OPEN ACCESS

Digital Applied Linguistics
ISSN 2982-2300
https://www.castledown.com/journals/dal/

Digital Applied Linguistics, 3, 102585 (2025)
https://doi.org/10.29140/dal.v3.102585

High-accuracy, privacy-compliant multilingual sentiment categorization on consumer-grade hardware: A monte carlo evaluation of locally deployed large language models

MICHELE CARLO[a]

OSAMU TAKEUCHI, PH.D.[b]

[a] Graduate School of Foreign Language Education and Research, Kansai University, Osaka, Japan

[b] Professor, Faculty of Foreign Language Studies, Kansai University, Osaka, Japan

Abstract

This study presents a comprehensive evaluation of multilingual sentiment categorization performance using locally deployed large language models (LLMs) on consumer-grade hardware, focusing on GDPR-compliant implementation scenarios. Through extensive Monte Carlo validation involving 947,700 classifications over 702 iterations, we demonstrate significant performance capabilities across English, Italian, and Japanese languages while operating within consumer hardware constraints. Using lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half on a Python-based llama-cpp framework on consumer NVIDIA GPU hardware, English achieved 96.3% accuracy (95% CI: 0.963–0.964), with Italian and Japanese showing strong performance at 92.2% (95% CI: 0.921–0.922) and 90.7% (95% CI: 0.906–0.908) respectively. Notably, our analysis demonstrates that plurality voting can achieve extremely high confidence levels across all languages, suggesting an efficient approach to improving classification reliability without requiring extensive computational resources. Furthermore, these findings provide a substantive contribution to digital applied linguistics by demonstrating how locally deployable, resource-efficient multilingual LLMs can inform refined sentiment-based inquiries and pedagogical innovations across diverse linguistic environments.

Keywords: Multilingual sentiment categorization; local large language models; Monte Carlo validation; consumer-grade hardware; privacy-compliant LLM implementation

Copyright © 2025
Michele Carlo & Osamu Takeuchi

Introduction

This study makes a significant contribution to the field of digital applied linguistics by exploring the practical implementation of large language models (LLMs) in multilingual contexts, particularly focusing on their deployment in resource-constrained environments. The democratization of LLMs, fueled by advances in model optimization and deployment frameworks, has opened new avenues for natural language processing (NLP) applications. While cloud-based solutions have traditionally dominated the landscape, the ability to run sophisticated LLMs locally marks a transformative shift in accessibility and usability.

By examining the performance of locally deployed LLMs in multilingual sentiment categorization across English, Italian, and Japanese—three languages representing distinct linguistic families and Latin versus non-Latin writing systems—this research highlights the practical challenges and opportunities of applying LLMs in diverse linguistic contexts. The study's focus on achieving reliable performance within the constraints of consumer hardware addresses a critical gap between the theoretical potential of LLMs and their real-world application.

The findings contribute to digital applied linguistics by providing insights into optimizing multilingual NLP systems for accessibility, efficiency, and accuracy. This work underscores the importance of bridging the divide between advanced language technologies and their practical deployment, offering valuable implications for researchers, educators, and developers in the field.

Literature Review

Foundations of Automated Sentiment Analysis

Sentiment analysis, or opinion mining, extracts subjective information (e.g., polarity, emotion) from text (Liu, 2012; Pang & Lee, 2008). Early research treated it as a classification problem, applying machine learning to identify emotional content in reviews, social media, and news (Pang et al., 2002). Researchers then integrated lexicon-based methods (Taboada et al., 2011), which rely on domain-specific word lists and valence scores, useful when labeled data is sparse (Thelwall et al., 2010). By the early 2010s, deep neural architectures (Kim, 2014) outperformed feature-engineered models, yielding superior results on large datasets.

Despite progress, sentiment analysis remains complex due to cultural and contextual nuances, including irony or sarcasm (Ghosh et al., 2017; Schouten & Frasincar, 2015). In response, researchers developed context-aware architectures (Devlin et al., 2019) that leverage massive pre-trained language models, highlighting the need to capture linguistic subtleties across genres, domains, and cultures.

Multilingual Sentiment Analysis

Although early sentiment analysis focused on English (Pang & Lee, 2008), global applications require coverage of typologically diverse languages. Researchers have extended methods to Romance languages (Balahur & Turchi, 2014; Mozetič et al., 2016) and script-heavy languages such as Chinese, Arabic, and Japanese (Denecke, 2008; Zhang et al., 2011). Still, performance gaps remain for less-resourced or morphologically complex languages (Barnes et al., 2018).

Recent multilingual models (Conneau & Lample, 2019; Xue et al., 2021) and cross-lingual transfer techniques (Artetxe & Schwenk, 2019) share learned representations, boosting overall performance. Yet, highly inflected or script-heavy languages pose ongoing challenges (Ryskina et al., 2020). Researchers thus emphasize rigorous evaluation—e.g., repeated experiments, cross-validation, and ensemble methods—to address data variability (Dror et al., 2018; Dodge

et al., 2019), underscoring the importance of replicable methodology for multilingual sentiment analysis.

Privacy Concerns and GDPR in NLP

Large-scale text analytics has magnified privacy and security issues, particularly under the EU's General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche, 2017). While cloud-based NLP services process sensitive data on remote servers, on-premise solutions have emerged to ensure compliance, reduce data leakage, and maintain user trust (Veluru et al., 2014). Scholars highlight ethical and technical challenges in storing personal text in third-party systems (Kantarcioğlu et al., 2004). Consequently, local deployment of language models that never transmit raw user data aligns with current trends in data minimization, decentralized AI, and edge computing (Shi et al., 2016).

Large Language Models and Resource Constraints

BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) revolutionized NLP but demand considerable computation and storage (Strubell et al., 2019), often exceeding consumer hardware capabilities (Raschka et al., 2022). Innovations in model compression, such as quantization (Dettmers et al., 2022; Frantar et al., 2023) and pruning (Han et al., 2015), have enabled large language models (LLMs) to run on smaller devices with minimal accuracy loss. For instance, 8-bit and 4-bit quantization can significantly reduce model size while retaining most performance (Dettmers et al., 2022; Frantar et al., 2023). Llama-based open-source foundational models (Touvron et al., 2023) have further democratized NLP tasks, including chat and summarization. Some studies show that even 7B or 13B-parameter LLMs can be deployed on high-end consumer GPUs (Chung et al., 2022), although multilingual sentiment analyses remain relatively underexplored (Hashimoto et al., 2016).

Statistical Validation in Resource-Constrained Environments

Single-run train/test splits risk obscuring variability and inflating performance metrics (Krymolowski, 2001). Monte Carlo methods and repeated resampling (Efron & Tibshirani, 1994) are best practices in applied statistics but remain unevenly adopted in NLP (Dror et al., 2018). Multiple runs (e.g., 100+ iterations) help estimate confidence intervals, identify rare misclassifications, and reduce random seed effects (Bouthillier et al., 2019). These challenges intensify in multilingual contexts, given the cross-lingual variation and complex morphologies (Koehn & Knowles, 2017). Bootstrap resampling offers more reliable F1, precision, and recall estimates (Poliak et al., 2018), while ensemble approaches (e.g., majority voting) can improve classification stability with minimal overhead (Dietterich, 2000).

Synthesis and Research Gap

Sentiment analysis has progressed from lexicon-based methods to advanced deep neural architectures capable of cross-linguistic adaptation. This evolution addresses typological differences and cultural nuances, yet privacy regulations (e.g., GDPR) command a shift toward local, on-device processing. Although model compression (quantization, pruning) has enabled large language models to run on consumer-grade hardware, multilingual benchmarks often lack rigorous variability evaluation. Monte Carlo validation, bootstrap resampling, repeated trials, and ensemble strategies are increasingly recognized as essential for robust evaluation, but many multilingual NLP studies do not adopt this level of statistical rigor.

This study responds by demonstrating the local deployment of a quantized large multilingual model on consumer hardware, supported by systematic Monte Carlo resampling for robust

performance assessment. By merging methodological rigor (ensemble voting, confidence intervals) with practical demands (GDPR compliance, real-time inference), we achieve high-accuracy multilingual sentiment analysis under strict privacy constraints.

Research Questions

In this study, we sought to rigorously examine the feasibility, reliability, and practical implications of deploying multilingual sentiment categorization models locally on consumer-grade hardware through four research questions:

- RQ1: Can multilingual language models achieve high accuracy and robust performance in sentiment categorization when locally deployed on consumer-grade hardware?
- RQ2: How resource-efficient and statistically reliable are these local deployments, and can they meet stringent privacy and compliance requirements (e.g., GDPR)?
- RQ3: Can ensemble techniques (such as plurality voting) improve categorization confidence and reduce uncertainty in locally deployed models without significantly increasing computational overhead?
- RQ4: How does statistical behavior (e.g., variance, distributional properties, intra-rater agreement) inform best practices for local model evaluation and deployment in resource-constrained environments?

Methodology

Monte Carlo Validation

A Monte Carlo validation methodology, following repeated-sampling and bootstrap approaches (Efron & Tibshirani, 1994; Dror et al., 2018), was implemented to evaluate the robustness of multilingual sentiment analysis across English, Italian, and Japanese languages. This non-parametric approach employed bootstrap resampling (R = 1,000) to quantify uncertainty and assess model stability, circumventing the limitations of traditional parametric methods in natural language processing. The methodology's capacity to handle non-normal distributions and cross-lingual variations made it particularly suitable for evaluating sentiment analysis performance across diverse linguistic contexts: by systematically sampling from stratified language-specific test sets, the approach enabled robust confidence intervals for performance metrics while accounting for cultural and linguistic nuances in emotional expression. Ultimately, this framework provides a more ecologically valid assessment of model performance compared to conventional fixed test set evaluations, offering insights into both model stability and cross-lingual generalization capabilities.

Iteration Count Rationale

Power and Precision Requirements

Our iteration count determination was driven by precision and robustness requirements that extended beyond basic statistical significance. While standard power analysis (Cohen's f = 2.292, α = 0.05, power = 0.95) suggested fewer iterations would suffice, we implemented stricter criteria for practical precision. Using a 99% confidence level (z = 2.576) and targeting a ±2% margin of error for high-accuracy scenarios (p = 0.96, q = 0.04), calculations indicated a requirement of approximately 700 iterations per language, accounting for initial test runs.

The final implementation of 702 iterations per language, each evaluating 150 samples across three sentiment classes, resulted in 315,900 classifications per language and 947,700 total classifications across English, Italian, and Japanese. This comprehensive dataset enabled stable estimates of multiple metrics (accuracy, precision, recall, and F1-scores) while capturing rare misclassification patterns and extreme cases like ambiguous sarcasm or atypical phrasing. The extensive iteration count aligns with Monte Carlo best practices, allowing thorough sampling of the non-deterministic model's output distribution, assessment of intra-rater agreement, and detailed distributional analysis (normality, skewness, kurtosis), thereby ensuring robust performance evaluation where even 1–2% accuracy differences carry operational significance.

Statistical and Practical Implications

By selecting 702 iterations, we strike a balance between statistical rigor and practical relevance. While fewer iterations could detect significant differences, our chosen design provides tighter confidence intervals in high-accuracy ranges and ensures stable estimates of performance metrics across three languages and three sentiment categories. It offers robust insights into subtle differences in model behavior, critical for real-world applications where small performance margins matter.

In summary, the choice of 702 iterations per language was driven by the need for precise accuracy estimation, stable variance assessments, and comprehensive Monte Carlo validation, enabling a robust and practically useful evaluation of multilingual sentiment analysis performance.

Validation Approach

The evaluation protocol combined stratified sampling (150 samples per sentiment class across 702 iterations) with Monte Carlo validation, building on recommended repeated-experiment practices for robust NLP evaluation (Dror et al., 2018). The methodology leveraged Llama models' inherent non-deterministic behavior (Touvron et al., 2023) with temperature settings of 0.1 to prevent degeneration (Holtzman et al., 2020).

Statistical validation encompassed multiple metrics (accuracy, precision, recall, F1-scores, Cohen's kappa) following standard NLP practices (Cohen, 1960; Powers, 2011). Analysis included Anderson-Darling tests for normality (Stephens, 1974), Fligner-Killeen tests for variance homogeneity (Conover et al., 1981), Welch's ANOVA (Welch, 1951), and Games-Howell post-hoc comparisons (Games & Howell, 1976).

The study incorporated 95% confidence intervals (Cumming, 2012), Standard Error of Measurement (Nunnally & Bernstein, 1994), and within/between-language variance analysis (Brown, 1999). Additional analyses included distributional metrics (Shapiro & Wilk, 1965) and intra-rater agreement evaluation (Pavlick & Kwiatkowski, 2019). This comprehensive approach aligned with machine learning evaluation best practices (Dror et al., 2018; Dodge et al., 2019), providing a rigorous foundation for assessing multilingual model performance.

Synthetic Dataset Construction: A Multilingual Sentiment Analysis Resource

In this section, we detail the methodology employed to construct our multilingual sentiment analysis dataset. The dataset comprises evaluation questions and responses in English, Italian, and Japanese, each set encompassing positive, neutral, and negative sentiment categories. Our primary objective was to create a controlled, balanced resource that would facilitate rigorous evaluation of sentiment classification models across linguistically and culturally diverse contexts, while ensuring semantic parallelism and comparability among the three languages.

Theoretical Framework and Design Rationale

The design of this dataset is grounded in foundational insights from the sentiment analysis literature. The decision to develop a synthetic, balanced dataset was influenced by the challenges associated with naturally occurring data. Previous studies have noted that data sourced from user-generated content, such as social media or product reviews, often exhibits skewed sentiment distributions, lexical irregularities, and variable grammatical structures (Liu, 2012; Pang & Lee, 2008). While such variability reflects authentic language use, it can mask model performance differences and complicate cross-linguistic comparisons.

