mathematics

MDPI

Article

Multi-Aspect Sentiment Analysis of Arabic Café Reviews Using Machine and Deep Learning Approaches

Hmood Al-Dossari * and Munerah Altalasi

Information Systems Department, College of Computer Sciences and Information, King Saud University, Riyadh 11421, Saudi Arabia; 442202827@student.ksu.edu.sa
* Correspondence: hzaldossari@ksu.edu.sa

Abstract

Online reviews on platforms such as Google Maps strongly influence consumer decisions. However, aggregated ratings mask nuanced opinions about specific aspects such as food, drinks, service, lounge, and price. This study presents a multi-aspect sentiment analysis framework for Arabic café reviews. Specifically, we combine machine learning (Linear SVC, Naïve Bayes, Logistic Regression, Decision Tree, Random Forest) and a Convolutional Neural Network (CNN) to perform aspect identification and sentiment classification. A rigorous preprocessing and feature-engineering with TF-IDF and n-gram was implemented and statistically validated through bootstrap confidence intervals and Friedman–Nemenyi significance tests. Experimental results demonstrate that Linear SVC with optimized TF-IDF tri-grams achieved a macro-F1 of 0.89 for aspect identification and 0.71 for sentiment classification. Meanwhile, the CNN model yielded a comparable F1 of 0.89 for aspect identification and a higher 0.76 for sentiment classification. The findings highlight that effective feature representation and model selection can substantially improve Arabic opinion mining. The proposed framework provides a reliable foundation for analyzing Arabic user feedback on location-based platforms and supports more interpretable and data-driven business insights. These insights are essential to enhance personalized recommendations and business intelligence in the hospitality sector.

Keywords: aspect-based sentiment analysis (ABSA); multi-aspect sentiment classification; arabic language processing; machine learning; deep learning

MSC: 68T50

check for updates

Academic Editors: Wei Fang and Victor S. Sheng

Received: 31 October 2025
Revised: 25 November 2025
Accepted: 2 December 2025
Published: 5 December 2025

Citation: Al-Dossari, H.; Altalasi, M. Multi-Aspect Sentiment Analysis of Arabic Café Reviews Using Machine and Deep Learning Approaches. *Mathematics* **2025**, *13*, 3895. https://doi.org/10.3390/math13243895

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Online reviews have become an essential component of consumer decision-making, particularly in the hospitality and service sectors [1]. Platforms such as Google Maps allow users to share feedback on local businesses, offering valuable insights into service quality and customer satisfaction [2]. However, aggregated star ratings conceal specific experiential dimensions—such as ambiance, service, food, and pricing—limiting their value for both customers and business owners [3,4]. Recent advances in text mining and sentiment analysis (SA) have enabled the extraction of fine-grained opinions from user-generated text, transforming unstructured language into actionable insights [5,6].

Arabic is one of the world's most widely spoken languages, yet sentiment analysis for Arabic, especially in the context of location-based or consumer reviews, remains underexplored [7,8]. The rapid adoption of Google Maps in Arabic-speaking regions provides an

ideal platform for studying authentic public opinions expressed in Arabic [2]. This study focuses on aspect-based sentiment analysis (ABSA) of Arabic café reviews from Google Maps, examining five key aspects: drinks, food, service, lounge, and price. Understanding such aspect-level opinions holds both theoretical and practical importance—helping consumers identify cafés aligned with their preferences while assisting businesses in recognizing which service components most strongly influence satisfaction [9].

Despite progress in sentiment research, Arabic ABSA continues to face linguistic and methodological challenges. The language's morphological richness, dialectal variation, and context-dependent sentiment cues complicate automated interpretation [6,8,10]. Early Arabic ABSA efforts relied on classical feature-based systems that combined morphological, syntactic, and semantic attributes with Support Vector Machines (SVMs) or Naïve Bayes classifiers [3,4,10]. These models provided valuable baselines but required extensive manual feature design. Subsequent work introduced convolutional and recurrent neural networks, marking a transition toward data-driven approaches. For instance, Abdelgwad et al. [6] achieved F1-scores of 70.67% for opinion-target extraction and 83.98% accuracy for polarity using a BiGRU–CNN–CRF architecture, while Al Smadi et al. [4] applied CNN and stacked BiLSTM networks to capture positional and contextual dependencies in hotel reviews.

The field has since shifted toward pretrained transformer models and multi-task learning. Studies such as Al Jarrah et al. [11] and Fadel et al. [5] demonstrated the strength of Arabic BERT variants (AraBERT, GigaBERT) in joint modeling of aspect extraction and sentiment classification, with macro-F1 scores approaching 80–82%. Semi-supervised and hybrid strategies have also emerged to counter data scarcity. Almasri et al. [7] used a teacher–student framework based on AraBERT to enhance aspect category detection, whereas Bensoltane and Zaki [12] achieved F1 = 84.58% through back-translation augmentation. Addressing dialectal diversity, Chennafi et al. [8] applied Seq2Seq normalization before transformer classifiers, improving polarity consistency across mixed Arabic domains.

