment analysis (ABSA) requires identifying opinion targets (aspects) and assigning sentiment at the aspect level, which is more challenging than document- or sentence-level polarity. Benchmarking efforts such as the SemEval ABSA tasks codified common subtasks (aspect term extraction, aspect category detection, aspect-level sentiment) and clarified typical failure modes (Pontiki et al., 2014, 2016). Persistent difficulties include sarcasm/irony, implicit aspects, and context-dependent polarity shifts. For example, sarcasm remains difficult for neural models due to its pragmatic nature (Ghosh & Veale, 2017), and surveys report that even modern approaches struggle when sentiment is conveyed implicitly or relies on extra-sentential context (Joshi et al., 2017). These issues are directly relevant to modular systems: if subtasks are split across agents (e.g., aspect detection and sentiment assignment), coordination must preserve contextual cues are needed to resolve such phenomena.

**2.4 Efficiency, Deployment, and Evaluation Beyond Accuracy**

A parallel literature argues that progress in NLP should be assessed by accuracy and by computational, financial, and environmental costs. The Green AI perspective recommends reporting metrics such as training/inference compute and energy alongside standard task scores (Schwartz et al., 2020). Parameter-efficient adaptation methods (e.g., LoRA) reduce the cost of customizing models to new tasks by training small adapter modules rather than full networks (Hu et al., 2021). Together with compression/distillation, these techniques provide a design space in which small or adapted models can deliver competitive performance under realistic deployment constraints.

**2.5 Summary and Research Gap**

Prior work establishes (i) the effectiveness of transformers for sentiment analysis, (ii) practical routes to efficiency via modular decomposition, compression/distillation, ensembling, and parameter-efficient adaptation, and (iii) the continued difficulty of ABSA phenomena such as sarcasm and implicit aspects. What is missing is a systematic, reproducible benchmark that quantifies the cost–quality–latency trade-off of a modular, small-model pipeline for aspect-level sentiment extraction against a strong LLM baseline on realistic e-commerce data. The present study addresses this gap by evaluating a coordinated set of small language model agents for ABSA and comparing them to a GPT-4–class baseline using accuracy (attribute-level F1), error (MAE), cost, and latency, with ablations that probe the contribution of each module.

**3. Methods**

**3.1 Datasets**

We will use three subsets of the Amazon 5-core review dataset: Electronics, Kindle Store, and Clothing/Shoes/Jewelry. Together these datasets provide diverse domains with attribute-rich reviews, totaling approximately 800 MB compressed. Each dataset contains at least five reviews per user and item, ensuring density and representativeness.

**3.2 Preprocessing**

From the combined datasets, 300,000 reviews will be sampled. Data will be split into training (80%), development (10%), and test (10%). Relevant fields retained will include reviewText, ratings, and helpfulness votes. For evaluation, 10,000 sentences will be auto-labeled using a lexicon-pattern approach, supplemented by 500 manually validated reviews to ensure inter-annotator reliability (target κ ≥ 0.8).

**3.3 Systems**

* **Baseline:** A GPT-4o one-shot pipeline, priced at $2.50 input / $10 output per million tokens.
* **SLM:**
  1. **Chunker** – Python rule-based segmentation of long reviews.
  2. **Feature-Finder** – Phi-3-Mini-128k-Instruct (3.8B parameters, Hugging Face).
  3. **Sentiment-Scorer** – Mistral-3B-Instruct (3B parameters, Mistral AI).
  4. **Coordinator/Cost-Logger** – Python module that merges outputs, deduplicates features, aggregates sentiment, and logs token usage and latency.

**3.4 Controls**

To ensure comparability, all systems will:

* Adhere to the same JSON schema for feature and sentiment outputs.
* Use identical prompts tailored per role.
* Run deterministically with temperature = 0 and fixed random seeds.

**3.5 Hardware**

Experiments will run on a single VM with 8 vCPUs and 32 GB RAM, or Colab Pro, with all SLMs executed in CPU-only mode. This reflects accessible, cost-conscious deployment conditions.

**4. Results**

Performance will be evaluated along four dimensions:

* **Attribute-level F1:** Computed with 5,000 bootstrap resamples to estimate 95% confidence intervals. McNemar’s test will be applied to paired comparisons of ΔF1.
* **Sentiment Mean Absolute Error (MAE):** Analyzed using paired t-tests.
* **Cost Efficiency:** Measured as cost per 100 reviews, logged directly by the coordinator.
* **Latency (p95):** Compared between systems using Wilcoxon signed-rank tests.
* **Effect Sizes:** Reported using Cohen’s d for MAE and relative cost ratios to contextualize differences.

**5. Discussion**

This research is expected to:

1. Establish the first public cost-versus-quality benchmark for SLM agents in aspect-based sentiment analysis.
2. Release an open-source pipeline (including notebooks, prompts, and cost logging) for reproducibility.
3. Provide an error taxonomy identifying common feature misses (e.g., paraphrases, sarcasm, rare attributes).
4. Deliver practical guidance on when micro-agents are sufficient substitutes for large LLMs in industry applications.

**6. Risks and Mitigations (add to Discussion)**

* **Prompt brittleness → low recall:** Mitigated with prompt tuning and few-shot examples.
* **Fuzzy feature labels → scoring noise:** Addressed with string normalization and alias lists.
* **Token cost fluctuations:** Prices will be recorded at the time of experiment and recalculated if APIs change.

**7. Reproducibility (add to Discussion)**

Reproducibility will be ensured through:

* Release of fixed model SHA hashes, seed values, and a run.sh script for dataset download and pipeline execution.
* A Dockerfile with pinned dependencies (mistral\_rs, transformers, accelerate, openai).
* Archiving of processed Parquet files (~1 GB) on Zenodo with a DOI.

need to add in your Conclusion

**References**

* 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*, 1877–1901. [https://arxiv.org/abs/2005.14165](https://arxiv.org/abs/2005.14165?utm_source=chatgpt.com)
* Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
* Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, 1–15. <https://doi.org/10.1007/3-540-45014-9_1>
* Ghosh, A., & Veale, T. (2017). Fracking sarcasm using neural network architectures. *ACL 2017*. <https://aclanthology.org/P17-2068>
* Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., … Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*. <https://arxiv.org/abs/2106.09685>
* Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., … Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. *Findings of EMNLP 2020*, 4163–4174. <https://aclanthology.org/2020.findings-emnlp.372>
* Joshi, A., Sharma, V., & Bhattacharyya, P. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys, 50*(5), 1–22. <https://doi.org/10.1145/3124420>
* Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. *SemEval 2014*, 27–35. <https://aclanthology.org/S14-2004>
* Pontiki, M., et al. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *SemEval 2016*, 19–30. <https://aclanthology.org/S16-1002>
* Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*. <https://arxiv.org/abs/1706.05098>
* Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. <https://doi.org/10.1145/3381831>