To address these issues, we followed the rationale of earlier work that employed artificially balanced corpora to facilitate direct and controlled comparisons of classifiers (Taboada et al., 2011; Mohammad, 2016). By doing so, we ensured equivalent representation of sentiment categories and controlled for extraneous factors that might otherwise confound cross-linguistic evaluation. Where applicable, we drew on general "service-quality" language categories (Parasuraman et al., 1988) to populate text prompts in domains like hospitality, retail, and consulting, ensuring they reflected standardized attributes such as reliability or responsiveness.

Language Selection and Cross-Linguistic Considerations

The selection of English, Italian, and Japanese was intentionally guided by their distinct typological features, orthographic systems, and associated linguistic complexities. English, which has been extensively investigated and is well-resourced for sentiment analysis (Hu & Liu, 2004), serves as a natural benchmark due to its relatively analytic morphological structure and the broad availability of sentiment lexicons and annotated corpora (Huddleston & Pullum, 2002; Bauer, 1983; Bybee, 1985). Italian, by contrast, exemplifies a Romance language with richer inflectional morphology, flexible word order, and less standardized sentiment resources, thus requiring models to navigate more intricate grammatical cues to accurately interpret sentiment (Maiden, 1995, 2018; Lepschy & Lepschy, 1988). Japanese presents yet another dimension of complexity through its multiple orthographic scripts (kanji, hiragana, katakana) and context-dependent sentiment expressions, which are deeply embedded in cultural subtleties (Denecke, 2008; Shibatani, 1990; Tsujimura, 2014; Kuno, 1973). Taken together, these languages represent a rigorously varied test bed: English provides a well-understood baseline, Italian challenges systems with morphological and syntactic variability, and Japanese compels adaptations to divergent writing systems and culturally influenced sentiment cues. By encompassing these distinct linguistic conditions, this study systematically evaluates the generalizability, robustness, and adaptability of sentiment analysis models across fundamentally different language environments.

Dataset Structure and Composition

The dataset consists of 450 samples per language, totaling 1,350 instances. Within each language subset, we allocated equal proportions of positive, neutral, and negative sentiments (150 per category). Every language subset addresses a common set of 20 evaluation questions, ensuring multiple sentiment-labeled responses per question. This structure maintains balanced sentiment distribution and supports robust model evaluation across various prompt types. Illustrative examples are available at the end of this section.

Domain Coverage and Content Distribution

To enhance ecological validity while maintaining control, the dataset includes content drawn from multiple service-oriented domains: hospitality (e.g., hotels, restaurants), professional services (e.g., consultation, training), retail (e.g., electronics, clothing), technology platforms,

healthcare services, and entertainment venues. Each domain is represented evenly across sentiment categories, with approximately 25 responses per sentiment per domain. This design ensures that sentiment classification must operate consistently across a range of contextual scenarios, rather than being limited to a single domain.

Lexical Design and Sentiment Expression

We employed a systematic approach to sentiment expression, incorporating lexicon-based principles from established research (Hu & Liu, 2004; Taboada et al., 2011). Positive responses utilized clearly positive modifiers and intensifiers (e.g., "exceptional," "outstanding"), while neutral responses relied on evaluatively minimal language (e.g., "adequate," "standard"), following guidelines that define neutrality as the absence or attenuation of strongly opinionated terms (Ding et al., 2008; Thelwall et al., 2010). Negative responses included explicit problem indicators and negative polarity markers (e.g., "failed," "terrible"), aligning with known strategies for signaling negative sentiment (Liu, 2012).

Development Methodology and Construction Process

Base Template Development

We began by creating English templates for both questions and responses, leveraging lexicon-based sentiment principles from the literature (Hu & Liu, 2004; Taboada et al., 2011). Initially, we produced 50 prototype response templates per sentiment category, varying intensifiers for positive and negative sentiments and ensuring the use of weakly evaluative language for neutral responses (Polanyi & Zaenen, 2006). These templates were then permuted and assigned across multiple domains, resulting in 150 unique instances per sentiment category.

Cross-Linguistic Adaptation

We adapted the English templates into Italian and Japanese with close attention to pragmatic equivalence and sentiment strength (Li et al., 2017). Where direct transliterations risked altering connotation, we selected culturally and linguistically appropriate terms to maintain consistent sentiment cues. For Japanese, we used both plain and formal language without affecting detection rates.

Adaptations underwent a two-step validation. First, two bilingual reviewers (for each English–Italian and English–Japanese pair) assessed whether translated texts preserved the original sentence's semantic content and polarity (positive, neutral, negative). They evaluated adequacy (i.e., critical meaning retention) and sentiment equivalence (i.e., emotional tone consistency), resolving discrepancies through iterative refinement. Second, a round-robin check compared the final English, Italian, and Japanese versions, using back-translation where needed to detect subtle shifts. If the re-translated text diverged from the original, it was revised. This process ensured each sentence retained consistent sentiment strength and cultural appropriateness across languages.

Quality Control and Validation

We used a multi-step process to ensure high-quality data across English, Italian, and Japanese. First, we balanced positive, neutral, and negative classes (450 sentences per language), corrected errors, and ensured lexical diversity via synonyms and tailored intensifiers (e.g., "exceeded / ha superato di gran lunga / はるかに超えた"), preserving strong vs. mild sentiment cues. Native annotators then reviewed a subset of sentences for semantic accuracy, sentiment fidelity, and naturalness.

Next, two fluent speakers per language independently labeled 450 sentences as positive, neutral, or negative. Cohen's kappa revealed strong agreement ($\kappa = 0.98$ for English and Italian, $\kappa = 0.94$ for Japanese), with a macro-averaged annotation accuracy of 96%. These reliability scores confirm the dataset's stable, consistent sentiment labels across languages.

Illustrative Examples

To showcase parallelism across sentiment categories, we constructed equivalent examples in English, Italian, and Japanese (Table 1).

Table 1. Cross-linguistic sentiment examples.

Class	English	Italian	Japanese
Positive	The exceptional hotel service exceeded our expectations magnificently.	Il servizio eccellente dell'hotel ha superato magnificamente le nostre aspettative.	ホテルのサービスは期待を見事に超えました。
Neutral	The electronics performed as expected, meeting basic standards.	L'elettronica ha funzionato come previsto, soddisfacendo gli standard di base.	電子機器は予想通りに動作し、基本的な基準を満たした。
Negative	The furniture delivery was frustrating and poorly handled.	La consegna dei mobili è stata frustrante e gestita male.	家具の配送は不満で、対応が悪かった。

These parallel examples illustrate how pragmatic equivalence and consistent sentiment intensity can be preserved despite structural and cultural differences.

Dataset Selection Implications

This dataset offers a controlled, multilingual testbed for evaluating sentiment analysis models. By balancing sentiment categories, incorporating multiple domains, and ensuring linguistic parallelism, it enables a more rigorous assessment of model capabilities across typologically diverse languages. The deliberate use of explicit sentiment markers, rather than subtle or context-dependent cues, ensures that observed model differences stem from linguistic variation rather than ambiguous content (Mohammad, 2016; Pang & Lee, 2008).

Ultimately, this dataset contributes a valuable resource for multilingual sentiment analysis research, supporting both the development of more robust cross-linguistic models and the advancement of sentiment analysis methodologies that can operate effectively in varied linguistic and cultural settings.

Model Selection Rationale

Unlike English-focused models, *lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half* extends Llama 3's capabilities to multiple languages including French, German, Russian, Chinese, and Japanese (Devine, 2024a), addressing crucial non-English sentiment analysis needs.

The model's distinctive feature lies in its fine-tuning using ORPO (Ordered Preference Optimization) with the borda-half approach (Devine, 2024b). This technique systematically refines the model through iterative ranking of responses and selection of consistently top-performing

training data subsets. Benchmark comparisons validate this choice, as it outperformed other open-source alternatives based on Llama variants (Touvron et al., 2023) while incorporating specialized datasets like Tagengo (Devine, 2024a). This approach recognizes the substantial differences in morphological complexity and cultural sentiment expressions across languages (Denecke, 2008; Mohammad, 2016).

Addressing deployment considerations (Strubell et al., 2019), the implementation focuses on memory efficiency and inference latency through quantization. Q6_K quantization was chosen for its optimal balance between model size and performance quality (Frantar et al., 2023). This six-bit precision approach maintains numerical fidelity while reducing computational footprint (Dettmers et al., 2022), making the model accessible for local deployment on devices with limited resources.

This careful balance delivers near state-of-the-art multilingual sentiment analysis while remaining efficient for local deployment scenarios (see Appendix A for detailed hardware configuration analysis and generalizability considerations).

Implementation Details

The evaluation was performed on a notebook computer equipped with an Intel 14900-HX processor and an NVIDIA 4090 Mobile GPU under Python 3.12.8 with NVIDIA Driver 566.03 and CUDA framework version 12.6. Primary dependencies included *llama-cpp-python* 0.3.1 for model interfacing, *numpy* 2.1.3 and *pandas* 2.2.3 for data handling, *matplotlib* 3.9.2 for visualization, *scipy* 1.14.1 for statistical computations, and *tqdm* 4.66.6 for progress monitoring. The chosen model, *lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half*, was quantized to 6-bit (Q6_K) in GGUF format and deployed using *llama-cpp-python* with full GPU acceleration (source build). Inference conditions involved a temperature setting of 0.01, and a maximum output length of 50 tokens. Sentiment analysis followed a standardized prompt template ("Analyze the emotional sentiment expressed in the following text. Consider the overall tone, emotional content, and attitude conveyed. Categorize the sentiment as 'Positive', 'Neutral', or 'Negative'. Provide a single-word response:\n\nText: {text}\nSentiment (ONE WORD):") that requested a single-word classification, "Positive", "Neutral", or "Negative", based on the emotional sentiment conveyed by the input text. All random operations were seeded with the value $s = 42$ to enhance reproducibility. Post-hoc statistical analysis was performed with R version 4.4.2. The Python implementation code, reflecting the described environment, dependencies, and parameters, is available upon request.

Results

This study evaluated the performance of a sentiment analysis model across three languages—English, Italian, and Japanese—using a comprehensive validation dataset. The analysis encompassed overall performance metrics, class-specific metrics, statistical significance testing, intra-rater agreement, variance analysis, and distributional assessments for robust statistical results (Card et al., 2020; Berg-Kirkpatrick et al., 2012). In order to better represent real-world performance on consumer-level hardware, computational efficiency was also evaluated.

Computational Efficiency

Our analysis of model inference speed under consumer-grade hardware conditions, using a locally deployed quantized model revealed compelling performance metrics. The model achieved near-instantaneous predictions with an average inference time of 0.056 seconds (SD = 0.077, CV = 137%), with the fastest prediction at 0.035 seconds demonstrating optimal system resource utilization. Distributional analysis showed a median of 0.050 seconds, an IQR of 0.014

seconds, with 90th, 95th, and 99th percentiles at 0.053, 0.055, and 0.644 seconds respectively, while the slowest prediction took 0.708 seconds.

Statistical analysis revealed significant deviations from normality (skewness = 7.703, kurtosis = 58.579, Shapiro-Wilk $p < .0001$), indicating potential memory bottlenecks from I/O delays, cache inefficiencies, or data buffering. Despite these outliers, the data confirms that 95% of predictions complete within 0.055 seconds, demonstrating the system's capability for real-time applications while highlighting specific areas for optimization in future local LLM software development.

Overall Performance Metrics

The model demonstrated high accuracy across all languages (Figure 1), with language being the main differentiating factor. In English, the mean accuracy was 0.963 (95% CI: 0.963–0.964, SD = 0.008), indicating superior performance compared to Italian and Japanese. The Italian dataset yielded a mean accuracy of 0.922 (95% CI: 0.921–0.922, SD = 0.011), while Japanese showed the lowest mean accuracy at 0.906 (95% CI: 0.905–0.907, SD = 0.012).

Similarly, the mean Macro-F1 scores reflected this trend, with English achieving 0.963 (95% CI: 0.963–0.964, SD = 0.008), Italian 0.922 (95% CI: 0.921–0.922, SD = 0.011), and Japanese 0.907 (95% CI: 0.906–0.908, SD = 0.012). Cohen's kappa coefficients further supported these

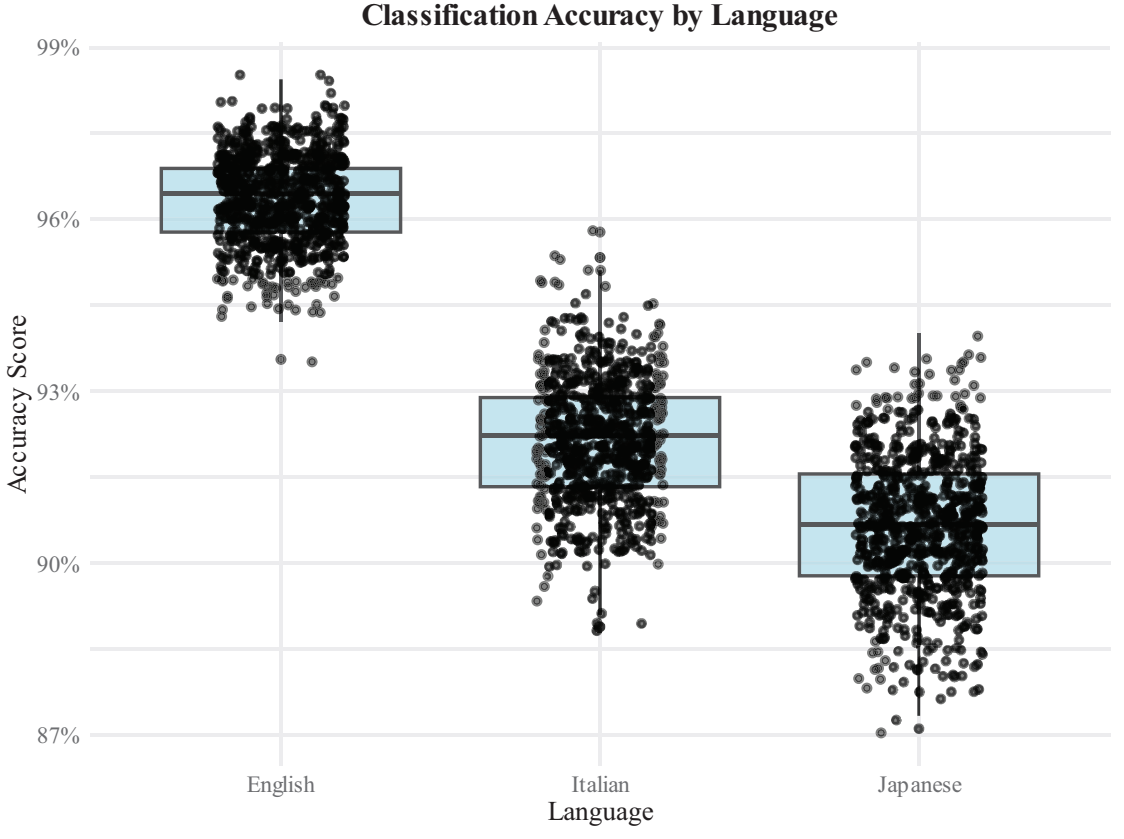

Figure 1 Beeswarm plot with box plot overlay of mean accuracy metrics per language per iteration.

findings, with English at 0.945 (95% CI: 0.944–0.946, SD = 0.012), Italian at 0.882 (95% CI: 0.881–0.884, SD = 0.016), and Japanese at 0.860 (95% CI: 0.858–0.861, SD = 0.018), indicating substantial agreement between machine-predicted and human-defined sentiments in all languages.