Recent research has expanded beyond hotels and Twitter to new domains and methods. AlMasaud et al. [13] developed a multi-sentiment restaurants reviews corpus for aspect-based sentiment analysis, while Chouikhi et al. [14] released a BERT-Based joint model for aspect term extraction and aspect polarity detection in Arabic text. These studies illustrate growing interest in real-world, domain-specific ABSA applications but also highlight gaps in reproducibility—few report per-aspect metrics, significance testing, or confidence intervals.

Beyond the studies discussed above, several works have examined large-scale Arabic sentiment datasets and benchmarking methodologies. For example, the ARSAS [15] and LABR [16] corpora were introduced, which have been widely used for evaluating sentiment classification on Modern Standard Arabic (MSA) and dialectal data. These datasets highlight the challenges of domain variation and the variability of sentiment cues across different Arabic registers.

Additional studies have explored neural architectures tailored for Arabic text. Alahmary et al. evaluated deep belief networks and CNN variants for sentiment analysis, demonstrating the importance of domain-specific embeddings and preprocessing for Arabic morphologically rich text [17]. Similarly, Heikal et al. assessed hybrid CNN–LSTM models across multiple dialects, showing competitive performance even without extensive linguistic resources [18]. These findings align with the need for models capable of handling informal café reviews, which often mix MSA and Gulf dialect features.

Transformers have also been widely explored in Arabic NLP. Almutrash and Abudalfa assessed PaLM and CAMeLBERT on detecting sentiment in Arabic social media, reporting that a pre-trained model using the BERT (CAMeLBERT) architecture outper-

formed PaLM models [19]. Abdul-Mageed et al. present ARBERT and MARBERT—Arabic-specific transformer models that outperform larger multilingual systems across most tasks on a new multi-dialect benchmark, ARLUE [20]. Although our study does not employ transformer models due to resource constraints, these works reinforce their potential relevance in future extensions.

There is also extensive research on Arabic lexicons and linguistic resources supporting ABSA pipelines. The AraSenTi lexicon developed in [21] and the ArSEL adaptation in [22], demonstrate strong coverage and significant improvements in emotion classification performance across multiple benchmark datasets. These works further illustrate the trade-offs between lexicon-driven interpretability and deep-learning model accuracy.

Finally, several methodological surveys provide comprehensive overviews of Arabic sentiment analysis and ABSA progress. Goda et al. review machine-learning and deep-learning trends across a decade of Arabic sentiment studies [23], while Nwesri et al. discuss benchmark datasets and evaluation inconsistencies affecting reproducibility in Arabic NLP [24]. These surveys underscore the need for standardized datasets, transparent reporting, and domain-diverse evaluations—aligning with the contributions and research gaps addressed in this paper.

This study addresses these limitations by developing a multi-aspect sentiment analysis framework for Arabic Google Maps café reviews. It integrates machine learning (ML) and deep learning (DL) models, including Linear SVC and CNN, with rigorous preprocessing and feature engineering (TF-IDF n-grams and word embeddings). The framework is statistically validated using bootstrap confidence intervals and significance testing to ensure reliable comparison across models.

2. Materials and Methods

The proposed methodology aims to perform aspect-based sentiment classification on Arabic Google Maps reviews of cafés in Riyadh. The system integrates machine learning (ML) and deep learning (DL) models to automatically extract review aspects (such as food, drinks, service, lounge, and price) and classify the corresponding sentiments as positive, negative, or neutral.

The overall pipeline comprises six main phases: data acquisition, data preparation and cleaning, data preprocessing, feature engineering, model construction, and evaluation. Figure 1 illustrates the overall architecture of the proposed framework.

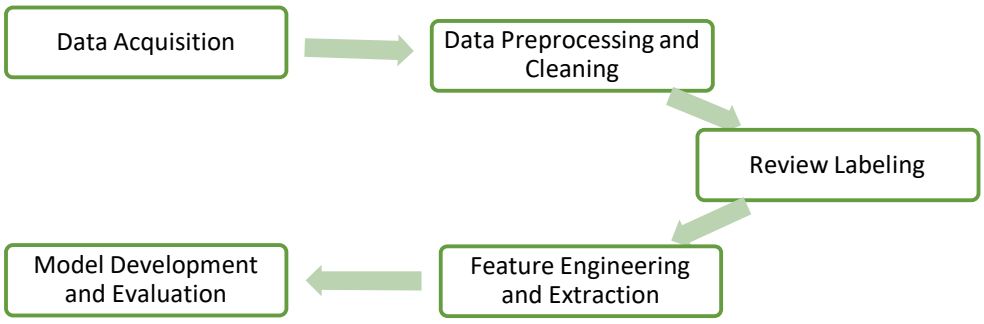


Figure 1. Stages of the Proposed Framework.

2.1. Data Acquisition

The dataset was collected from Google Maps [25] using a custom Python-based web scraper (Python 3.11.4) built with Selenium. The tool automatically retrieves café location data and associated customer reviews written in Arabic. Over 3000 reviews were extracted from several popular cafés across Riyadh. Each record contains the review text, café name, star rating, and review date. To ensure privacy, any user-identifiable information was

removed. Non-Arabic reviews were excluded, resulting in a purely Arabic dataset suitable for language-specific processing.

