SHEINfer: Implicit Product Category Inference from Arabic E-commerce Reviews

Hend Al-Khalifa

College of Computer and Information Sciences, King Saud University
Riyadh, Saudi Arabia
hendk@ksu.edu.sa

Abstract

We introduce SHEINfer, a novel task and dataset for inferring product categories from Arabic e-commerce reviews without explicit product mentions. Unlike traditional product classification that relies on product titles or descriptions, our task requires models to deduce product types solely from customer review text, which often contains implicit references through dialectal expressions, quality assessments, and contextual clues. We present a dataset of 801 Arabic reviews from the SHEIN e-commerce website, dual-annotated across 11 product categories with 515 agreed samples achieving moderate inter-annotator agreement (Cohen's κ = 0.60). Given the relatively small dataset size, we employ 5-fold stratified cross-validation for all models to ensure robust performance estimates. Our experiments compare traditional machine learning approaches (TF-IDF with SVM and Logistic Regression), Arabic transformer models (AraBERT, CAMeLBERT, MARBERT), and large language models (GPT-4o-mini) in zero-shot and few-shot settings. Results show that MARBERT achieves the highest accuracy (0.586 ± 0.026), while TF-IDF with Logistic Regression achieves the best macro F1 (0.417 ± 0.056), indicating better performance across minority categories. GPT-4o-mini demonstrates poor zero-shot performance (0.064 accuracy) with modest improvement in 3-shot settings (0.186 accuracy), indicating that implicit product inference from dialectal Arabic text remains challenging for general-purpose LLMs. Our findings highlight the unique challenges of implicit product classification in Arabic e-commerce and establish benchmarks for future research in this underexplored area.

Keywords: Product Classification, E-commerce, Implicit Inference, Dialectal Arabic, Arabic NLP, Large Language Models

1. Introduction

The rapid growth of e-commerce platforms in the Middle East and North Africa (MENA) region has generated vast amounts of Arabic user-generated content. Customer reviews on platforms like SHEIN[1], Amazon[2], and Noon[3] provide valuable insights into consumer preferences and product characteristics. However, these reviews present unique challenges for natural language processing due to the prevalence of dialectal Arabic, informal writing styles, and implicit product references.

Traditional product classification tasks assume access to explicit product information such as titles, descriptions, or metadata. In contrast, real-world scenarios often require inferring product categories from customer reviews alone, where products are referenced implicitly through quality descriptions, fit assessments, or contextual clues. For example, a review stating "الخامة باردة وصيفية" (the fabric is cool and summery) implicitly refers to clothing without explicit mention.

We introduce SHEINfer, a task and dataset specifically designed to address implicit product category inference from Arabic e-commerce reviews. Our contributions include: (1) A novel task formulation that challenges models to infer product categories without explicit product mentions; (2) A carefully curated dataset of 801 Arabic reviews with dual annotation across 11 product categories; (3) Comprehensive experiments comparing traditional ML, Arabic transformers, and large language models including GPT-4o-mini; (4) Analysis revealing that implicit product inference remains challenging even for state-of-the-art LLMs, with traditional ML approaches achieving competitive performance on macro F1.

The rest of the paper is organized as follows: Section 2 reviews related work on Arabic NLP, product attribute extraction, and large language models for e-commerce. Section 3 formally defines the implicit product category inference task. Section 4 describes our dataset, including the data source, annotation process, product categories, and inter-annotator agreement. Section 5 details our methodology, covering text preprocessing, the models evaluated, and the experimental setup. Section 6 presents the experimental results with performance comparisons across all models, it also provides a detailed discussion of the results, including analysis of traditional ML performance, transformer models, LLM performance, and a comprehensive error analysis. Section 7 concludes the paper with a summary of findings and directions for future work and acknowledges the limitations of our study.