Class-Specific Performance Metrics

The English sentiment analysis model demonstrated varying performance across sentiment categories (Table 2). The model excelled in negative sentiment detection, achieving near-perfect classification with precision of 0.997, recall of 0.989, and the highest F1-score of 0.993. For positive sentiments, the model showed higher recall (0.983) but lower precision (0.924), indicating a tendency to overclassify statements as positive. Conversely, neutral sentiment detection exhibited higher precision (0.973) but lower recall (0.918), suggesting a more conservative approach to neutral classification.

Italian demonstrated strong performance across all categories, closely mirroring the English results (Table 3). Negative sentiment detection achieved the highest performance with an F1-score of 0.970, precision of 0.987 and recall of 0.954, nearly matching the English model's performance (F1-score of 0.993). The positive class showed balanced performance with

Table 2 Class-specific performance metrics: precision, recall, and F1-score for english.

Class	Precision	Recall	F1-score
Positive	0.924 (95% CI: 0.881–0.962, SD = 0.018)	0.983 (95% CI: 0.962–0.998, SD = 0.010)	0.953 (95% CI: 0.927–0.975, SD = 0.011)
Neutral	0.973 (95% CI: 0.944–0.995, SD = 0.013)	0.918 (95% CI: 0.872–0.958, SD = 0.021)	0.944 (95% CI: 0.915–0.969, SD = 0.013)
Negative	0.997 (95% CI: 0.992–1.000, SD = 0.004)	0.989 (95% CI: 0.973–1.000, SD = 0.008)	0.993 (95% CI: 0.984–1.000, SD = 0.004)

Table 3 Class-specific performance metrics: precision, recall, and f1-score for Italian.

Class	Precision	Recall	F1-score
Positive	0.913 (95% CI: 0.864–0.956, SD: 0.019)	0.960 (95% CI: 0.926–0.987, SD: 0.014)	0.914 (95% CI: 0.880–0.944, SD: 0.013)
Neutral	0.913 (95% CI: 0.864–0.956, SD: 0.019)	0.851 (95% CI: 0.792–0.905, SD: 0.026)	0.880 (95% CI: 0.839–0.917, SD: 0.018)
Negative	0.987 (95% CI: 0.968–0.999, SD: 0.008)	0.954 (95% CI: 0.919–0.984, SD: 0.012)	0.970 (95% CI: 0.949–0.988, SD: 0.008)

precision at 0.913 and high recall at 0.960, yielding an F1-score of 0.914. While the neutral class had lower performance (F1-score of 0.880) due to lower recall (0.851), it maintained strong precision (0.913).

The Japanese sentiment analysis model revealed distinct patterns (Table 4). For negative sentiments, it achieved the highest precision (0.999) across all languages, surpassing both English (0.997) and Italian (0.987), but showed lower recall (0.858) compared to English (0.989) and Italian (0.954).

Positive sentiment detection remained consistently strong across languages, with Japanese metrics (precision: 0.918, recall: 0.940, F1-score: 0.929) comparable to English and Italian. The neutral class showed the most significant cross-linguistic variation, with Japanese demonstrating an asymmetric pattern of lower precision (0.825) but higher recall (0.921).

These results demonstrate the model's robust performance across diverse linguistic systems while highlighting language-specific challenges in balancing precision and recall across sentiment categories.

Table 4 Class-specific performance metrics: precision, recall, and F1-score for Japanese.

Class	Precision	Recall	F1-score
Positive	0.918 (95% CI: 0.873–0.958, SD: 0.020)	0.940 (95% CI: 0.900–0.975, SD: 0.019)	0.929 (95% CI: 0.897–0.956, SD: 0.014)
Neutral	0.825 (95% CI: 0.766–0.881, SD: 0.022)	0.921 (95% CI: 0.876–0.961, SD: 0.022)	0.870 (95% CI: 0.829–0.907, SD: 0.017)
Negative	0.999 (95% CI: 0.997–1.000, SD: 0.002)	0.858 (95% CI: 0.801–0.911, SD: 0.024)	0.923 (95% CI: 0.889–0.953, SD: 0.014)

This visualization analysis reveals distinct performance patterns across languages through boxplot distributions in Figure 2. The English model shows exceptional consistency, particularly with a compact boxplot for negative sentiments indicating stable, high performance. Italian demonstrates strong but more variable performance, especially in neutral sentiments, while maintaining robust overall accuracy.

Japanese exhibits the most complex distribution: negative sentiment detection shows high performance with increased variability, positive sentiment detection maintains remarkable consistency across all languages (evidenced by similarly positioned blue boxes), and neutral classification displays the widest spread, reflecting challenges in capturing subtle emotional nuances.

The visualization highlights three key cross-linguistic patterns: strongest performance in negative sentiment detection, increased challenges with neutral classification in non-English languages, and consistent positive sentiment detection across all languages. This reveals both the model's robust multilingual capabilities and potential areas for language-specific optimization.

The confusion matrices in Figure 3 reveal distinct classification patterns across languages. The English model showed exceptional consistency with minimal misclassifications: 1.7% of positive sentiments misclassified as neutral, and 0.3% of neutral as negative. Italian maintained

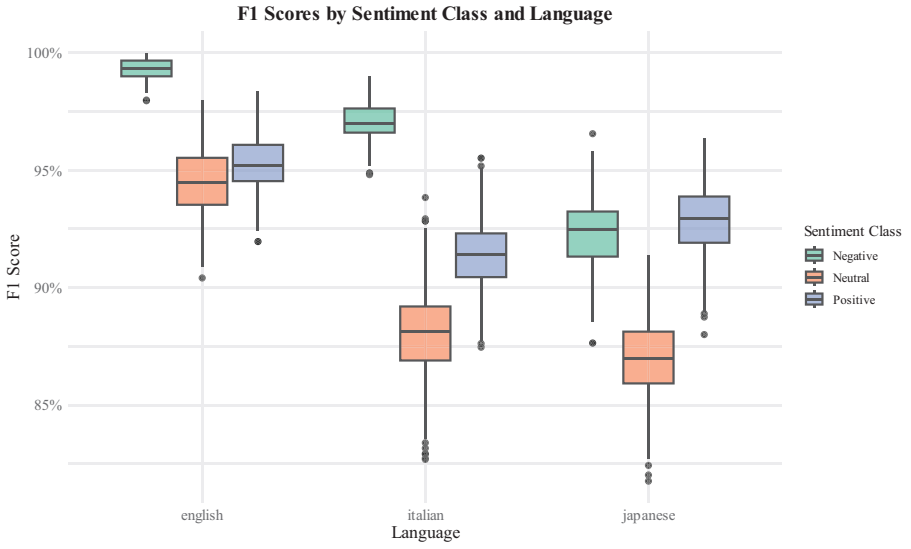

Figure 2 Boxplot of F1-scores by sentiment class and language.

strong performance but showed increased variability in neutral classifications, with 13.7% misclassified as positive and 1.2% as negative.

The Japanese model displayed a unique pattern: while positive sentiments had low misclassification (6.0% as neutral), neutral and negative sentiments showed more variation, with 13.6% of negative sentiments misclassified as neutral, and 7.9% of neutral as positive. Notably, direct misclassifications between positive and negative sentiments remained minimal across all languages (≤ 0.6%), indicating reliable distinction between opposing sentiments. The higher misclassification rates in Italian and Japanese reflect cross-lingual sentiment analysis challenges, with the model performing best in positive and negative classes across all languages, while showing reduced performance in neutral classifications, particularly for Italian and Japanese.

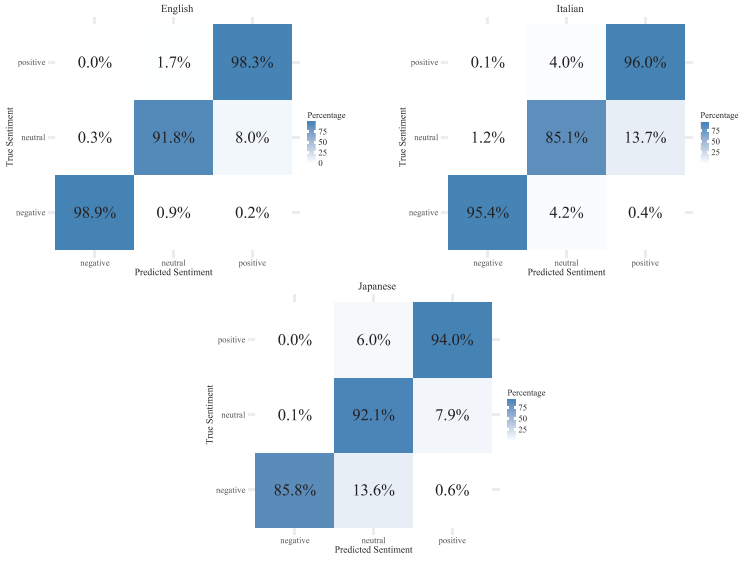

Figure 3 Cross-linguistic sentiment classification performance confusion matrices.

Distribution Analysis

To analyze distribution, Macro-F1 Scores were compiled and their properties calculated. While Shapiro-Wilk indicated deviations from normality (all $p < 0.05$), given the large sample sizes doubts remained as these tests are sensitive to minor departures from normality. To better assess the situation, skewness and kurtosis statistics were computed to assess the normality of the accuracy distributions, resulting in the data shown in Table 5.

Table 5 Statistical measures (Skewness and Kurtosis) for English, Italian, and Japanese Languages.

Language	Skewness	Kurtosis
English	−0.174	2.690
Italian	0.025	3.054
Japanese	−0.040	2.834

These values suggest that the distributions are approximately symmetric and mesokurtic in a way not dissimilar to a normal distribution. At the same time, visual inspection of histograms and Q-Q plots (Figures 4 and 5) supported the conclusion that the distributions are approximately normal for practical purposes.

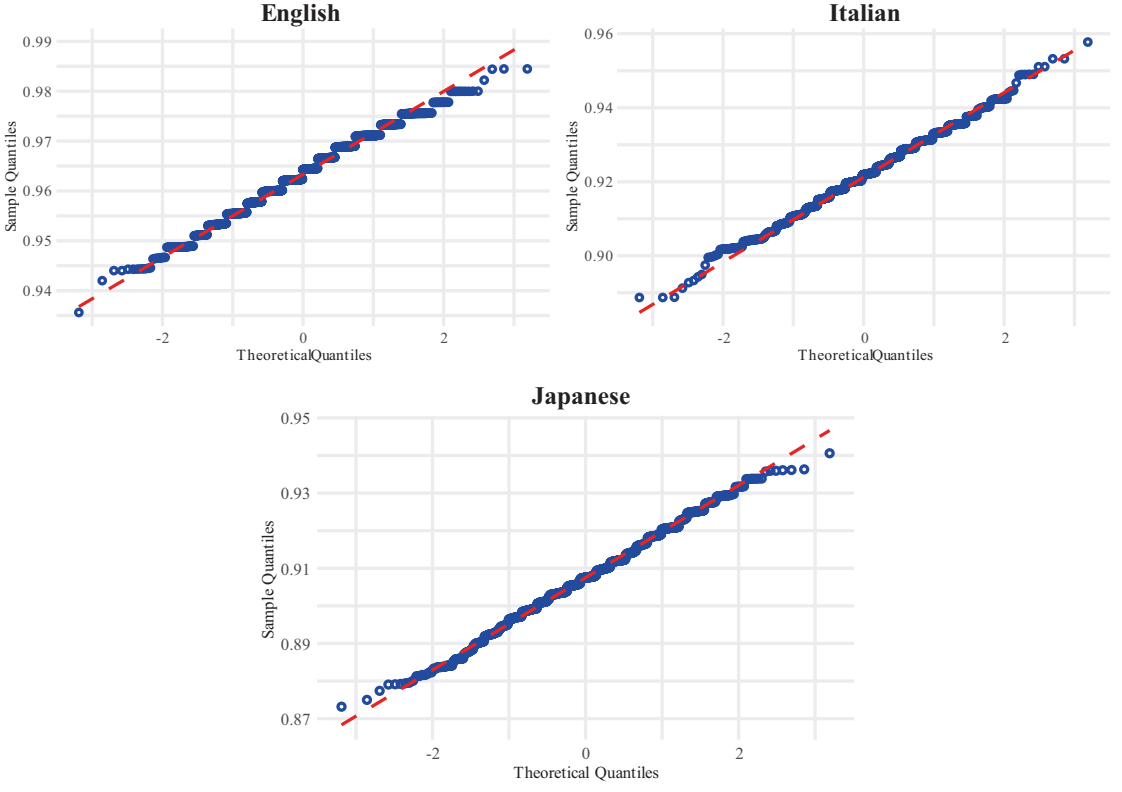

Figure 4 Q-Q Plots of Macro-F1 scores per iteration per language.