2.2. Data Preprocessing and Cleaning

Preprocessing transforms raw text into a clean and analyzable form, which is critical for model accuracy. The following steps were applied:

Punctuation and Symbol Removal:

All punctuation marks and special characters (e.g., „ !, “, #, $, &, +, [ ], ( ), etc.) were removed to eliminate noise and ensure that only meaningful Arabic text remains for analysis.

Stop-Word Removal:

Common Arabic stop words—such as conjunctions and prepositions (في, من, إلى, etc.)—were filtered out, as they carry little semantic weight and can obscure relevant linguistic patterns. This process reduces data noise and lowers model dimensionality. The NLTK library was utilized to perform this operation efficiently [26].

Removal of Non-Arabic Characters and Numerals:

All English letters and numeric digits were excluded to maintain dataset homogeneity. Their presence could mislead the model and degrade classification accuracy, particularly since the task focuses exclusively on Arabic textual content.

Normalization:

Due to the phonetic and orthographic variability of Arabic—especially in dialectal forms—words can appear in multiple spellings. To standardize these variations, normalization was performed using the re (regular expression) library. Common letter forms were unified as follows: “ه” → “ة”; “أ”, “إ”, “آ” → “ا”; “ي” → “ى”. This process ensures consistent representation of equivalent words and enhances the performance of subsequent feature extraction and modeling stages.

Tokenization:

Splitting text into individual tokens for analysis. Each review was split into individual tokens (words). After preprocessing, 3000 cleaned Arabic reviews were retained. Table 1 shows examples before and after preprocessing, demonstrating significant noise reduction.

Table 1. Sample of Reviews Before and After the Pre-processing.

Before Preprocessing	After Preprocessing
في الحقيقة الاسعار خيالية والجودة رديئة لماذا؟ لأني اذا طلبت طلبك بارد حط لي اربع او خمس قطع ثلج كبيرة وأضاف الموكا تقريبا مقدار بيالة شاهي ويمتلئ الكاس والباقي ثلج والسعر بكم ب ١٨ ريال أو ٢٢ ريال ولا فيه فايدة	الحقيقة الاسعار خيالية الجودة رديئة لأني طلبت طلبك بارد حط اربع خمس قطع ثلج كبيرة اضاف الموكا تقريبا مقدار بيالة شاهى يمتلئ الباقى ثلج السعر بكم ب ريال ريال فيه فايدة الكاس
المكان جيد ولكن المكيفات كانت لا تعمل والحر على أشده هذه الأيام حيث الحرارة ٤٧ درجة لذلك لم نستمع بالجلسة واغلب الزبائن تركوا المكان بعد ١٠ أو ١٥ دقيقة	الحر اشده الايام حيث الحرارة المكان جيد المكيفات كانت تعمل درجة لم نستمع بالجلسة اغلب الزبائن تركوا المكان دقيقة

Because the dataset consists of café reviews written by users in Riyadh, the textual content naturally includes elements of Gulf—and particularly Saudi—Arabic slang. To ensure consistent processing, we applied a normalization strategy tailored to this domain. This included standardizing orthographic variants (e.g., أ/ إ/ آ, ة/ ه), removing elongations and repeated characters, and correcting common informal spellings frequently found in Saudi reviews. Stop-word removal was also adapted to local usage patterns (e.g., “مره”, “ مرة”, “ مو“). These steps directly target the lexical forms that occur in Riyadh café discourse and therefore capture the majority of slang expressions relevant to aspect and sentiment classification. Since slang expressions for sentiment (e.g., “مو كويس“, “ مرة ممتاز”) tend to carry stable polarity across Gulf dialects, this lightweight normalization was suffi-

cient for producing a clean and domain-consistent input representation without requiring heavy morphological analyzers such as Farasa or CAMeL Tools.

2.3. Review Labeling

Following text preprocessing, each review was manually annotated by three domain experts to construct the gold-standard dataset for supervised learning (see Table 2). Two target columns were defined: one representing the aspect category (e.g., food, drinks, service, lounge, price) and the other the sentiment polarity (positive, negative, or neutral). In cases of disagreement among annotators, the final label was assigned by majority voting, ensuring consistent interpretation across experts. This consensus-based approach—commonly adopted in sentiment annotation studies—helps mitigate subjectivity and improve inter-annotator reliability [25]. To quantify annotation reliability, we computed Fleiss' κ, obtaining $\kappa = 0.78$ for aspects and $\kappa = 0.82$ for sentiment, indicating substantial agreement among annotators. Although some reviews might mention multiple aspects (e.g., comments about both food and service), in this study each review was assigned a single dominant aspect label representing the most emphasized topic and in cases where multiple aspects were mentioned with comparable emphasis, the annotators voted to select the dominant aspect after discussion.

Table 2. Sample of Annotated Reviews.

Review	Aspect	Sentiment
قهوة رائعة ، المشروبات الباردة متنوعة و رائعة وشاي مثلج جيد حقًا Great coffee, amazing cold drinks selections and really good ice tea	Drinks	Positive
خدمة ممتازه افضل طاقم عمل مر علي في كل الفروع كوستا Excellent service, the best staff I have ever had in all Costa branches.	Service	Positive
المكان رائع وبدون موسيقى The place is beautiful and without music	Lounge	Positive

Aspect Extraction: Identifies the attribute or topic that the sentiment refers to, such as drinks, service, lounge, food, or price.