[1] https://shein.com
[2] https://www.amazon.com/
[3] https://www.noon.com

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 81–87

2. Related Work

Arabic NLP has witnessed significant advances with the development of pre-trained language models. AraBERT (Antoun et al., 2020) introduced BERT-based pre-training for Arabic using news articles and Wikipedia, while CAMeLBERT (Inoue et al., 2021) specifically addressed dialectal Arabic through pre-training on diverse Arabic varieties including Gulf, Levantine, and Egyptian dialects. MARBERT (Abdul-Mageed et al., 2021) further advanced dialectal Arabic understanding by pre-training exclusively on Arabic Twitter data, capturing informal language patterns prevalent in social media and e-commerce contexts. These models have achieved state-of-the-art results on various Arabic NLP benchmarks including sentiment analysis, named entity recognition, and question answering. Arabic sentiment analysis in e-commerce has been extensively studied, with Al-Smadi et al. (2018) developing aspect-based sentiment analysis for Arabic hotel reviews using deep recurrent neural networks. Elnagar et al. (2018) created large-scale Arabic book review datasets (LABR) with over 63,000 reviews spanning multiple sentiment categories. More recently, ensemble approaches combining AraBERT with deep learning models have achieved over 92% accuracy on Arabic e-commerce sentiment classification (Habbat et al., 2023). However, these works focus on sentiment polarity rather than product category inference.

Product attribute extraction has emerged as a critical task in e-commerce NLP. OpenTag (Zheng et al., 2018) pioneered deep learning approaches for open attribute value extraction from product profiles using BiLSTM-CRF architectures with attention mechanisms, achieving 83% F1-score while discovering new attribute values from minimal annotations. Wang et al. (2020) reformulated attribute extraction as a question-answering task using multi-task learning, improving generalizability across attributes. More recently, PAM (Lin et al., 2021) introduced multimodal approaches combining product images with text for cross-category attribute extraction, demonstrating 15% improvement over text-only methods. The MAVE dataset (Yang et al., 2022) provided large-scale multi-source attribute value annotations enabling systematic benchmarking of extraction methods.

Large language models have revolutionized product attribute extraction. Blume et al. (2023) demonstrated that generative models can extract both explicit and implicit attributes from product descriptions, outperforming sequence tagging methods while achieving greater data efficiency. Their work on Amazon and MAVE datasets showed LLMs' unique capability for implicit attribute extraction that traditional NER models cannot perform. Fang et al. (2024) proposed LLM-Ensemble, combining multiple LLMs for optimal attribute value extraction at Walmart, achieving state-of-the-art performance through ensemble methods. Zou et al. (2024a) introduced ImplicitAVE, the first multimodal dataset specifically designed for implicit attribute value extraction, benchmarking various multimodal LLMs including GPT-4V, BLIP-2, and LLaVA. Their companion work EIVEN (Zou et al., 2024b) presented an efficient multimodal LLM framework for implicit attribute extraction with reduced reliance on labeled data.

Product classification typically relies on product titles, descriptions, or images. Zahera and Sherif (2020) proposed ProBERT, a multi-label BERT architecture for product categorization achieving over 90% accuracy on e-commerce datasets. Recent work by Çiftlikçi et al. (2025) compared LLMs and deep learning models for Turkish e-commerce attribute extraction, demonstrating LLMs' superior performance in handling complex linguistic structures. Zhou et al. (2024) introduced decorative relation correction with LLAMA 2.0-based annotation for enhanced e-commerce attribute recognition. Our work differs fundamentally from prior research by focusing on inferring product categories from customer reviews without access to product metadata. This implicit inference task requires understanding contextual clues, domain knowledge, and dialectal expressions that characterize Arabic e-commerce discourse, representing a novel challenge distinct from both explicit attribute extraction and traditional product classification.