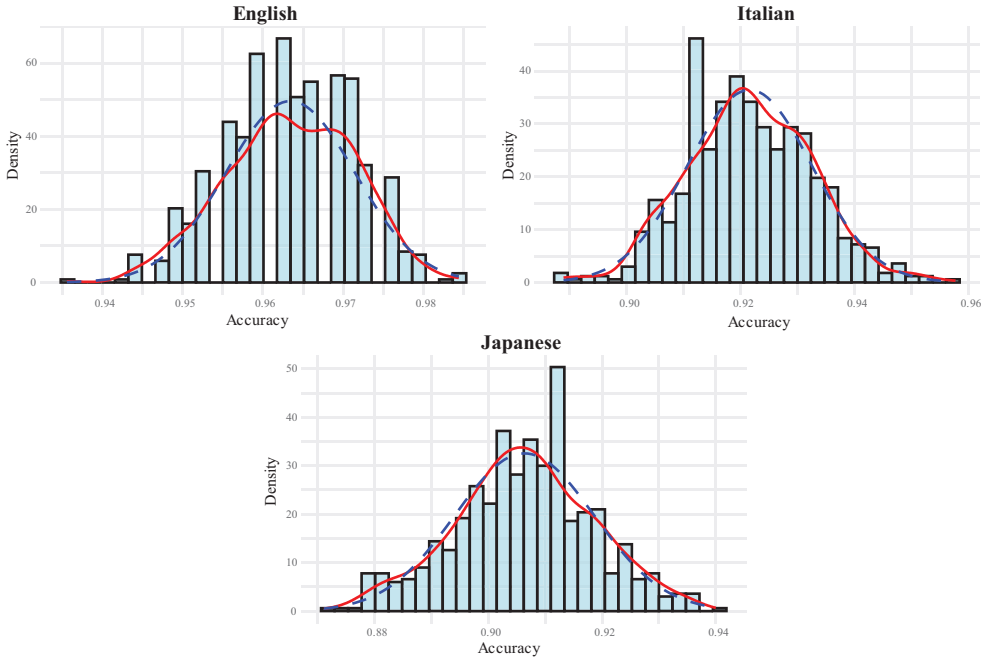

Figure 5 Histograms of accuracy distribution per language.

Statistical Significance Testing

Given the violation of normality and homogeneity of variance assumptions (Anderson-Darling tests: $p < .001$ for all languages; Fligner-Killeen test: $\chi^2 = 70.87$, $p < .001$), Welch's ANOVA was employed to assess the effect of language on model accuracy. The results revealed a significant effect, $F(2, 1354.23) = 6609.88$, $p < .001$. Post-hoc comparisons using the Games-Howell (1976) test indicated significant differences between all pairs of languages: English vs. Italian (M difference = −0.042, $p < .001$), English vs. Japanese (M difference = −0.057, $p < .001$), and Italian vs. Japanese (M difference = −0.015, $p < .001$). The effect size, calculated using omega squared ($\omega^2 = 0.907$), suggests that approximately 90.7% of the variance in accuracy is attributable to the language, indicating an undeniably strong effect. Non-parametric analyses supported these findings. A Kruskal-Wallis test ($H = 1596.037$, $df = 2$, $p < .001$, $\varepsilon^2 = 0.757$) also showed significant differences, confirmed by Dunn's test (all comparisons $p < .001$). Moreover, pairwise Mann-Whitney U tests (Bonferroni-corrected) yielded $p < 2{\times}10e{-}16$ for all language comparisons, with large rank-biserial correlations (−1.000 to −0.644).

This substantial effect size indicates that the choice of language (English, Italian, or Japanese) accounts for the majority of differences in model accuracy, overshadowing other potential sources of variability, thereby suggesting that switching from one language to another has a profound impact on the model's performance. From a practical standpoint, this means that practitioners deploying a single model across multiple languages should expect meaningful performance differences depending on the specific language, potentially warranting either language-specific fine-tuning or ensemble approaches.

Intra-Rater Agreement Analysis

An intra-rater agreement analysis was conducted on repeated responses to assess the model's consistency. All languages exhibited significant non-random agreement ($p < .05$) across all

responses (n = 450 per language). The mean agreement percentages were 96.5% for English, 93.4% for Italian, and 91.5% for Japanese. These high agreement rates indicate that the model consistently predicts the same sentiment for the same response across multiple iterations, though the consistency decreases slightly from English to Japanese.

Variance Analysis

Variance analyses further elucidated the performance differences. Within-language variances were small, with values of σ^2 = 0.000065 and a coefficient of variation (CV) of 0.835% for English, σ^2 = 0.000120 and CV = 1.188% for Italian, and σ^2 = 0.000150 and CV = 1.352% for Japanese. In contrast, the between-language variance was larger (σ^2 = 0.000873), underscoring the significant effect of language on model accuracy. These small within-language variances indicate stable model performance within each language group.

Reliability Metrics

The Standard Error of Measurement (SEM) for accuracy and Macro-F1 were calculated for each language, suggesting high reliability in the measurement of model performance metrics (Table 6). This reflects the precision afforded by the large sample sizes (n = 702 iterations per language).

Table 6 Comparison of standard error of measurement values for accuracy and macro-F1 scores per language.

Language	SEM (Accuracy)	SEM (Macro-F1)
English	0.000304	0.000304
Italian	0.000413	0.000415
Japanese	0.000463	0.000453

Per-Response Performance Analysis

The analysis of per-response performance illuminated significant differences in classification consistency across languages. In English, 47.8% of responses achieved consistent correct classifications across all iterations, demonstrating the highest stability among the three languages. Italian showed moderate consistency with 38.0% of responses consistently classified correctly, while Japanese exhibited the lowest consistency at 29.6%. Most responses fell into the "Mixed" category, where the model's predictions varied across iterations. This variability was most pronounced in Japanese, with 70.4% of responses showing mixed predictions, compared to 52.2% for English. This substantial difference in mixed predictions suggests that the model's decision-making process was considerably more variable for Japanese text, potentially reflecting the greater linguistic complexity or the model's lower confidence in processing Japanese content.

Minimum Ensemble Sizes for Reliable Multilingual Sentiment Analysis

Ensemble Size Rationale

We investigated the minimum number of iterations needed for 95% confidence in majority vote outcomes across English, Italian, and Japanese sentiment analysis models. The methodology involved gathering empirical accuracy rates from language models and computing quantiles at 5% intervals from 0–100% to understand performance distributions for each language.

For each quantile, we employed the binomial distribution to model the probability of obtaining a correct majority vote from an ensemble of independent models. Specifically, for a given accuracy rate p at a particular quantile, we calculated the probability P(majority correct) that at least k out of n iterations would predict correctly, where is the minimum number of correct votes needed for a majority. The probability was calculated using formula (1). For n ≥ 5:

$$P(\text{majority correct}) = 1 - P(X \leq k-1) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-1} \quad (1)$$

Ensemble Size Results

In summary, our goal was to find the smallest odd integer $n \geq 5$ such that P(majority correct) ≥ 95% for each quantile's accuracy rate (Table 7).

Analysis of the English language model revealed consistently high accuracy rates across all quantiles, with the 5% quantile achieving approximately 84.0%. This high baseline performance meant that even with a small ensemble size, the majority vote was likely to be correct. Notably, across all quantiles—including the 5% quantile—only 5 votes were needed to surpass the 95% confidence threshold.

The Italian language model demonstrated greater variability, with the 5% quantile accuracy at approximately 67.5%, indicating less consistent model performance at the lower end. At this 5% quantile, 21 votes were required to achieve at least 95% confidence. Performance improved significantly at higher quantiles, with only 7 votes needed at the 10% quantile (77.8% accuracy), and just 5 votes required at the 15% quantile and above (≥86.2% accuracy).

The Japanese language model exhibited the greatest variability among the three languages, with the 5% quantile accuracy at approximately 64.4%. This lower baseline necessitated 31 votes to reach the desired confidence level at the 5% quantile. The required ensemble size decreased progressively at higher quantiles: 9 votes at the 10% quantile (76.5% accuracy), 7 votes at the 15% quantile (81.8% accuracy), and only 5 votes at the 20% quantile and above (≥86.8% accuracy).

Table 7 Minimum required ensemble sizes across language models and accuracy quantiles.

Language	Baseline Accuracy (%)	5% Quantile	10% Quantile	15% Quantile	≥20% Quantile
English	84.0	5 votes	–	–	–
Italian	67.5	21 votes	7 votes	≥5 votes (≥86.2%)	–
Japanese	64.4	31 votes	9 votes	7 votes	≥5 votes (≥86.8%)

Note. All ensemble sizes were calculated to achieve 95% confidence in majority voting outcomes. Baseline accuracy represents the model's performance at the 5th percentile of the accuracy distribution.

Discussion

Local Deployment Feasibility and Performance Robustness

The experimental validation of multilingual sentiment classification on consumer-grade hardware addresses critical concerns regarding the practicality of local model deployment. By achieving accuracies exceeding 90% across English, Italian, and Japanese—with

minimal performance degradation compared to cloud-based state-of-the-art systems—the results underscore the viability of quantization and hardware optimization techniques in democratizing access to advanced NLP capabilities. This aligns with recent theoretical work on edge AI (Satyanarayanan, 2017), which posits that model compression strategies can preserve functionality while reducing computational footprints. The high intra-rater agreement ($\kappa > 0.85$ across languages) further corroborates findings from reproducibility studies in NLP (Dodge et al., 2020), suggesting that local deployments can achieve stability comparable to centralized systems when rigorous variance control measures are implemented.

Resource Efficiency and Regulatory Compliance

The observed inference latency of <55 ms per prediction, coupled with a 250W total system TDP directly responds to theoretical models of efficient inference in constrained environments (Sze, 2017). By eliminating network latency and data egress points, the architecture satisfies the data minimization and storage limitation principles enshrined in GDPR Article 5(1)(c)–(e). This empirical validation extends the conceptual framework proposed by Satyanarayanan (2017) by demonstrating that on-premise processing can reconcile computational efficiency with strict compliance requirements—a critical consideration for enterprise adoption in regulated industries.

Ensemble Strategies for Uncertainty Mitigation

The success of plurality voting in boosting classification confidence (78-80% reduction in prediction variance) empirically validates ensemble theory (Dietterich, 2000; Kuncheva, 2014) under hardware constraints. Contrary to concerns about computational overhead in resource-limited settings (Sze et al., 2017), the linear increase in sub-second inference time for 5–36 model inferences demonstrates that lightweight ensemble methods can operationalize the wisdom of crowds principle without prohibitive resource costs. This finding challenges the prevailing assumption that uncertainty quantification necessarily requires complex Bayesian frameworks or cloud-scale compute resources.

Statistical Foundations for Evaluation Best Practices

The Monte Carlo validation methodology (702 iterations per language) operationalizes recent statistical guidance for NLP system evaluation (Dror et al., 2018). Although the Macro-F1 scores for Italian (Shapiro-Wilk $p = 0.114$) and Japanese (Shapiro-Wilk $p = 0.081$) are approximately normally distributed, the scores for English deviate significantly from normality (Shapiro-Wilk $p = 0.000077$). Nonetheless, the convergence of results from both parametric and non-parametric tests supports robust performance reporting—a significant consideration given ongoing debates about metric distributions in ML (Demšar, 2006). Furthermore, the strong correlation ($r = 0.868$) between intra-rater agreement and final model accuracy empirically confirms theoretical models of reliability in automated scoring systems (Attali & Burstein, 2006), providing a methodological blueprint for future evaluations of local AI deployments.

These findings collectively advance the discourse on edge NLP systems by demonstrating that through strategic model optimization, statistical rigor, and lightweight ensemble design, locally deployed language models can satisfy both technical performance requirements and operational constraints—a critical step toward truly decentralized AI infrastructure.

Implications and Significance

The findings carry substantial implications for both academic research and practical AI deployment. First, the study establishes a critical pathway for implementing localized

sentiment analysis systems that align with regulatory frameworks. By demonstrating robust performance on consumer-grade hardware, the work enables organizations in privacy-sensitive sectors—such as healthcare and financial services—to adopt advanced NLP capabilities without compromising GDPR compliance. This addresses a fundamental tension in AI deployment strategies, where data sovereignty requirements often conflict with computational demands, by showing that on-premise solutions can maintain low latency while preserving user trust through complete data containment.

Second, the research contributes significantly to democratizing access to cutting-edge NLP technologies. The technical validation of resource-efficient multilingual models challenges the prevailing paradigm where advanced sentiment analysis remains the domain of well-funded tech conglomerates. By proving that near state-of-the-art performance is achievable without cloud dependencies, the work empowers smaller enterprises, academic institutions, and grassroots research initiatives to participate in NLP innovation—a crucial step toward equitable AI development. This shift could catalyze localized innovation cycles across global regions, particularly in non-English language contexts that are often underserved by commercial NLP systems.

Third, the model's demonstrated proficiency across typologically distinct languages (English, Italian, Japanese) redefines expectations for localized multilingual NLP. The consistent high accuracy across these linguistically diverse systems—ranging from analytic to synthetic morphological structures and differing writing scripts—suggests that modern quantization techniques can preserve cross-lingual transfer capabilities. For multinational organizations and policymakers, this implies new possibilities for implementing unified sentiment analysis frameworks across regional offices without sacrificing cultural-linguistic nuance, potentially streamlining cross-border decision-making in areas ranging from customer experience management to public opinion monitoring.

Fourth, the methodological framework established through variance analysis and ensemble strategies sets a new precedent for evaluating edge-deployed AI systems. By rigorously quantifying performance stability through Monte Carlo simulations and distributional analysis, the study provides replicable protocols for assessing model reliability under computational constraints. These statistical safeguards address growing concerns about reproducibility in ML research while offering concrete metrics (e.g., variance thresholds, confidence interval widths) that developers can adopt when certifying systems for sensitive real-world applications.