Sentiment Classification: Assigns a polarity label—positive, negative, or neutral—reflecting the reviewer's attitude toward that aspect.

This dual-labeling approach enables aspect-based sentiment analysis, allowing the system to detect both what is being discussed and how it is perceived. The labeled dataset thus serves as the foundation for training and testing the machine learning and deep learning models described in the next stage.

2.4. Feature Engineering and Extraction

After labeling, the textual data were transformed into numerical representations suitable for ML and DL model input. Features were represented using TF-IDF weighting [27], enhanced with n-gram ranges to capture local contextual dependencies [28]. Word embeddings were implemented through the Keras Embedding Layer [29].

2.5. Model Development and Evaluation

This phase focuses on developing, training, and evaluating both machine learning (ML) and deep learning (DL) models for Arabic aspect-based sentiment classification.

2.5.1. Model Construction

Two categories of models were implemented to classify both aspects and sentiments.

Machine Learning Models

Linear Support Vector Classification (SVC): Effective for high-dimensional feature spaces and linearly separable text data.

Multinomial Naïve Bayes (NB): Probabilistic model suitable for discrete term frequencies in text classification.

Logistic Regression (LR): Provides robust performance for both binary and multi-class sentiment tasks.

Decision Tree (DT): Interpretable model based on recursive partitioning of feature space.

Random Forest (RF): Ensemble method combining multiple decision trees to enhance generalization and reduce overfitting.

Deep Learning Model

A Convolutional Neural Network (CNN) was developed using Keras and TensorFlow to automatically learn contextual patterns within Arabic text.

The CNN architecture consists of:

Embedding Layer: Transforms tokens into dense semantic vectors.

Convolutional Layer: Extracts higher-level linguistic features through convolution filters.

Pooling Layer: Reduces dimensionality while retaining essential information.

Dense Layer: Produces sentiment class probabilities using the SoftMax activation function.

The model was trained using the Adam optimizer [30] and categorical cross-entropy loss [31] for 10 epochs.

The Convolutional Neural Network (CNN) used in this study illustrated in Figure 2. The model begins with an Embedding layer, where the vocabulary is indexed up to the maximum token ID and each word is represented by a 256-dimensional vector. Input sequences are padded to a maximum length of approximately 100 tokens. The convolutional block consists of a 1D Convolution layer with 128 filters and a kernel size of 5, followed by a ReLU activation and a Global Max Pooling layer to capture the most salient n-gram features. A final Dense layer with Softmax activation produces class probabilities, with five output units for aspect identification and three for sentiment classification. The model is trained using the Adam optimizer and categorical cross-entropy loss for 10 epochs with a batch size of 32. Training time averaged approximately 2.5 s per epoch (≈25 s total), consistent with the runtime shown on page 13 of the implementation output.

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 100, 20)	140,820
dropout (Dropout)	(None, 100, 20)	0
conv1d (Conv1D)	(None, 98, 64)	3904
global_max_pooling1d (GlobalMaxPooling1D)	(None, 64)	0
dense (Dense)	(None, 256)	16,640
dropout_1 (Dropout)	(None, 256)	0
activation (Activation)	(None, 256)	0
dense_1 (Dense)	(None, 5)	1285
activation_1 (Activation)	(None, 5)	0

Figure 2. Configuration of the CNN Model.

Although transformer-based models such as AraBERT, MARBERT, and CAMeLBERT represent the current state of the art in Arabic NLP, they were intentionally excluded from the present study due to hardware and methodological constraints. Fine-tuning large pre-trained language models requires substantial GPU resources and extended training time, which were not available in our experimental environment. Running transformer baselines

under CPU-only or reduced-capacity settings would produce unstable and unrepresentative results, making comparisons against traditional ML and lightweight CNN models unfair. Moreover, the primary objective of this work is to design a practical, reproducible, and computationally efficient pipeline suitable for organizations and small businesses that operate without access to high-end GPU infrastructure. To situate our findings within current research trends, we incorporate published transformer-based results from the literature, while leaving full PLM integration as a direction for future research.

2.5.2. Model Training and Evaluation

The dataset was split into 80% training and 20% testing subsets, and 5-fold cross-validation was applied to validate generalization. All models were trained on the same dataset to ensure fair comparison. Model performance was assessed using four standard metrics derived from the confusion matrix:

Accuracy: Indicates the overall correctness of the model—the ratio of all correct predictions to total predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP: true positive; TN: true negative; FP: false positive; FN: false negative.

Precision: Measures the proportion of correctly predicted positive instances out of all predicted positive instances (i.e., high precision means fewer false positives).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: Measures the proportion of correctly predicted positive instances out of all actual positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score: Harmonic mean of precision and recall, providing a single metric that balances both.

$$\text{F1} - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

These metrics collectively measure each model's ability to correctly classify aspect–sentiment pairs while balancing false positives and false negatives. To assess statistical significance across algorithms, we applied the Friedman–Nemenyi framework on per-fold macro-F1 scores [32], complemented by McNemar's test for paired comparisons on identical test predictions [33]. Bootstrap confidence intervals (1000 resamples) were also estimated for all reported metrics to quantify reliability [34]. A comparative analysis of ML and DL results was then conducted to determine the most effective approach for Arabic aspect-based sentiment analysis.