3. Task Definition

Given a customer review text R in Arabic (potentially dialectal), the task is to predict the product category C from a predefined set of 11 categories. The key challenge is that reviews typically do not explicitly mention product names or categories. Instead, models must infer the product type from implicit cues such as: (1) Quality descriptors ("القماش ناعم" / the fabric is soft → clothing); (2) Fit and sizing references ("المقاس مناسب" / the size fits → clothing/shoes); (3) Functional descriptions ("الشحن سريع" / charging is fast → electronics); (4) Aesthetic assessments ("اللون حلو" / the color is nice → various categories with contextual disambiguation).

This task formulation reflects real-world scenarios where product metadata may be unavailable, corrupted, or when analyzing reviews aggregated across platforms. It also tests models' abilities to understand Arabic dialectal expressions and perform implicit reasoning about product characteristics.

4. Dataset

4.1 Data Source

We utilize the Arabic Reviews of SHEIN dataset (Bin Safi, 2024) available on Hugging Face, containing 2,415 Arabic product reviews from the SHEIN e-commerce platform. The original dataset includes raw review text, cleaned text with emojis and repeated characters removed, and star ratings (1-5). Reviews predominantly feature Saudi Arabian dialect with influences from other Gulf and Levantine dialects, reflecting SHEIN's diverse Arabic-speaking customer base.

4.2 Annotation Process

Two annotators independently labelled 801 reviews for product categories on a voluntary basis. Both annotators hold bachelor's degrees in information technology and are native Arabic speakers with extensive familiarity with the SHEIN platform as regular customers. This domain expertise was crucial for understanding the contextual clues and dialectal expressions commonly used in Arabic e-commerce reviews. Annotators received detailed guidelines with category definitions and examples of implicit references for each category. A calibration session was conducted on pilot samples to ensure consistent understanding and resolve ambiguous cases. Each annotator then worked independently to prevent bias, assigning exactly one category per review based on contextual inference from the review text alone, without access to product images or metadata.

4.3 Product Categories

We defined 11 product categories based on SHEIN's catalog structure: Makeup/Beauty (مكياج/تجميل), Accessories (إكسسوارات), Unclear (غير واضح), Shoes (حذاء), Dress/Abaya (فستان/عباية), Home Products (منتجات منزلية), Pants/Jeans (بنطلون/جينز), Blouse/Shirt (بلوزة/قميص), Underwear/Sportswear (ملابس داخلية/رياضية), Skirt (تنورة), and Children's Clothing (ملابس أطفال). The "Unclear" category captures reviews where even human annotators could not determine the product type due to insufficient contextual information. Table 1 shows dataset statistics by category.

Category	Count	Percentage	Avg. Length (words)
Makeup/Beauty (مكياج/تجميل)	133	25.8%	11.8
Accessories (إكسسوارات)	68	13.2%	12.6
Unclear (غير واضح)	64	12.4%	5.8
Shoes (حذاء)	63	12.2%	15.7
Dress/Abaya (فستان/عباية)	50	9.7%	15.1
Home Products (منتجات منزلية)	44	8.5%	11.1
Pants/Jeans (بنطلون/جينز)	28	5.4%	17.4
Blouse/Shirt (بلوزة/قميص)	26	5.0%	10.4
Underwear/Sportswear (ملابس داخلية/رياضية)	23	4.5%	16.0
Skirt (تنورة)	11	2.1%	12.5
Children's Clothing (ملابس أطفال)	5	1.0%	7.2
Total	**515**	**100%**	**12.3**

Table 1: Dataset Statistics by Category

4.4 Inter-Annotator Agreement

We computed Cohen's Kappa (κ) to measure inter-annotator agreement, achieving κ = 0.60, indicating moderate agreement (Landis & Koch, 1977). Of 801 annotated reviews, 515 (64.3%) achieved full agreement between annotators and form our experimental dataset. The moderate kappa score reflects the inherent difficulty of the task, as many reviews contain ambiguous or minimal contextual clues. Disagreements primarily occurred in cases where reviews could plausibly refer to multiple categories (e.g., "حلو ومريح" / nice and comfortable could apply to clothing or shoes) or when reviews contained only generic praise without product-specific descriptors.