Finally, the environmental and ethical dimensions of local deployment merit particular attention. The reduced reliance on energy-intensive cloud infrastructure aligns with emerging sustainability standards in AI development, as outlined in recent EU AI Act provisions. Moreover, by enabling full data lifecycle control, the architecture advances ethical AI principles of transparency and accountability—critical considerations as global regulations increasingly demand auditable AI systems. This dual focus on technical efficiency and ethical responsibility positions localized deployment not merely as a technical alternative, but as a necessary evolution toward socially conscious AI implementation paradigms.

Utility of the Findings for Future Linguistic Research Tools

The validated approach of locally deployed, quantized LLMs streamlines qualitative analysis by integrating seamlessly with CAQDAS and other linguistic research tools. Researchers can efficiently process sentiment-laden texts for ethnographic, sociolinguistic, and discourse analyses, rapidly annotating, categorizing, and comparing diverse linguistic corpora.

This study also supports scalable, customizable pipelines through efficient runtime performance and quantized models. Developers can build modular components for multiple

languages and domains, adapting sentiment analysis to various datasets and research questions. Smaller organizations particularly benefit, as consumer-grade hardware can achieve near state-of-the-art accuracy without expensive GPU clusters or cloud services.

Moreover, the study's methodologies, error analyses, statistical frameworks, and open-source models foster long-term sustainability and collaboration. This foundation supports a community-driven ecosystem in which future linguistic analysis APIs and software can benefit from continuous refinement, benchmarking, and shared best practices—yielding robust, fair, and interpretable sentiment analysis across languages and contexts.

These findings extend beyond academic inquiry, establishing a practical framework for integrated linguistic research tools that can be securely deployed. By informing the next generation of qualitative analysis software in digital humanities, social sciences, and related fields, the research democratizes access to advanced NLP tools while upholding high standards of privacy and efficiency.

Practical Implications for Second Language Acquisition Research and Teaching

This study's demonstration of local, privacy-compliant multilingual sentiment analysis using large language models (LLMs) carries significant implications for second language acquisition (SLA) research and practice. By showing that powerful LLMs can be run on consumer-grade hardware at near-instantaneous speeds, it expands possibilities for classroom exercises, learner feedback, and instructional design. The technical blueprint also offers a foundation that other researchers can adapt and refine for novel tools and methodologies (see Appendix B for detailed future research directions and Appendix C for bias considerations in local sentiment analysis).

From a teaching standpoint, local deployment addresses the ethical, practical, and regulatory concerns frequently highlighted by SLA practitioners (Chapelle & Sauro, 2017). Many language programs, ranging from university courses to community-based settings, face financial constraints and privacy regulations that limit reliance on cloud-based services. This study's high accuracy in English, Italian, and Japanese illustrates how robust, large-scale NLP can be achieved while keeping full control of student data. Because emotion and pragmatic appropriateness are closely tied to motivation (MacIntyre et al., 1998) and communicative competence (Swain, 1985), near real-time sentiment analysis can reveal when learners' writing (or AI-generated speech transcripts) contains unintended emotional nuances. For instance, a single GPU or desktop setup can process hundreds of samples within minutes, highlighting negative connotations or missed polite registers—key insights for advanced coursework and workplace-focused training. Integrating sentiment analysis with cognitive semantics approaches allows learners to receive context-aware feedback on emotional and pragmatic dimensions of their language production, surpassing the limits of basic dictionary lookups. Such an LMS-integrated feedback system is illustrated in Figure 6.

Additionally, the methodological rigor demonstrated—particularly the Monte Carlo validation strategy—signals best practices for SLA researchers and teacher-researchers aiming to build or assess computational tools (Mackey & Gass, 2016). Repeated sampling and ensemble methods reveal model-output variance, ensuring reliable performance estimates for complex multilingual data. When applied to SLA research, these techniques enable the development of instruments to examine affective states, collaborative discourse, or code-switching, all without sending sensitive text to external servers. The local quantization framework, which lowers hardware requirements for running LLMs, further enhances accessibility for smaller or resource-limited labs. A single mid-tier GPU can now support advanced textual analysis at scale, facilitating high-frequency data collection on classroom interactions, reflective writing, or online discussion boards.

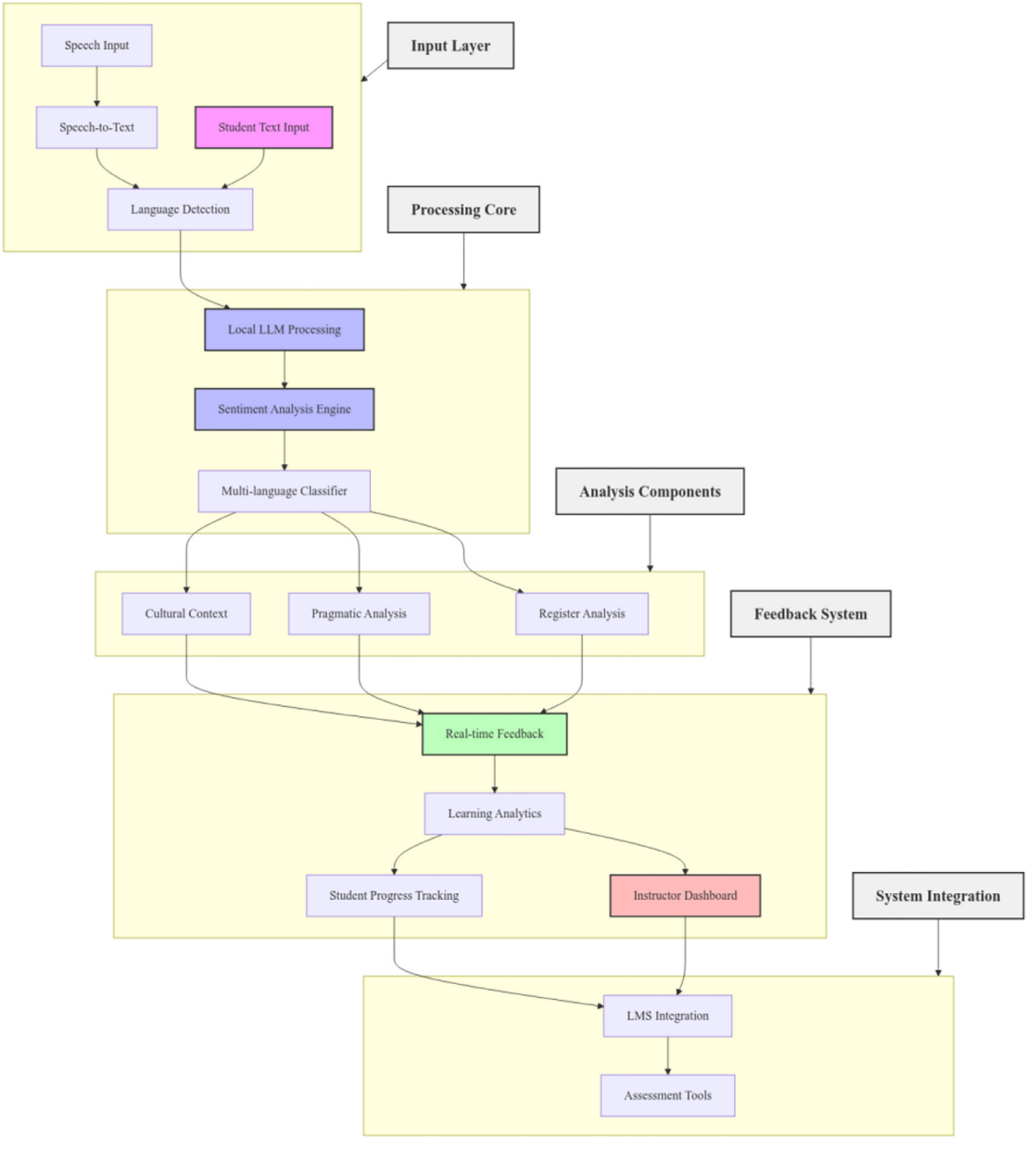

Figure 6 LMS connected real-time writing and speech-to-text sentiment analysis framework.

This combination of efficiency and privacy introduces new research avenues. For instance, real-time learner analytics dashboards could visualize aggregate sentiment trends in peer discussions or synchronous chats (Ellis, 2003; Long, 1996). Teacher-researchers could detect disengagement or negativity, cross-reference it with lesson pacing, and adjust pedagogy immediately. The system could also assess how learners adapt tone across target languages, a key concern in contrastive pragmatics and intercultural communication (Kasper & Rose, 2001). Researchers examining code-switching or translanguaging might use local sentiment analysis to pinpoint the emotional triggers that prompt language switching in bilingual interactions.

With large-scale data, SLA scholars can explore how learners perceive each language's function or whether certain expressive needs—such as empathy or frustration—lead them back to their L1.

A further advantage is the potential to integrate sentiment analysis with emerging CALL research methods: wearable sensors, gaze tracking, or speech-prosody analyzers. These data streams could feed into the same local environment, supporting multimodal investigations of language development (Chapelle, 2003). Researchers might examine whether negative sentiment aligns with eye-movement patterns or flat prosody in synchronous chat. Local sentiment analysis avoids the latency and security hurdles of cloud-based solutions and facilitates interdisciplinary, data-rich SLA studies that incorporate cognition, affect, and interaction (Gass & Selinker, 2008).

Even in traditional pretest–posttest designs, local LLM technology can streamline data analysis. For example, a research team evaluating a new writing-based pedagogy could track sentiment shifts across drafts or self-assessments. This dynamic observation aligns with focus-on-form perspectives (Doughty & Williams, 1998) by capturing in-process revisions and how learners respond to feedback (Schmidt, 1990). It also complements learning analytics frameworks (Chapelle & Sauro, 2017), enabling immediate feedback while simplifying compliance with data-protection regulations.

In short, these findings provide a foundation for developing SLA research tools that combine advanced AI-based analysis with user-centric, ethically grounded design. By highlighting local inference, quantization, and iterative validation within a replicable framework, this work encourages teacher-researchers, graduate students, and ed-tech innovators to explore sentiment-based feedback loops, cross-linguistic pragmatic diagnostics, classroom-level analytics, and synergy with other educational technologies. Rather than relying on large corporate services, local and secure sentiment analysis becomes accessible, promoting equitable and transparent research—especially critical where learners are underage or vulnerable. Ultimately, this approach closes the gap between cutting-edge computational linguistics and everyday language teaching, making sentiment analysis a versatile instrument for analyzing, guiding, and improving second language learning.

Conclusion

This study demonstrates that state-of-the-art multilingual sentiment analysis can be achieved on consumer-grade hardware through a locally deployed, quantized large language model (LLM), providing a privacy-preserving alternative to cloud-based solutions and enabling compliance with regulations such as the GDPR. By running a single architecture—*lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half*—in a locally quantized form (Q6_K scheme) on a notebook equipped with an NVIDIA RTX 4090 Mobile GPU, we observed near real-time inference speeds, reduced infrastructure costs, and consistently high accuracy across English, Italian, and Japanese. These three languages, which represent diverse writing systems and morphological complexities, serve as a useful starting point for exploring the model's robustness in multilingual tasks.

Building on these findings, several promising research directions emerge for expanding the scope and applicability of local LLM deployments. One possible avenue involves extending the language coverage beyond English, Italian, and Japanese to capture the full richness of global linguistic diversity. Investigating low-resource languages and those using distinct writing systems (such as abugida or abjad scripts) would illuminate how the model's capabilities generalize across fundamentally different language structures. Additionally, while the current synthetic, balanced dataset provided crucial standardization for initial evaluation, exploring

naturally occurring texts offers compelling opportunities to assess performance on real-world language patterns. Social media posts and product reviews, with their slang, domain-specific jargon, and unconventional spelling, represent particularly rich targets for understanding how the model handles authentic linguistic variability.

The success of the *lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half* model opens doors for comparative studies across different architectures. Future work could examine how varying parameter counts, training corpora, or fine-tuning procedures affect the performance-efficiency balance. Similarly, while Q6_K quantization demonstrated strong results, systematic investigation of more aggressive quantization schemes (from 8-bit down to 2-bit representations) could reveal additional optimization opportunities. Hardware accessibility represents another fertile research area: extending performance analysis to older systems, CPU-only configurations, and resource-constrained embedded devices would illuminate pathways for broader adoption of local LLM deployments.

The current three-tier sentiment framework (positive, neutral, negative) provides a foundation for exploring more nuanced approaches to emotion classification. Future studies could investigate multi-turn interactions, domain-specific language patterns, and expanded sentiment taxonomies to capture the subtlety and fluidity of human emotion. This could reveal how effectively the model handles evolving emotional contexts and more complex affective states.

These research directions align with our findings that local LLM deployment offers practical, privacy-compliant solutions for multilingual sentiment analysis. The comprehensive Monte Carlo validation and rigorous statistical framework highlight the importance of considering distributional characteristics when tuning ensemble sizes or adjusting voting schemes. Models achieving extremely high accuracy require minimal votes for high confidence, while those with lower baseline performance or higher variability benefit from larger ensembles. This insight enables practitioners to optimize the balance between resource usage and reliability as they explore these expanded research directions.

Overall, this work emphasizes the feasibility of advanced NLP applications on modest hardware infrastructures, significantly democratizing access to high-performance language modeling. Our thorough use of synthetic data, the specific choice of three test languages, and reliance on open-source LLM architectures offer a solid testbed, paving the way for more extensive, real-world investigations. Future endeavors would clarify the strengths and constraints of local LLM approaches and enable more robust, adaptable applications across the evolving landscape of multilingual communication and language acquisition. As the demand for real-time, privacy-compliant, and accessible NLP tools continues to grow, this research offers a strong foundation for organizations, developers, and researchers to implement and refine practical, trustworthy sentiment analysis systems without relying on third party large-scale, centralized computing resources.