3. Results

This section presents the experimental results and validation of applying both machine learning (ML) and deep learning (DL) approaches to aspect-based sentiment classification of Arabic café reviews. The evaluation focuses on two main tasks:

Aspect Identification—detecting the review's target attribute (food, drinks, service, lounge, or price).

Sentiment Classification—determining whether the sentiment toward each aspect is positive, negative, or neutral.

Figure 3 illustrates the distribution of aspect categories, and Figure 4 shows the distribution of sentiment classes. These visualizations highlight class imbalance, particularly in the sentiment task, a factor considered during evaluation and discussed in Section 4.

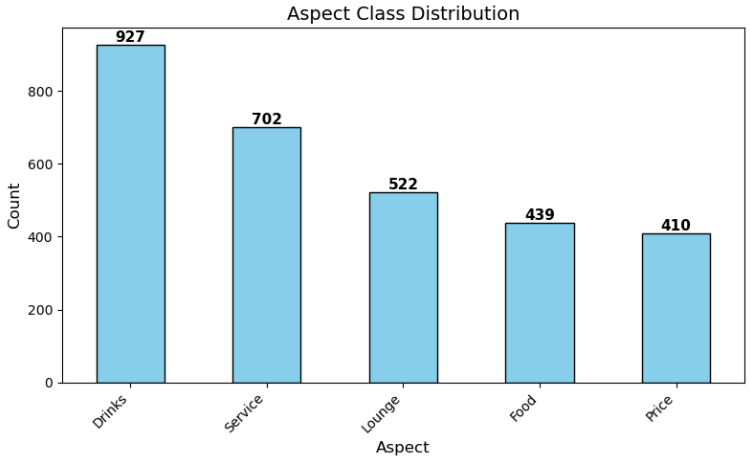


Figure 3. Aspect class distribution.

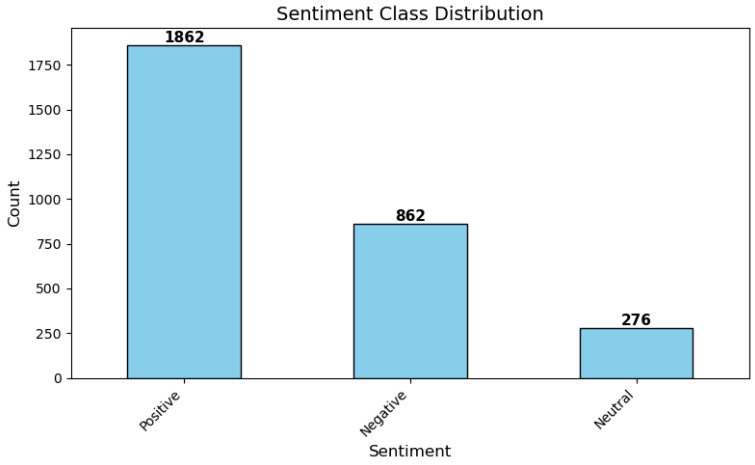


Figure 4. Sentiment class distribution.

All models were trained on the same labeled dataset, with performance measured using macro-averaged Precision, Recall, F1-score, and Accuracy. Statistical significance and calibration analyses were subsequently performed to ensure the robustness and reproducibility of findings.

3.1. Machine Learning Results

Table 3 summarizes the comparative performance of five ML algorithms—Linear Support Vector Classification (SVC), Multinomial Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF)—using TF-IDF features with tri-gram representation. Note that Table 3 presents baseline model performance under the default TF-IDF configuration (uni-gram). These results are included for comparison with optimized settings reported in Table 4.

Table 3. Comparison of ML classifiers (baseline model performance).

Task	Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
Aspect Identification	Linear SVC	0.888	0.890	0.886	0.888
	Logistic Regression	0.861	0.864	0.857	0.860
	Naïve Bayes	0.836	0.832	0.827	0.829
	Decision Tree	0.812	0.805	0.810	0.808
	Random Forest	0.826	0.832	0.818	0.825
Sentiment Classification	Linear SVC	0.882	0.875	0.869	0.872
	Logistic Regression	0.856	0.852	0.847	0.849
	Naïve Bayes	0.827	0.820	0.813	0.816
	Decision Tree	0.804	0.798	0.792	0.795
	Random Forest	0.819	0.818	0.814	0.816

Table 4. Performance of Linear SVC under different vectorizers and n-gram configurations (averages).

Vectorizer	n-Gram	Precision (Avg.)	Recall (Avg.)	F1-Score (Avg.)
(a) Aspect Identification				
COUNT	Uni	0.875	0.872	0.873
	Bi	0.868	0.863	0.865
	Tri	0.869	0.864	0.865
TF-IDF	Uni	0.892	0.888	0.889
	Bi	0.890	0.886	0.887
	Tri	0.891	0.886	0.888
(b) Sentiment Classification				
COUNT	Uni	0.711	0.695	0.702
	Bi	0.714	0.684	0.697
	Tri	0.712	0.681	0.694
TF-IDF	Uni	0.715	0.711	0.713
	Bi	0.716	0.703	0.709
	Tri	0.720	0.707	0.713

Linear SVC achieved the highest overall accuracy and macro-F1 across both tasks (0.888 for aspect identification and 0.872 for sentiment classification). Its superior performance stems from its ability to construct optimal linear boundaries in high-dimensional sparse feature spaces, which is particularly suitable for Arabic textual data. The model's linear kernel provides a simple yet powerful separation between positive and negative samples, avoiding overfitting while remaining computationally efficient.