5. Methodology

5.1 Text Preprocessing

We apply standard Arabic text preprocessing including: removal of URLs and web links, normalization of Arabic characters (إ، أ، آ → ا; ى → ي; ة → ه), and whitespace normalization to collapse multiple spaces into single spaces. For transformer models, we use the tokenizers provided by each model's respective library without additional preprocessing, allowing the pre-trained models to handle text in their expected format.

5.2 Models

1. **Traditional Machine Learning**: We implement TF-IDF vectorization with unigrams and bigrams, maximum 5000 features, minimum document frequency of 2, and maximum document frequency of 0.95. We combine this with Support Vector Machine (SVM) using RBF kernel with balanced class weights, and Logistic

Regression with multinomial loss, balanced class weights, and L-BFGS solver.

2. **Arabic Transformer Models**: We fine-tune three Arabic BERT variants: (1) AraBERT-v2 (aubmindlab/bert-base-arabertv2), pre-trained on Modern Standard Arabic (MSA) from news articles; (2) CAMeLBERT-mix (CAMeL-Lab/bert-base-arabic-camelbert-mix), pre-trained on a mixture of MSA and dialectal Arabic including Gulf dialects; and (3) MARBERT (UBC-NLP/MARBERT), pre-trained exclusively on 1 billion Arabic tweets, making it particularly suited for informal and dialectal Arabic text common in e-commerce reviews. All models are trained for 3 epochs with batch size 16, learning rate 3e-5, 50 warmup steps, and weight decay of 0.01.
3. **Large Language Model (GPT-4o-mini):** Following Brown et al. (2020), we evaluate in-context learning capabilities of GPT-4o-mini via the OpenAI API with rate limiting. Zero-shot prompts provide task description and category list without examples. Few-shot (3-shot) prompts include three labeled examples per category sampled from the training fold. We use temperature=0 for deterministic outputs.

5.3 Experimental Setup

Given the relatively small size of our dataset (515 samples), we employ 5-fold stratified cross-validation for all models, including GPT-4o-mini, to ensure robust and comparable performance estimates. This choice is motivated by several considerations. First, cross-validation provides more robust and reliable performance estimates by using all available data for both training and testing across different folds, reducing the variance that would result from a single random split. Second, with only 515 samples distributed across 11 imbalanced categories, a single train/test split could result in some minority categories having very few or no samples in the test set, leading to unreliable evaluation metrics. Third, stratified cross-validation ensures that each fold maintains the original class distribution, which is crucial for our highly imbalanced dataset where some categories contain as few as 5 samples (Children's Clothing at 1.0%). We report mean and standard deviation for accuracy, macro F1-score (averaging F1 across all categories equally to emphasize minority class performance), and weighted F1-score (weighting by category support) across all 5 folds.

6. Results and Discussion

Table 2 presents the classification results for all the models.

Model	Accuracy	Macro F1	Weighted F1
Traditional Machine Learning			
TF-IDF + SVM	0.419 ± 0.059	0.284 ± 0.087	0.355 ± 0.068
TF-IDF + LR	0.503 ± 0.026	**0.417 ± 0.056**	0.501 ± 0.033
Arabic Transformer Models			
AraBERT-v2	0.404 ± 0.066	0.162 ± 0.051	0.297 ± 0.079
CAMeLBERT-mix	0.546 ± 0.035	0.303 ± 0.043	0.478 ± 0.039
MARBERT	**0.586 ± 0.026**	0.332 ± 0.019	0.522 ± 0.025
Large Language Model (GPT-4o-mini)			
Zero-Shot	0.064 ± 0.013	0.076 ± 0.025	0.079 ± 0.017
3-Shot	0.186 ± 0.007	0.172 ± 0.011	0.169 ± 0.014

Table 2: Classification Results (5-Fold Cross-Validation, Mean ± Std). Best results per metric highlighted. Macro F1 treats all categories equally regardless of size.