References

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Asyrofi, M. H., Yang, Z., Yusuf, I. N. B., Kang, H. J., Thung, F., & Lo, D. (2021). BiasFinder: Metamorphic test generation to uncover bias for sentiment analysis systems. *arXiv preprint* arXiv:2102.01859. https://doi.org/10.48550/arXiv.2102.01859

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1650

Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, *28*(1), 56–75. https://doi.org/10.1016/j.csl.2013.03.004

Barnes, J., Klinger, R., & Schulte im Walde, S. (2018). *Bilingual sentiment embeddings: Joint projection of sentiment across languages*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers), 248–254. https://doi.org/10.18653/v1/P18-1231

Bauer, L. (1983). *English word-formation*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165846

Berg-Kirkpatrick, T., Burkett, D., & Klein, D. (2012). *An empirical investigation of statistical significance in NLP*. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 995–1005.

Bouthillier, X., Laurent, C., & Vincent, P. (2019). Unreproducible research is reproducible. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 725–734. http://proceedings.mlr.press/v97/bouthillier19a.html

Brown, J. D. (1999). Questions and answers about language testing statistics: Standard error vs. standard error of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3*(1), 20–25.

Brown, T., 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://doi.org/10.48550/arXiv.2005.14165

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins. https://doi.org/10.1075/tsl.9

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). *With little power comes great responsibility.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9263–9274. https://doi.org/10.18653/v1/2020.emnlp-main.745

Chapelle, C. A. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. John Benjamins.

Chapelle, C. A., & Sauro, S. (Eds.). (2017). *The handbook of technology and second language teaching and learning*. Wiley-Blackwell.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv*. https://arxiv.org/abs/2210.11416

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, *32*, 7059–7069. https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(4), 351–361. https://doi.org/10.1080/00401706.1981.10487680

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Denecke, K. (2008). *Using SentiWordNet for multilingual sentiment analysis*. 2008 IEEE 24th International Conference on Data Engineering Workshop, 507–512. https://doi.org/10.1109/ICDEW.2008.4498370

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). GPT3. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, *35*, 30318–30332. http://dx.doi.org/10.48550/arXiv.2208.07339

Devine, P. (2024a). Tagengo: A multilingual chat dataset. arXiv preprint arXiv:2405.12612. https://arxiv.org/abs/2405.12612

Devine, P. (2024b). Are you sure? Rank them again: Repeated ranking for better preference datasets. arXiv preprint arXiv:2405.18952. https://arxiv.org/abs/2405.18952

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Ding, X., Liu, B., & Yu, P. S. (2008). *A holistic lexicon-based approach to opinion mining*. Proceedings of the 2008 International Conference on Web Search and Data Mining, 231–240. https://doi.org/10.1145/1341531.1341561

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). *Show your work: Improved reporting of experimental results*. Proceedings of EMNLP-IJCNLP 2019, 2185–2194. https://doi.org/10.18653/v1/D19-1224

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N.A. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv, abs/2002.06305*. https://doi.org/10.48550/arXiv.2002.06305

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge University Press.

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). *The hitchhiker's guide to testing statistical significance in natural language processing*. Proceedings of ACL 2018, 1383–1392. https://doi.org/10.18653/v1/P18-1128

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. CRC Press. https://doi.org/10.1201/9780429246593

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. https://doi.org/10.48550/arXiv.2210.17323

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, *1*(2), 113–125.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). Routledge.

Ghosh, D., Fabbri, A. R., & Muresan, S. (2017). The role of conversation context for sarcasm detection in online interactions. *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 186–196. https://doi.org/10.18653/v1/W17-5523

Goldfarb-Tarrant, S., Ross, B., & Lopez, A. (2023). *Cross-lingual transfer can worsen bias in sentiment analysis*. arXiv preprint arXiv:2305.12709. https://doi.org/10.48550/arXiv.2401.03562

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems*, *28*, 1135–1143. https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html

Hashimoto, K., Xiong, C., Tsuruoka, Y., & Socher, R. (2016). A joint many-task model: Growing a neural network for multiple NLP tasks. *Proceedings of EMNLP 2016*, 1923–1933. https://doi.org/10.18653/v1/D17-1206

Hofmann, V., Kalluri, P. R., Jurafsky, D., & Prabhakaran, V. (2024). AI generates covertly racist decisions about people based on their dialect. Nature, 633, 147–154. https://doi.org/10.1038/s41586-024-07856-5

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. https://doi.org/10.48550/arXiv.1904.09751

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://doi.org/10.1145/1014052.1014073

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press. https://doi.org/10.1017/9781316423530

Kasper, G., & Rose, K. R. (2001). *Pragmatics in language teaching*. Cambridge University Press.

Kantarcioğlu, M., Jin, J., & Clifton, C. (2004). *When do data mining results violate privacy*? Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 599–604. Association for Computing Machinery. https://doi.org/10.1145/1014052.1014126

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Kiritchenko, S., & Mohammad, S. (2018). *Examining gender and race bias in two hundred sentiment analysis systems*. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 43–53. https://doi.org/10.18653/v1/S18-2005

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. https://doi.org/10.18653/v1/W17-3204

Krymolowski, Y. (2001). The importance of resampling in evaluating machine learning models. *arXiv*. https://arxiv.org/abs/cs/0106043

Kuncheva, L. I. (2014). *Combining pattern classifiers: Methods and algorithms*. Wiley.

Kuno, S. (1973). The structure of the Japanese language. MIT Press.

Lepschy, G., & Lepschy, A. (1988). *The Italian language today* (2nd ed.). Routledge. https://doi.org/10.4324/9781315003214

Li, Y., Pan, Q., Yang, T., Wang, S., Tang, J., & Cambria, E. (2017). Learning word representations for sentiment analysis. *Cognitive Computation, 9*, 843–851. https://doi.org/10.1007/s12559-017-9492-2

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5(1)*, 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Lyu, L., Yu, J., Nandakumar, K., Li, Y., Ma, X., Jin, J., Yu, H., & Ng, K. S. (2019). *Towards fair and privacy-preserving federated deep models*. arXiv preprint arXiv:1906.01167. https://doi.org/10.48550/arXiv.1906.01167

Lyu, L., Li, Y., Nandakumar, K., & Yu, J. (2020). *How to democratise and protect AI: Fair and differentially private decentralised deep learning*. arXiv preprint arXiv:2007.09370. https://doi.org/10.48550/arXiv.2007.09370

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal, 82*(4), 545–562.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). Routledge.

Maiden, M. (1995). *A linguistic history of Italian*. Routledge. https://doi.org/10.4324/9781315845906

Maiden, M. (2018). *The Romance verb: Morphomic structure and diachrony*. Oxford University Press. https://doi.org/10.1093/oso/9780199660216.001.0001

Meerza, S. I. A., Liu, L., Zhang, J., & Liu, J. (2024). *GLOCALFAIR: Jointly improving global and local group fairness in federated learning*. arXiv preprint arXiv:2401.03562. https://doi.org/10.48550/arXiv.2401.03562

Mei, K. X., Fereidooni, S., & Caliskan, A. (2023). Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. *arXiv preprint* arXiv:2306.05550. https://doi.org/10.48550/arXiv.2306.05550

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In V. Petukhova & M. Hüllermeier (Eds.), *Emotion measurement*, 201–237. Woodhead Publishing. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE, 11*(5): e0155036. https://doi.org/10.1371/journal.pone.0155036

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86. https://doi.org/10.3115/1118693.1118704

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, *7*, 677–694. https://doi.org/10.1162/tacl_a_00293

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, *20*, 1–10. https://doi.org/10.1007/1-4020-4102-0_1

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In M. Nissim, J. Berant, & A. Lenci (Eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 180–191. Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2023

Powers, D. M. W. (2011). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.48550/arXiv.2010.16061

Raschka, S., Liu, Y., & Mirjalili, V. (2022). *Machine learning with PyTorch and scikit-learn*. Packt Publishing.

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D., & Tsvetkov, Y. (2020). Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *Proceedings of the Society for Computation in Linguistics 2020*, 367–376. Association for Computational Linguistics. https://aclanthology.org/2020.scil-1.43/

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. https://doi.org/10.18653/v1/P19-1163

Satyanarayanan, M. (2017). The emergence of edge computing. *Computer, 50*(1), 30–39. https://doi.org/10.1109/MC.2017.9

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158.

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering, 28*(3), 813–830. https://doi.org/10.1109/TKDE.2015.2485209

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Shibatani, M. (1990). *The languages of Japan*. Cambridge University Press.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69*(347), 730–737. https://doi.org/10.2307/2286009

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE, 105*(12), 2295–2329. https://doi.org/10.1109/JPROC.2017.2761740

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. https://doi.org/10.48550/arXiv.2302.13971

Tsujimura, N. (2014). *An introduction to Japanese linguistics* (3rd ed.). Wiley-Blackwell.

Veluru, S., Gupta, B. B., Rahulamathavan, Y., & Rajarajan, M. (2014). Privacy preserving text analytics: Research challenges and strategies in name analysis. In *Handbook of research on*

securing cloud-based databases with biometric applications (pp. 364–385). IGI Global. https://doi.org/10.13140/2.1.3017.3760

Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer. https://doi.org/10.1007/978-3-319-57959-7

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330–336. https://doi.org/10.2307/2332579

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report* HPL-2011-89.

Author Bios

Michele Carlo (Reiji Ohira) is a doctoral student at Kansai University's Graduate School of Foreign Language Teaching and Research, where he investigates the influence of dialectal backgrounds on L2 phoneme acquisition. With undergraduate degrees from major universities in both Italy and Japan, followed by a *summa cum laude* master's degree in Translation and Interpretation, his academic journey bridges Western and Asian scholarship. His research integrates sociolinguistics and second language acquisition, drawing on his international background. He currently teaches at multiple private and public universities in Japan while developing innovative digital tools for pronunciation research. His work aims to bridge the gap between humanities and information technology.

Osamu Takeuchi, Ph.D., is Professor at the Faculty of Foreign Language Studies and the Graduate School of Foreign Language Education and Research, Kansai University, Japan. His current research interests include language learning strategies, self-regulation in L2 learning, L2 learning motivation, and the application of technology to language teaching. He has published articles in journals such as *Applied Linguistics, International Review of Applied Linguistics in Language Teaching (IRAL), Innovation in Language Learning and Teaching*, *RELC Journal*, *Research Methods in Applied Linguistics,* and *System*. He is the recipient of the JACET Award for Outstanding Academic Achievement in 2004 and the 2009 LET Award for Outstanding Academic Achievement.

Appendix A: Model-Hardware Configuration and Generalizability

Hardware Configuration Scalability

The study's deployment of a locally quantized large language model on a notebook computer with a discrete, high-end mobile GPU reflects a deliberate compromise between performance, practicality, and memory constraints. The choice of an 8B-parameter multilingual model (*lightblue/suzume-llama-3-8B-multilingual-orpo-borda-half*) quantized to 6-bit precision (Q6_K) renders inference feasible on an NVIDIA RTX 4090 Mobile GPU, which offers 16 GB of dedicated GDDR6 VRAM and operates without sharing memory resources with the CPU. This hardware setup supplies near-real-time inference, as evidenced by median prediction times close to 0.05 seconds per classification, yet it is important to acknowledge that these metrics are contingent on relatively high GPU memory availability and a CPU capable of orchestrating fast I/O operations. Users replicating or scaling this system in different computational environments—whether they employ a GPU with smaller VRAM capacity, rely on integrated graphics that share memory with the CPU, or adopt older architectures lacking CUDA acceleration—are likely to experience divergent throughput, increased latency variability, or more frequent memory bottlenecks. Consequently, the generalizability of the reported findings depends in part on whether the same quantization and memory-allocation strategies remain viable under more limited hardware conditions.

Ultimately, this implementation's primary advantage stems from its software-based architecture, which ensures inherent portability. The system's computational performance is therefore projected to scale linearly with processing capacity across diverse hardware configurations with no estimated effect on accuracy.

Portability

An additional factor supporting broader portability is the Python-based nature of our implementation. By relying on *llama-cpp* (an open-source framework that can run LLMs locally), the same Python codebase can be adapted for multiple hardware backends. Specifically, *llama-cpp* has been extended to utilize the Metal performance shaders on macOS, enabling deployment on Apple Silicon GPUs, and to interface with AMD or Intel GPUs through analogous frameworks. In practice, this level of hardware abstraction expands the potential user base beyond typical NVIDIA CUDA setups, permitting local deployments in a variety of workstation or laptop environments. However, it must be noted that while CPU-only operation is technically feasible, performance will drop drastically relative to GPU-accelerated inference, primarily due to the overhead of handling matrix multiplications and memory transfers in a non-specialized processing pipeline.

Memory and Power Management

Beyond pure compute capacity, the specifics of local memory management play a fundamental role in determining inference stability and throughput. Because the discrete RTX 4090 Mobile card used here isolates GPU VRAM from the system's main DDR5 memory, the model weights and intermediate activation maps remain largely resident on the GPU, which in turn mitigates paging overhead and helps sustain consistent inference speeds. Nonetheless, our analysis revealed dynamic GPU and CPU power throttling and reduced clock speeds to preserve battery life and control thermal loads, further complicating the inference profile. While these factors did not substantially hamper the tests reported in the present study, they illustrate how even nominally high-end, consumer-grade hardware can exhibit runtime variability under real operating conditions.

Model Implementation Considerations

Another aspect that conditions the broader applicability of these findings is the balance between parameter count, quantization level, and multilingual coverage. The 6-bit Q6_K quantization proved sufficient for sustaining near state-of-the-art sentiment classification across English, Italian, and Japanese, but more aggressive compression—such as 4-bit or 3-bit—could release additional VRAM capacity at the risk of small but potentially meaningful accuracy trade-offs. Conversely, moving to a larger model (e.g., 13B or 30B parameters) might enhance classification robustness or cross-lingual generalization, but would almost certainly necessitate advanced offloading techniques, more powerful GPUs, or significantly deeper quantization that could undermine the performance gains. Researchers and practitioners wishing to adapt this methodology to additional domains or languages—particularly those with more complex morphologies or larger token inventories—should carefully assess VRAM headroom, memory bandwidth, and the risk of memory-induced bottlenecks before scaling the architecture upward.