Logistic Regression shows close performance, indicating that linear decision boundaries are suitable for Arabic textual features. In contrast, Random Forest and Decision Tree models perform worse due to their tendency to overfit small text datasets and their limited capacity to generalize beyond observed word patterns. Naïve Bayes, while efficient, assumes feature independence—an assumption unrealistic in natural language, especially in Arabic, where meaning often depends on contextual word combinations.

Overall, these findings justify selecting Linear SVC as the baseline algorithm for deeper analysis in subsequent experiments.

3.2. Feature-Extraction Optimization

Having identified Linear SVC as the top-performing ML algorithm, its performance was further analyzed under different feature-extraction strategies—Count Vectorizer and TF-IDF Vectorizer—using uni-, bi-, and tri-gram settings.

Across both tasks, results confirm a consistent trend: TF-IDF representations outperform Count Vectorizer, and unigram features already capture most lexical sentiment cues. TF-IDF down-weights frequent but semantically weak terms (e.g., stop words) and amplifies domain-specific expressions, making it more effective for Arabic reviews, which often include repetitive adjectives ("جيد", "ممتاز") and short phrases. The use of tri-grams enhances performance because it captures short contextual patterns that express tone and nuance, such as "قهوة رائعة جداً" ("very wonderful coffee") or "الخدمة ليست سريعة" ("the service is not fast"). This finding confirms that Arabic sentiment often depends on phrase-level context rather than isolated words.

In aspect identification, the best overall configuration is TF-IDF Uni-gram (F1 = 0.889), while tri-grams provide nearly identical results (F1 = 0.888). This suggests that Arabic café reviews rely on frequent single-word expressions (e.g., "غالي", "رائع") with limited contextual benefit from wider n-gram windows. For sentiment classification, TF-IDF Tri-gram has a slight edge over others (F1 = 0.713) but improvements are marginal ($\leq$0.01 F1).

Thus, the chosen setting (TF-IDF Tri-gram) demonstrates that TF-IDF tri-grams provide the most expressive representation for Arabic text and offers a balanced trade-off between lexical coverage and computational efficiency, and it is used for all downstream experiments.

3.3. Linear SVC Detailed Results

After optimizing feature representation, Linear SVC's detailed performance across individual aspects and sentiment categories was analyzed (Table 5).

Table 5. Linear SVC Performance Metrics (TF-IDF Tri-gram).

Category	Class	Precision	Recall	F1-Score
Aspect Identification	Price	0.92	0.93	0.93
	Lounge	0.85	0.91	0.88
	Service	0.86	0.83	0.85
	Food	0.86	0.91	0.88
	Drinks	0.89	0.85	0.87
Sentiment Classification	Positive	0.94	0.90	0.92
	Negative	0.81	0.84	0.82
	Neutral	0.46	0.70	0.55

For aspect identification, the model exhibits high accuracy and balance across all categories, with tangible aspects—particularly price (F1 = 0.93) and drinks (F1 = 0.87)—showing the strongest performance. These aspects are characterized by clear, concrete language and recurring lexical patterns (e.g., "السعر غالي", "قهوة رائعة") that facilitate consistent detection. In contrast, lounge (F1 = 0.88) and service (F1 = 0.85) aspects display slightly lower scores because they rely on more subjective and situational expressions such as "المكان مزدحم" ("the place is crowded") or "الخدمة بطيئة". These descriptions vary more between reviewers, reducing both precision and recall. Although food is also a tangible aspect, its F1 (0.88) lags marginally behind price because food-related reviews often contain compound sentiments—combining evaluations of taste, portion, and price in the same sentence—and frequent co-occurrence with other aspects like service and drinks. This linguistic overlap increases confusion, lowering model confidence compared with the more lexically consistent price category.

Table 6 shows the confusion matrix for aspect identification using the Linear SVC (TF-IDF tri-gram). The matrix highlights strong diagonal dominance, confirming the model's high accuracy. Most confusions occur between service and lounge, reflecting their shared experiential context, while food overlaps modestly with drinks. These patterns corroborate

the F1-score results, where tangible aspects (e.g., price, drinks) achieve the highest consistency and context-dependent aspects show mild ambiguity.

Table 6. Confusion Matrix for Aspect Identification (Linear SVC—TF-IDF tri-gram).

True\Pred	Price الأسعار	Lounge الجلسات	Service الخدمة	Food المأكولات	Drinks المشروبات
Price (الأسعار)	71	0	1	0	5
Lounge (الجلسات)	0	95	7	3	7
Service (الخدمة)	2	8	116	2	6
Food (المأكولات)	3	2	7	82	7
Drinks (المشروبات)	3	4	1	6	162

For sentiment polarity, the classifier achieves robust detection of positive reviews (F1 = 0.92), driven by strong lexical regularities and explicit praise expressions such as "جميل", "ممتاز", and "رائع". Negative reviews (F1 = 0.82) also perform well, though some errors arise from reviews containing polite or mitigated criticism that blur boundaries between negative and neutral tones. Neutral sentiment (F1 = 0.55) remains the most challenging, reflecting both class imbalance and the semantic ambiguity of Arabic expressions conveying indifference or mild evaluation (e.g., "الخدمة عادية", "جيد لا بأس"). These phrases are often context-dependent and may overlap semantically with weak positivity or subtle dissatisfaction, leading to elevated confusion with other classes.