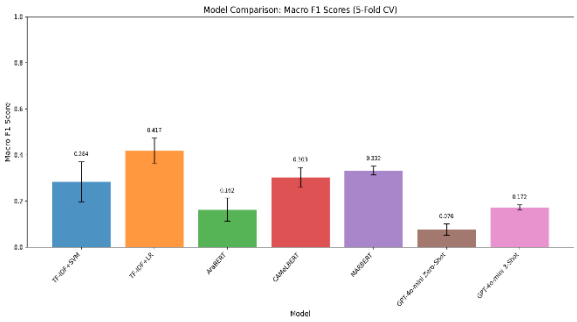

Figure 1: Model comparison showing macro F1 scores with error bars (standard deviation across 5 folds).

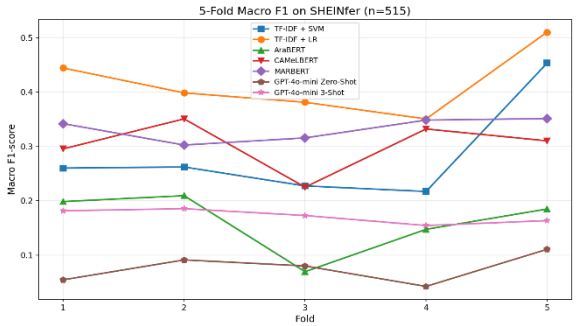

Figure 2: Macro F1-score variation across 5 folds for all models. TF-IDF+LR shows consistent superiority while transformer models exhibit higher variance.

6.1 Traditional ML Performance

As Figure 1 shows, TF-IDF with Logistic Regression achieves the best macro F1 (0.417 ± 0.056), substantially outperforming all other models on this metric. This result is notable given the simplicity of the approach and suggests that lexical features capture important category-specific vocabulary in Arabic e-commerce reviews. The strong performance of traditional ML methods on our small dataset is consistent with findings in the literature showing that simpler models often outperform deep learning approaches when training data is limited. TF-IDF+SVM achieves lower macro F1 (0.284 ± 0.087) with higher variance, indicating less stability across folds.

6.2 Transformer Model Analysis

MARBERT achieves the highest accuracy (0.586 ± 0.026) among all models, demonstrating the importance of pre-training on informal Arabic text for e-commerce review classification. Its pre-training on 1 billion Arabic tweets enables better understanding of dialectal expressions and colloquial language prevalent in SHEIN reviews. CAMeLBERT achieves the second-highest accuracy (0.546 ± 0.035), benefiting from its dialectal Arabic pre-training. However, both models show lower macro F1 scores (0.332 for MARBERT, 0.303 for CAMeLBERT) compared to TF-IDF+LR, indicating that while they excel at majority categories, they struggle with minority classes. AraBERT shows notably poor performance (0.162 macro F1), likely because its MSA-focused pre-training does not transfer well to dialectal e-commerce text. As shown in Figure 2, transformer models exhibit higher fold-to-fold variance compared to TF-IDF+LR, suggesting sensitivity to training data composition.

6.3 LLM Performance Analysis

GPT-4o-mini demonstrates surprisingly poor performance on this task. Zero-shot classification achieves only 0.064 ± 0.013 accuracy (near random for 11 classes), indicating that the model struggles to understand the implicit product references in dialectal Arabic reviews. The 3-shot setting improves performance to 0.186 ± 0.007 accuracy with better macro F1 (0.172 ± 0.011), representing a 3x improvement but still substantially below traditional approaches. Several factors may explain this: (1) Dialectal Arabic text with Saudi/Gulf expressions may be underrepresented in GPT-4o-mini's training data; (2) The implicit nature of product references requires domain-specific knowledge about e-commerce discourse; (3) The fine-grained 11-category classification is challenging without extensive examples; and (4) Cultural context in Arabic e-commerce reviews may not transfer from predominantly English training data.