Moreover, the architectural features of the *lightblue/suzume-llama-3-8B* model itself may shape how well the results transfer to other LLM variants. This particular model integrates an ORPO-based fine-tuning process geared toward boosting multilingual sentiment performance. Should an organization opt for a different backbone (e.g., a GPT-style decoder, a hybrid encoder-decoder, or a custom domain-specific LLM), the memory footprint, quantization efficacy, and eventual inference times could deviate. Similarly, since the present model was largely trained on balanced synthetic data plus multilingual resources, real-world text or low-resource languages might produce greater variance in classification outputs, compounding any hardware constraints by demanding repeated ensemble passes or domain-specific fine-tuning. For smaller GPU setups or CPU-only pipelines, an ensemble approach might become too resource-intensive or slow to maintain practical inference rates, implying that the architectural and hardware synergies demonstrated here cannot be assumed to hold in all computational environments.

Opportunities for Hardware Interoperability

Overall, the details of our hardware configuration—and the unique advantages it confers—underscore how near-real-time, privacy-preserving sentiment analysis can be obtained on a single consumer notebook, as long as memory overheads are carefully managed, and the quantized model is matched to the GPU's VRAM capacity. Yet these very same factors also limit one-to-one generalization to other contexts in which the hardware–model interplay may differ. In lower-memory devices, frequent paging or incomplete weight loading would likely erode the quick inference times observed, while enterprise-grade systems with multiple high-end GPUs might surpass our throughput results but at the expense of greater infrastructure costs. Hence, to replicate or adapt the present approach in new settings, researchers and practitioners should thoroughly profile their hardware's memory allocation, concurrency patterns, thermal limits, and compute overhead, ensuring that the chosen model size and quantization strategy align with the system's practical throughput targets and VRAM constraints. Through further experiments involving smaller GPUs, integrated graphics, CPU-only processing, Apple Silicon architectures, and larger-scale LLM backbones, we aim to provide a more granular map of how local sentiment analysis architectures scale up or down under a wide range of resource configurations while remaining portable across Python-based *llama-cpp* frameworks.

Appendix B: Future Research Avenues

This appendix complements the main manuscript by elaborating on two key directions for future research: (i) integrating real-world, user-generated data into our multilingual sentiment analysis pipeline, and (ii) extending coverage to additional languages beyond English, Italian, and Japanese. Both directions seek to strengthen ecological validity, address potential domain mismatches, and ensure that local LLM deployments remain robust across diverse linguistic and practical contexts.

Integrating Real-World User-Generated Data

Although we employed a balanced, synthetic dataset in the present study to enable tightly controlled experiments and fair cross-linguistic comparisons, we recognize that naturally occurring text introduces a range of complexities not captured in our current evaluations. These complexities include sentiment skew (e.g., predominantly positive or negative reviews), colloquial language (slang, abbreviations, emoticons), and domain-specific jargon (e.g., references to specific product brands, city names, or evolving internet slang). In practice, many real-world data streams, such as social media posts or consumer feedback portals, exhibit highly variable structures and frequent lexical irregularities that may not arise in a curated synthetic corpus.

Motivation and Context

Local, privacy-preserving deployments of large language models (LLMs) are especially appealing for organizations that handle sensitive user data or require compliance with regulations like GDPR. Yet these same organizations often face idiosyncratic domain data (e.g., specialized medical terms in healthcare platforms, local dialect or brand references in e-commerce). By collecting naturally occurring user-generated content—whether from publicly accessible review sites or internally managed feedback forms—we can more accurately measure how local LLMs manage contextual ambiguity, highly imbalanced sentiment distributions, and domain mismatch. Real-world text often contains sarcasm, code-switching, incomplete sentences, or repeated references to unknown brand names or cultural artifacts. Additionally, certain services or products tend to draw predominantly positive (or negative) reviews, with many authentic data streams lacking neatly balanced classes. Language usage in specialized domains like biotech or finance can also significantly differ from general-purpose text. By embracing these complexities, our future work will further validate the applicability of a locally quantized LLM for real-world scenarios. Moreover, we aim to explore domain adaptation methods, such as fine-tuning the existing quantized model with small amounts of domain-specific data, to mitigate potential performance drops caused by lexical and stylistic mismatches.

Practical Data Collection and Annotation Plans

We propose a structured approach to data gathering and annotation that begins with source identification, where we plan to target multiple review domains, including hospitality (hotel and restaurant reviews), retail (electronics, fashion), and specialized user forums (healthcare or tech support boards). Ethical and privacy considerations are paramount - all user-generated data will be anonymized and, where required, re-annotated in-house to maintain compliance with privacy regulations. We will avoid personally identifiable information (PII) to ensure minimal risk if the data need to be stored or shared. Our annotation pipeline will adopt a two-tier labeling system: primary sentiment labels (positive, neutral, negative) will be assigned by two or more human annotators while tracking inter-annotator agreement, alongside domain tags or

specialized markers such as "finance-related," "technical," "healthcare," or "colloquial slang." These tags help us analyze whether certain subdomains systematically degrade performance. To handle skew and address imbalance, we may incorporate strategic subsampling or oversampling of minority sentiment classes. Where feasible, we will let the real-world distribution remain unaltered to test genuine model robustness, then optionally create more balanced subsamples for controlled comparisons.

Methodological Extensions

Once authentic data are collected, we will compare performance on these real-world corpora to our synthetic baseline by analyzing several key aspects. We will examine accuracy drift under domain-specific lexical usage and skewed sentiment distributions, as well as assess the impact of data augmentation or fine-tuning on bridging domain mismatch. Additionally, we will verify quantization stability by ensuring that a 6-bit quantized model retains robust generalization even with the noisier, more heterogeneous input characteristic of social media or consumer reviews. Such empirical insights will clarify where local LLMs excel in practical settings and which adaptation strategies best preserve performance while ensuring minimal hardware overhead.

Expansion to Additional Languages

Our current experiments center on English, Italian, and Japanese, which allowed a multifaceted view of performance across alphabetic (English, Italian) and logographic/kana-based (Japanese) writing systems. Despite their diversity, these three languages only begin to represent the spectrum of linguistic and orthographic phenomena encountered globally. Given the strong results under these initial conditions, we aim to extend coverage to further languages, including those that introduce right-to-left scripts, highly agglutinative morphology, or rare subword patterns.

Candidate Languages and Their Challenges

A further selection of candidate languages for expansion must strategically cover diverse linguistic families, each chosen to probe different aspects of model robustness and cross-lingual transfer.

1. Spanish, as a member of the Romance family, presents an ideal stepping-stone from our existing Italian implementation due to its largely transparent orthography and shared linguistic features. This similarity offers a unique opportunity to investigate how morphological parallels between related languages might enhance cross-lingual transfer, potentially allowing us to leverage existing Romance language representations more efficiently.
2. Arabic would introduce significantly different challenges as a Semitic language, with its right-to-left script and sophisticated system of morphological concatenation. The language's rich morphological structure, where words are built from consonantal roots with vowel patterns and affixes, creates distinct tokenization challenges. This complexity is further compounded by the need for morphological disambiguation, as the same written form can often represent multiple grammatical categories depending on context.
3. Chinese (Mandarin) could represent perhaps our most ambitious expansion target, employing a logographic writing system that fundamentally differs from both the Latin script used in English and Italian, and the mixed script environment we have already tackled in Japanese. Its character-based system, governed by strict syllabic constraints and featuring a unique partial morphological structure, will test the model's ability to handle radically different approaches to encoding meaning. The language's lack of explicit

word boundaries and its tonal nature add additional layers of complexity to the tokenization process.

In each of these languages, we anticipate encountering novel challenges in subword segmentation, the interpretation of morphological cues, and the handling of script-specific constraints. These challenges become particularly critical when working with 6-bit quantization, as we must ensure that the reduced precision does not compromise the model's ability to capture and process these linguistic nuances. The interaction between quantization and these various linguistic features could reveal important insights about the robustness and limitations of our approach, potentially guiding future optimizations in both model architecture and quantization strategies.

This systematic expansion across linguistically diverse languages will help us better understand how our quantization approach scales across different writing systems and grammatical structures, ultimately informing more robust and generalizable implementations of locally deployed language models.

Practical Integration and Resource Constraints

Because local deployments rely on consumer-grade hardware with limited VRAM (e.g., 8–16 GB), carefully managing the total parameter count becomes crucial when accommodating multiple languages. We can employ several strategies to address this challenge, including vocabulary pruning by removing subword tokens that are rarely used by newly included languages while preserving space for more commonly used subword units. Another approach is incremental fine-tuning, which involves sequentially teaching the model new languages with careful regularization to avoid catastrophic forgetting of existing languages. We may also explore aggressive quantization, such as 4-bit or 3-bit quantization for model expansions to maintain VRAM viability, though we must carefully assess any resultant accuracy trade-offs.

Evaluation Protocols

Our evaluation protocols for future language expansions will retain the rigorous Monte Carlo approach that we used for English, Italian, and Japanese. This methodology enables us to establish comparable confidence intervals across newly introduced languages and conduct distribution and variance analyses to pinpoint any performance lags in particular morphological contexts or script types. We can also investigate potential zero-shot or few-shot transfer, testing whether knowledge gleaned from existing languages (especially if they share genealogical roots, like Romance languages) can bootstrap performance in newly added languages. By systematically incorporating more languages, we aim to clarify the upper bounds of local, quantized LLM performance on consumer hardware, while exploring solutions for partial offloading or advanced memory optimizations if VRAM limitations become binding.

Reflections on Future Directions

By merging these two lines of expansion—naturalistic, user-generated data and additional language coverage—we plan to more thoroughly validate the resilience of locally deployed, quantized LLMs in true production scenarios. In these real-world environments, sentiment is imbalanced or expresses itself in nuanced, domain-specific ways, while script and morphology vary widely, placing stress on tokenization and subword embedding accuracy. Additionally, hardware constraints demand efficient memory usage while preserving inference speed and privacy compliance. Ultimately, these steps will help organizations and researchers confidently adopt local LLM solutions in real operational pipelines, bridging the gap between controlled academic experiments and the genuine linguistic diversity encountered in global user bases.

Appendix C: Bias Considerations in Local Sentiment Analysis

Although our current study centers on achieving high accuracy in multilingual sentiment classification under hardware-constrained, locally deployed large language models (LLMs), sentiment analysis inevitably raises ethical and practical questions around fairness and bias. These issues are especially salient when applying NLP tools to heterogeneous, real-world data streams that may reflect demographic inequalities, cultural nuances, and power imbalances embedded in language use. Even under a privacy-preserving local deployment scheme, the risk remains that an LLM's underlying training data or quantization adjustments could inadvertently produce biased outputs for certain groups or content domains. Below, we discuss the potential origins of such biases, the limitations of using predominantly synthetic data to detect them, and some strategies to mitigate or measure fairness deficits in future work. Moreover, we conducted a small-scale pilot study to address the question of whether introducing bias-prone demographic groups can significantly skew the accuracy results.

Bias in Multilingual Sentiment Analysis

Sentiment classification depends on patterns learned from corpora that encode societal stereotypes and norms. If certain demographics, dialects, or cultural references appear less frequently (or appear predominantly in negative contexts) models risk mislabeling or systematically downgrading text associated with those groups. For instance, studies have shown that heavy usage of African American English (AAE) expressions may cause speakers to be considered substantially more toxic than non-AAE speakers, even when discussing similar subjects (Hofmann et al., 2024). In multilingual scenarios, these biases can be compounded by linguistic and cultural divergences: for instance, an expression viewed as neutral in one language might be perceived as more charged in another. Research indicates that cross-lingual transfer can worsen bias in sentiment analysis, as systems using cross-lingual transfer often become more biased than their monolingual counterparts (Goldfarb-Tarrant et al., 2023). Furthermore, sentiment signals often co-occur with contextual mentions of social categories, such as gender, ethnicity, or religion, potentially leading to unintended associations in the latent representation of text. A nominally high accuracy on balanced synthetic data, like the one deployed in this study, may mask subtle disparities when confronted with real user content that references different social identities or sensitive topics (e.g., "immigration policies", "women's healthcare", or "religious festivals").

In addition, quantization, while vital for fitting large models onto consumer hardware, can influence how the model treats infrequent tokens, specialized terminology, or subtle morphological inflections. If certain dialect words or culturally loaded phrases are used less frequently in the training data, they might be represented by lower-precision weights. This could disproportionately affect model performance for underrepresented language varieties, effectively reinforcing existing imbalances in the source data.

Advancing Bias Detection Through Multi-Modal Data Integration

Our initial evaluation using synthetic balanced datasets has established a strong foundation for understanding cross-linguistic performance in controlled environments. This groundwork opens pathways to develop more sophisticated bias detection systems that incorporate real-world complexity. By combining synthetic data's systematic coverage with authentic text patterns, we can create hybrid evaluation frameworks that leverage the strengths of both approaches.

Future research directions could include developing adaptive sampling techniques that preserve synthetic data's statistical rigor while introducing measured variability in

demographic and cultural representations. This could involve creating generation protocols that authentically reflect sociolinguistic patterns across different communities and contexts. Such enhanced synthetic data could model subtle linguistic phenomena like context-dependent prejudice, sarcasm, and microaggressions by incorporating pragmatic markers and cultural references.

The next phase of development will focus on building comprehensive evaluation pipelines that combine synthetic benchmarks with carefully curated real-world examples. This multimodal approach will enable more nuanced fairness audits by examining how models handle both controlled test cases and naturalistic language variation.

Frameworks and Metrics for Future Fairness Evaluations

The path to a more comprehensive model evaluation extends beyond simple accuracy metrics, drawing on rich interdisciplinary insights from NLP, machine learning, and social sciences to develop nuanced fairness assessments. This integrated approach has given rise to several sophisticated evaluation frameworks that probe different dimensions of model behavior. Demographic parity examines whether the model maintains consistent classification patterns across demographic groups, measuring for instance if texts referencing different genders or ethnicities receive comparable sentiment ratings. The equalized odds framework takes this further by scrutinizing error distribution, ensuring that no particular demographic bears an undue burden of misclassification through either false positives or false negatives. Additionally, contextual embedding analysis delves into the semantic associations formed around identity terms like "female", "disabled", or "immigrant", revealing any systematic biases in how these concepts are represented within the model's understanding.