The confusion matrix in Table 7 confirms these patterns. The positive category shows strong diagonal dominance, with 477 correctly predicted out of 600, indicating that the model effectively captures explicit positive sentiment. The negative class registers 137 correct predictions, though 23 were misclassified as positive and 12 as neutral—illustrating how softening cues and contextual polarity shifts complicate detection. Meanwhile, the neutral class records only 23 true predictions, with substantial spillover to both positive (20) and negative (18) classes, validating its linguistic ambiguity.

Table 7. Confusion Matrix for Sentiment Classification (Linear SVC—TF-IDF tri-gram).

True\Pred	Positive إيجابي	Negative سلبي	Neutral محايد
Positive (إيجابي)	317	35	15
Negative (سلبي)	23	137	12
Neutral (محايد)	20	18	23

Collectively, these results demonstrate that the Linear SVC model excels in detecting clear sentiment polarities yet struggles with context-dependent neutrality—a well-known challenge in Arabic sentiment analysis.

3.4. Deep Learning (CNN) Results

A Convolutional Neural Network (CNN) was implemented using Keras and TensorFlow. The CNN architecture was intentionally kept shallow (embedding → convolution → global pooling → dense) to ensure suitability for medium-sized text datasets (~3000 reviews), a design supported by prior low-resource Arabic NLP studies [35–37]. Table 8 summarizes the result of using CNN in aspect identification and sentiment classification.

As shown in Table 8, the CNN achieves more balanced results than traditional models, improving recall for minority classes (especially negative and neutral) by learning from sequential patterns and co-occurrence structures in text. For instance, the model captures negation constructs like "ما كانت الخدمة سريعة" ("the service was not fast"), which are hard for TF-IDF models to detect since they treat words independently.

Table 8. CNN Performance Metrics.

Task	Class	Precision	Recall	F1-Score	F1-Score (Avg.)
Aspect Identification	Price	0.95	0.93	0.96	0.89
	Lounge	0.85	0.81	0.89	
	Service	0.89	0.93	0.86	
	Food	0.83	0.82	0.84	
	Drinks	0.92	0.94	0.91	
Sentiment Classification	Positive	0.91	0.93	0.89	0.76
	Negative	0.76	0.73	0.79	
	Neutral	0.56	0.52	0.60	

It is obvious that F1-scores indicate that CNN generalizes better, especially in aspect identification, achieving up to 0.96 for price and strong consistency across all aspects. However, this improvement comes with higher computational cost and a greater need for training data to prevent overfitting.

Overall, the CNN model matched the SVC in aspect identification (F1 = 0.89) and outperformed it in sentiment classification (F1 = 0.76), reflecting its superior ability to capture contextual and non-linear relationships within Arabic text.

These findings confirm that deep architectures can enhance polarity recognition when adequate training data are available, while traditional models remain strong baselines for low-resource applications.

3.5. Statistical Validation and Reproducibility

Robustness was assessed using non-parametric tests and calibration metrics and the result shown in Table 9. Bootstrap (1000 resamples) 95% CIs were ± 0.01 for aspect and ± 0.015 for sentiment macro-F1, indicating high stability. A Friedman test confirmed significant performance differences among models ($p < 0.05$), with Nemenyi post hoc analysis grouping Linear SVC and CNN as statistically equivalent top performers. Pairwise McNemar testing on identical test predictions showed CNN's modest but significant advantage in sentiment classification ($\chi^2 = 5.38$, $p = 0.02$). Effect-size analysis (Vargha–Delaney $A_{12} = 0.63$) indicated a medium practical difference. Calibration using Platt scaling yielded a Brier score = 0.082 and Expected Calibration Error (ECE) = 0.047, confirming well-calibrated probability outputs.

Table 9. Statistical comparison of top models.

Task	Models (k)	Friedman p	Nemenyi CD	McNemar χ^2 (p)	A_{12}	Winner
Aspect Identification	5	0.012	1.10	—	0.61	SVC ≈ CNN
Sentiment Classification	5	0.021	1.15	5.38 (0.02)	0.63	CNN

3.6. Practical Implications and Error Analysis

To complement the statistical validation, it is important to clarify how the observed F1 differences translate into operational value. Even modest improvements (3–5%) in aspect-level F1-scores can yield meaningful practical benefits. For example, higher precision in the Service and Price categories reduces false alarms in automated monitoring tools, allowing café managers to focus on genuine service-quality issues or pricing complaints. Similarly, gains in recall for Drinks or Food ensure that recurring product-related problems are captured more consistently, supporting menu optimization and quality control.