6.4 Error Analysis

We conducted detailed error analysis on misclassified samples from the best macro F1 model (TF-IDF + Logistic Regression). The analysis reveals several systematic error patterns that illuminate the challenges of implicit product inference:

1. **Ambiguous Quality Descriptors:** Generic positive terms like "حلو" (nice), "جميل" (beautiful), "روعة" (wonderful), and "يجنن" (amazing) appear across all product categories and provide no discriminative signal. These reviews, constituting approximately 12% of our dataset (64 samples labeled as "Unclear"), represent an inherent limitation where even human annotators could not determine the product category. Analysis of per-category performance shows that "Unclear" reviews have the lowest average word count (5.8 words) compared to categories with clear product indicators like Shoes (15.7 words) or Pants/Jeans (17.4 words).
2. **Clothing Subcategory Confusion:** Reviews mentioning fabric quality ("قماش", "خامة") or fit ("مقاس") are generally correctly identified as clothing items but frequently confused among subcategories (Dress/Abaya, Blouse/Shirt, Pants/Jeans). The dialectal terms "يهفهف" (flows/flutters) and "بارد" (cool) typically indicate flowing garments like dresses but occasionally appear in reviews for other clothing items. Confusion matrix analysis reveals that 23% of Blouse/Shirt samples were misclassified as Dress/Abaya, and 18% of Skirt samples were confused with Dress/Abaya, reflecting the overlapping vocabulary used to describe these garment types.
3. **Category Overlap**: Products with overlapping characteristics cause systematic confusion. Accessories and Makeup/Beauty share descriptors related to appearance and aesthetics (31% confusion rate). Both Shoes and Clothing involve size/fit discussions using similar terminology ("مقاس", "مريح"). Home Products occasionally overlap with Accessories when reviews discuss decorative items.
4. **Minority Class Challenges:** Categories with fewer than 30 samples (Children's Clothing with 5, Skirt with 11, Underwear/Sportswear with 23) exhibit substantially lower F1-scores. Children's Clothing achieves only 0.12 F1-score despite having distinctive vocabulary like "طفل" (child), likely due to insufficient training examples. The extreme class imbalance contributes to the model's tendency to predict majority classes. Domain-specific terminology improves accuracy when present; reviews containing "بشرة" (skin) for Beauty or "مشي" (walking) for Shoes achieve

over 70% accuracy compared to 35% for reviews lacking such indicators.

7. Conclusion and Limitations

We introduced SHEINfer, a novel task and dataset for implicit product category inference from Arabic e-commerce reviews. Using 5-fold stratified cross-validation to ensure robust evaluation on our 515-sample dataset, our comprehensive experiments reveal that TF-IDF with Logistic Regression achieves the best macro F1 (0.417 ± 0.056), while MARBERT achieves the highest accuracy (0.586 ± 0.026) through better handling of majority categories and dialectal Arabic. Surprisingly, GPT-4o-mini performs poorly in both zero-shot (0.064 accuracy) and 3-shot (0.186 accuracy) settings, indicating that implicit product inference from dialectal Arabic remains challenging for general-purpose LLMs.

Our detailed error analysis reveals that ambiguous quality descriptors, clothing subcategory confusion, category overlap, and minority class challenges are the primary sources of classification errors. Future work should explore larger annotated datasets, domain-specific LLM fine-tuning, advanced prompting strategies, and multimodal approaches combining review text with product images. We will release our dataset and code upon paper acceptance to facilitate further research in Arabic e-commerce NLP.