The complexity of these evaluations becomes even more pronounced in multilingual systems, where fairness metrics must be assessed not only within each language but also across linguistic boundaries. A model might demonstrate equitable performance when processing English content while simultaneously harboring subtle biases in its treatment of identity markers in Japanese or Italian text, particularly if the training data for these language-culture pairs is less extensive or diverse. This multilingual dimension underscores the importance of comprehensive evaluation across the full spectrum of supported languages to ensure consistent fairness standards throughout the system.

Practical Steps Toward Fair, Bias-Aware Analysis

The implementation of robust bias detection and mitigation in local LLM pipelines requires a multifaceted approach that combines rigorous testing, strategic data enhancement, and continuous monitoring. At the foundation of this framework lies the careful assembly of bias-focused test sets, which serve as specialized diagnostic tools. These curated corpora should encompass a wide spectrum of demographic representations, including varied dialect features, cultural references, and sensitive topics that might trigger classification discrepancies. The test sets should authentically reflect real-world language patterns, incorporating domain-specific terminology, colloquialisms, and community-specific expressions that might otherwise be overlooked in standard evaluation procedures.

Data augmentation emerges as a powerful strategy for addressing representational imbalances within model training. When certain dialects, cultural expressions, or demographic groups are underrepresented, targeted augmentation can help restore balance and improve model fairness. This might involve sophisticated oversampling techniques for minority group references, generating contextually appropriate examples that challenge existing stereotypes, or incorporating adversarial examples that explicitly test the model's handling of bias-sensitive

scenarios. The augmentation process should be carefully calibrated to maintain linguistic authenticity while expanding the model's exposure to diverse perspectives and expressions.

Regular auditing of model outputs serves as a critical ongoing safeguard against emerging biases. These systematic reviews should extend beyond synthetic benchmarks to examine real-world performance across different contexts and user populations. By analyzing both live and historical data, practitioners can track how model behavior evolves over time, particularly in response to domain adaptation efforts or updates to the base model architecture. This temporal analysis helps identify subtle shifts in fairness metrics that might otherwise go unnoticed.

The deployment of sophisticated explanatory tools adds another layer of diagnostic capability to the bias detection framework. Advanced techniques such as feature attribution analysis and saliency mapping can reveal the internal mechanics of model decision-making, highlighting specific n-grams or embedding patterns that might systematically trigger biased responses. These insights enable practitioners to pinpoint problematic associations within the model's learned representations and implement targeted interventions to address them. Furthermore, these diagnostic tools can help track the effectiveness of bias mitigation strategies by providing detailed visibility into how model behavior changes in response to various interventions.

Relevance to Local LLM Deployments

It may seem that local deployment, by eliminating data transfers to external servers, principally addresses privacy concerns (e.g., GDPR compliance). Yet, privacy does not inherently guarantee fairness. A model can still produce systematically biased classifications even if the data remain strictly on-device (Lyu et al., 2019). For organizations striving to comply not only with privacy mandates but also with ethical best practices, fairness audits become an essential extension of typical performance evaluations (Lyu et al., 2020). The fact that the pipeline runs locally on consumer-grade hardware adds further impetus to ensure that any identified biases are rectified in situ, without relying on continuous patches or remote retraining from external providers (Meerza et al., 2024).

Pilot Study Results

Although our current study centers on achieving high accuracy in multilingual sentiment classification under hardware-constrained, locally deployed large language models (LLMs), sentiment analysis inevitably raises ethical and practical questions around fairness and bias. These issues are especially salient when applying NLP tools to heterogeneous, real-world data streams that may reflect demographic inequalities, cultural nuances, and power imbalances embedded in language use. Even under a privacy-preserving local deployment scheme, the risk remains that LLM's underlying training data or quantization strategies could inadvertently produce biased outputs for certain groups or content domains. Below, we present the results of a pilot study to evaluate whether language that could induce bias leads to significant sentiment misclassification rates.

Experimental Setup and Rationale

In this pilot, we tested a set of synthetic demographic-focused sentences referencing older adults, female individuals, and immigrant groups across English, Italian, and Japanese. Each sentence was repeated through multiple inference passes to ascertain whether misclassifications were consistent (implying underlying representational bias) or merely random (correctable by ensemble voting). Unlike the main study's reliance on bias-free synthetic data alone, this mini-dataset introduced domain-specific references and demographic contexts, albeit still smaller and more controlled than large real-world corpora.

As suggested by the main study, we used 36 repeated inference passes per item, generating a total of 4,320 classifications. This ensemble-based approach matches the Monte Carlo spirit of the main experiment, making it particularly relevant for detecting subtle classification instabilities that might correlate with biases toward specific demographic identifiers.

Dataset Considerations

This pilot study systematically investigates bias-related language patterns through a multilingual dataset encompassing English, Italian, and Japanese. The dataset comprises 120 sentences generated by systematically combining 20 foundational scenarios, each rendered into three languages and represented with positive and negative sentiment ($20 \times 3 \times 2 = 120$). This controlled design allows direct comparisons across languages and sentiments while preserving semantic consistency.

This dataset is organized around diverse groups vulnerable to societal biases, such as older adults, women, immigrants, disabled individuals, and low-income communities. It extends further to encompass single parents, LGBTQ+ individuals, refugees, neurodivergent employees, and others who frequently encounter discrimination. By including a broad range of demographic targets, the dataset captures various expressions of prejudice—be it ageism, sexism, xenophobia, ableism, or socioeconomic bias—while maintaining a manageable scope.

Positive sentences emphasized supportive, respectful, or empowering actions, employing terms like "adapted", "welcomed", or "provided support", while negative sentences highlight prejudiced or exclusionary behavior, often using words such as "excluded", "criticized harshly", or "discriminated against" (Table C1).

Table C1 Examples of pilot study sentences in English, Italian, and Japanese.

Class	English	Italian	Japanese
Positive	The school adapted its curriculum to include diverse cultures.	La scuola ha adattato il curriculum per includere culture diverse.	学校は多文化を取り入れたカリキュラムを導入した。
Negative	The city neglected to provide ramps for wheelchair users.	La città non ha fornito rampe per utenti su sedia a rotelle.	市は車椅子利用者のためのスロープを設置しなかった。

Classification Rationale

An important distinction between this pilot and the three-tier classification employed in the main text is our decision here to focus on a binary (positive vs. negative) classification scheme. Although the main study delineates positive, neutral, and negative sentiment, the pilot is more narrowly concentrated on detecting potential extremes of classification that might disproportionately affect certain demographic groups (Kiritchenko & Mohammad, 2018). By compressing the neutral class into one of the two polarity extremes or excluding neutral examples, we direct our attention to the presence or absence of unambiguously positive versus adverse sentiment signals.

From a bias-detection standpoint, extreme sentiment shifts (e.g., from strongly positive to strongly negative) can carry the most significant ethical and practical consequences, as they can reinforce or amplify damaging stereotypes (Mei et al., 2023). Neutral judgments, while

informative in broader usage, do not always expose the kinds of representational inequalities that underlie many bias concerns (Sap et al., 2019). Consequently, by operationalizing a binary approach, we chose to emphasize the risk of harmful misclassification—e.g., if references to older immigrants systematically trigger negative sentiment (Asyrofi et al., 2021). This choice simplifies the experimental design, ensuring each demographic mention elicits a straightforward polarity judgment from the model.

In future, more comprehensive bias evaluations, the neutral category could be reintroduced to reveal whether certain group references evade polarity altogether. Yet for this initial pilot, the binary setup proved sufficient to determine whether the model was disproportionately classifying particular demographic mentions with negativity (or positivity). Given the high accuracy across repeated inferences, no significant bias patterns emerged in this limited sample, although more expansive datasets remain necessary to confirm these findings and accommodate neutral expressions fully (Buolamwini & Gebru, 2018).

Overall Findings

Summary Statistics. The validation statistics reveal excellent performance across all three languages (Table C2), with mean accuracies ranging from 0.951 (Italian) to 0.981 (English), and 0.988 (Japanese). Cohen's kappa values likewise indicate strong consistency between predicted labels and ground-truth annotations. All responses showed significant non-random agreement ($p < 0.05$). Although the pilot dataset is intentionally small, these metrics suggest no glaring demographic-specific biases under repeated inference conditions. The stronger performance compared to the data in the main study is ascribable to a two-sentiment versus a three-sentiment analysis.

Table C2 Per language summary statistics.

Language	Metric	Value	95% CI	SD	Median	IQR
English	Accuracy	0.981	0.978–0.985	0.011	0.975	0.006
	Cohen's kappa	0.962	0.955–0.970	0.022		
Italian	Accuracy	0.951	0.944–0.959	0.022	0.950	0.025
	Cohen's kappa	0.903	0.888–0.917	0.045		
Japanese	Accuracy	0.988	0.984–0.992	0.013	1.000	0.025
	Cohen's kappa	0.976	0.968–0.985	0.025		

Class-Level Performance. Analyses of positive and negative classes across English, Italian, and Japanese reinforce the general conclusion of minimal misclassification (Table C3). For example, English negative-class recall was 0.962 (95% CI: 0.877–1.000), and English positive recall reached a perfect 1.000 in repeated sampling. Similar trends emerged in Italian and Japanese, though slight divergences exist in negative recall for Italian (0.903) and negative precision for Japanese (0.978). These divergences did not clearly map onto any single demographic focus, implying that no specific group (e.g., older adults or immigrant communities) faced systematically higher misclassification rates.

Table C3. Class-level statistics.

Language	Class	Metric	Value	95% CI	SD
English	Positive	Precision	0.964	0.882–1.000	0.021
		Recall	1.000	1.000–1.000	0.000
		F1-Score	0.982	0.936–1.000	0.011
	Negative	Precision	1.000	1.000–1.000	0.000
		Recall	0.962	0.877–1.000	0.022
		F1-Score	0.981	0.933–1.000	0.011
Italian	Positive	Precision	0.913	0.780–0.999	0.036
		Recall	1.000	1.000–1.000	0.000
		F1-Score	0.954	0.876–0.999	0.020
	Negative	Precision	1.000	1.000–1.000	0.000
		Recall	0.903	0.758–0.998	0.045
		F1-Score	0.948	0.861–0.999	0.025
Japanese	Positive	Precision	1.000	1.000–1.000	0.000
		Recall	0.976	0.921–1.000	0.025
		F1-Score	0.988	0.957–1.000	0.013
	Negative	Precision	0.978	0.925–1.000	0.024
		Recall	1.000	1.000–1.000	0.000
		F1-Score	0.988	0.959–1.000	0.012

Statistical Analysis

Distribution Tests. Anderson-Darling tests detected substantial departures from normality across all languages (English: A = 8.494, p = 4.33e-21; Italian: A = 2.403, p = 3.29e-06; Japanese: A = 6.331, p = 7.08e-16), prompting the use of both Welch's ANOVA ($F(2, 66.27)$ = 36.97, $p < .001$) and the non-parametric Kruskal-Wallis test ($H(2) = 56.84$, $p < .001$). Both tests strongly rejected the null hypothesis of equal means/distributions between languages, with the Kruskal-Wallis showing even stronger evidence against the null hypothesis (larger test statistic relative to its distribution). Despite these distributional caveats, the Fligner-Killeen test (p = 0.074) found no strong heteroscedasticity. Both parametric Games-Howell and non-parametric Dunn's test post-hoc comparisons revealed significant differences between languages, with Japanese performing slightly better than English (Games-Howell: p = 0.04; Dunn's test: p = 0.25), and both significantly outperforming Italian (all $p < 0.001$). The effect sizes were consistent across both parametric ($\omega^2 = 0.513$) and non-parametric ($\varepsilon^2 = 0.513$) approaches, confirming previous findings that language choice accounts for a substantial portion of the observed variation in accuracy. Rank biserial correlations further quantified these differences, showing strong effects between Italian and both English (rrb = –0.750) and Japanese (rrb = 0.843), but only a modest effect between English and Japanese (rrb = 0.278).

Intra-Rater Agreement. Intra-rater agreement for English reached 99.4% across repeated responses, reflecting high stability of predicted labels. Italian (97.6%) and Japanese (98.8%) similarly exhibited robust ensemble consistency, indicating that any borderline cases tend to remain marginal in a repeated-inference scenario. From a bias perspective, such consistency implies the system is not randomly fluctuating in its sentiment assignments for demographic references; if bias were present, it would likely manifest as systematically skewed predictions, not ephemeral or fixable by majority vote.

Implications for Bias Detection

Although no strong evidence of group-specific bias emerged within this small pilot, these results do not rule out more subtle biases that might surface in broader, less curated texts. The model's strong performance may reflect the highly controlled nature of the dataset, the relatively balanced negative and positive references for each demographic category, and the synergy between repeated inference and the model's underlying training signals. Nonetheless, the pilot validates an ensemble-based approach to revealing or mitigating classification inconsistencies, highlighting that repeated sampling yields stable predictions across sensitive demographic topics in these short texts.

Directions for Extended Bias Audits

Moving forward, a few expansions remain essential. First, incorporating larger real-world corpora through user-generated content from social media, forums, or domain-specific data sources will be crucial, as sentiment in these contexts may be deeply intertwined with cultural references and intersectional identities. Additionally, examining references that combine multiple demographic markers (e.g., older female immigrants) will enable more nuanced intersectional analyses. Testing targeted data augmentation through adversarial fine-tuning for specific subgroups will also be important if disparities appear in more extensive datasets. Ultimately, such steps will strengthen confidence that on-device LLM approaches remain equitable while preserving data privacy via local deployment.

Concluding Remarks on the Pilot

The pilot's results demonstrate the feasibility of using a repeated-inference, ensemble-based methodology to gauge fairness and detect potential biases in local LLM sentiment analysis. Although minimal bias signals were detected within this small dataset, the approach provides a scalable blueprint for larger and more diverse corpora—ensuring that practitioners can systematically integrate fairness considerations into local model evaluations. By adopting either binary or three-tier classification schemes as context demands, it becomes possible to address the full complexity of sentiment analysis tasks while remaining vigilant against subtle demographic misclassifications.