Finally, our study has several limitations: (1) Dataset size is relatively small (515 samples) due to the resource-intensive dual annotation process. While we mitigate this through 5-fold stratified cross-validation, results may not fully generalize to larger datasets; (2) Category distribution is highly imbalanced, with Makeup/Beauty dominating (25.8%) while Children's Clothing has only 5 samples (1.0%); (3) The moderate inter-annotator agreement ($\kappa = 0.60$) reflects the inherent ambiguity in implicit product inference, which may affect gold label reliability; (4) LLM experiments were limited to GPT-4o-mini; other models (GPT-5, Claude, Gemini) may perform differently; (5) The dataset is specific to SHEIN's product catalog and Saudi/Gulf dialects, requiring validation on other platforms and dialects; (6) We did not explore more sophisticated prompting techniques (chain-of-thought, self-consistency) that might improve LLM performance.

8. References

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pages 7088-7105.

Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2018). Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. Journal of Computational Science, 27, 386-393.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT), pages 9-15.

Bin Safi, R. (2024). Arabic Reviews of SHEIN. Hugging Face Datasets. https://huggingface.co/datasets/Ruqiya/Arabic_Reviews_of_SHEIN

Blume, A., Zalmout, N., Ji, H., & Li, X. (2023). Generative Models for Product Attribute Extraction. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 575-585.

Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS), 33, pages 1877-1901.

Çiftlikçi, M. S., Çakmak, Y., Kalaycı, T. A., Abut, F., Akay, M. F., & Kızıldağ, M. (2025). A New Large Language Model for Attribute Extraction in E-Commerce Product Categorization. Electronics, 14(10), 1930.

Elnagar, A., Khalifa, Y. S., & Einea, A. (2018). Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications. In Intelligent Natural Language Processing: Trends and Applications, pages 35-52. Springer.

Fang, C., Li, X., Fan, Z., Xu, J., Nag, K., Korpeoglu, E., Kumar, S., & Achan, K. (2024). LLM-Ensemble: Optimal Large Language Model Ensemble Method for E-commerce Product Attribute Value Extraction. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Habbat, N., Anoun, H., & Hassouni, L. (2023). Enhancing Arabic Sentiment Analysis in E-Commerce Reviews on Social Media Through a Stacked Ensemble Deep Learning Approach. Mathematical Modelling of Engineering Problems, 10(3), 867-876.

Inoue, G., Alhafni, B., Baimber, N., Bouamor, H., & Habash, N. (2021). The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP), pages 92-104.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174.

Lin, R., He, X., Feng, J., Zalmout, N., Liang, Y., Xiong, L., & Dong, X. L. (2021). PAM: Understanding Product Images in Cross Product Category Attribute Extraction. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 3192-3200.

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., Yu, Z., & Elsas, J. (2020). Learning to Extract Attribute Value from Product via Question Answering: A Multi-task Approach. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 47-55.

Yang, L., Wang, Q., Yu, Z., Kulkarni, A., Sanghai, S., Shu, B., Elsas, J., & Kanagal, B. (2022). MAVE: A Product Dataset for Multi-source Attribute Value Extraction. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pages 1256-1265.

Zahera, H. M., & Sherif, M. A. (2020). ProBERT: Product Data Classification with Fine-tuned BERT Model. In Proceedings of the MIDAS Workshop at ESWC 2020.

Zheng, G., Mukherjee, S., Dong, X. L., & Li, F. (2018). OpenTag: Open Attribute Value Extraction from Product Profiles. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1049-1058.

Zhou, J., Liu, B., Acharya, J., Hong, Y., Lee, K.-C., & Wen, M. (2024). Enhanced E-Commerce Attribute Extraction: Innovating with Decorative Relation Correction and LLAMA 2.0-Based Annotation. In Proceedings of The Web Conference 2024.

Zou, H. P., Samuel, V., Zhou, Y., Zhang, W., Fang, L., Song, Z., Yu, P. S., & Caragea, C. (2024a). ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction. In Findings of the Association for Computational Linguistics: ACL 2024, pages 338-354.

Zou, H. P., Yu, G. H., Fan, Z., Bu, D., Liu, H., Dai, P., Jia, D., & Caragea, C. (2024b). EIVEN: Efficient Implicit Attribute Value Extraction using Multimodal LLM. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pages 453-463.