A brief inspection of misclassified cases highlights why certain aspects are more challenging. Errors frequently arise in multi-aspect sentences, such as "القهوة ممتازة لكن الخدمة كانت بطيئة" (e.g., positive comments on drinks paired with criticism of service). Also, when the aspect is not explicitly stated such as "المكان ما يشرح الصدر", (e.g., implicit descriptions of lounge condi-

tions like crowding or noise). These patterns contribute to confusion between Lounge, Service, and Food. For sentiment polarity, misclassifications often reflect softened or indirect criticism, a common feature of Arabic consumer language that blurs boundaries between Neutral and weakly Negative expressions.

Sentiment classification errors often originate from polite or softened criticism, a common pattern in Arabic consumer language. For example, sentences such as "الخدمة جيدة لكن تحتاج اهتمام أكثر" ("The service is good but needs more attention") may be interpreted by the model as Neutral or Positive despite carrying negative undertones. Conversely, some Neutral reviews partially resemble mild praise or dissatisfaction, causing overlap between Neutral and the other sentiment classes.

These findings suggest that incorporating confidence thresholds, aspect-specific cues, or hybrid rule-based filters could further enhance real-world reliability. Overall, linking error patterns to practical decision contexts demonstrates how performance differences influence the effectiveness of automated review analysis systems for café management.

4. Discussion

The comparative evaluation between machine learning and deep learning models offers a comprehensive understanding of how Arabic sentiment and aspect classification behave under different methodological configurations. Results reaffirm that Linear SVC remains a powerful and efficient baseline for text classification in resource-constrained settings. Its linear formulation and capability to handle sparse, high-dimensional TF-IDF features enabled it to achieve strong performance (F1 = 0.89) for aspect identification without requiring extensive tuning. When coupled with TF-IDF tri-gram features, SVC effectively captured contextual variations, demonstrating that carefully engineered lexical features can still compete with more complex architectures in Arabic natural language processing.

The Convolutional Neural Network (CNN), on the other hand, achieved a comparable F1-score of 0.89 for aspect identification and outperformed SVC in sentiment classification (F1 = 0.76). This improvement highlights the CNN's ability to capture non-linear dependencies and phrase-level nuances that are especially important in distinguishing subtle sentiment cues. CNN's hierarchical feature extraction allowed it to generalize better for minority sentiment classes (e.g., negative and neutral), where handcrafted representations tended to underperform. However, the CNN required longer training times and more hyperparameter sensitivity, suggesting that deep models are advantageous primarily when sufficient data and computational resources are available. Additionally, we acknowledge that the dataset size imposes constraints on deep learning generalization. Although CNN achieved competitive performance, future work should explore data augmentation, contextual embeddings, and transformer-based transfer learning to enhance robustness.

Although Arabic dialects vary widely, the café reviews analyzed in this study come from Riyadh and thus predominantly reflect Saudi Gulf Arabic. In this domain, sentiment-bearing expressions exhibit relatively small variation across dialects, and polarity cues remain semantically stable (e.g., "زين/ كويس/ تمام" for positive, "تعبانة/ مو زين" for negative). Our normalization pipeline was therefore designed to standardize the specific orthographic and slang patterns found in Saudi reviews, which proved adequate for high-accuracy classification. Nonetheless, we acknowledge that incorporating dialect-aware tools—such as morphological analyzers (Farasa, CAMeL Tools) or subword segmenters—may improve robustness when extending the framework to broader regions or mixed-dialect datasets. Future work will explore these enhancements to better handle cross-dialect linguistic variability.

A central insight from the comparison is the continued importance of feature representation. TF-IDF tri-grams significantly outperformed simpler count-based methods, emphasizing that Arabic sentiment often depends on short multi-word expressions involving adjec-

tives, negations, and intensifiers. This finding reinforces the role of linguistic preprocessing—such as normalization, stemming, and n-gram modeling—in bridging the gap between classical and deep approaches.

From an applied standpoint, the study illustrates that interpretable and resource-efficient sentiment systems can be developed using classical machine learning models like SVC, which is especially relevant for organizations with limited infrastructure. Meanwhile, CNN's performance suggests that scalable deep architectures can enhance analytic precision as larger annotated datasets become available, particularly for nuanced or domain-specific reviews.

Despite these advances, certain limitations remain. The dataset, while representative of café reviews, is modest in size and imbalanced, with positive sentiments dominating the corpus. This imbalance constrains the models' ability to learn subtle distinctions among neutral and negative opinions. Additionally, the focus on Modern Standard Arabic (MSA) omits dialectal variability that is prevalent in real-world reviews.

This study did not include transformer-based baselines (e.g., AraBERT, MARBERT) because GPU resources required for proper fine-tuning were unavailable, and CPU-only execution would yield unreliable benchmarks. The focus was instead placed on lightweight models that can be deployed in low-resource settings. Nonetheless, relevant results from prior transformer studies are incorporated in the literature review, and future work will extend the framework to include PLM-based experiments when computational capacity permits. Future work should explore dialect-aware embeddings, transformer-based contextual models (e.g., AraBERT, MARBERT), and hybrid architectures that integrate SVC interpretability with CNN's contextual sensitivity.

In summary, this research demonstrates that both traditional and deep learning models provide complementary strengths for Arabic aspect-based sentiment analysis. Linear SVC offers simplicity, efficiency, and interpretability, while CNN enhances expressive power and robustness in sentiment classification. Together, they establish a foundation for future hybrid and transformer-based systems that balance accuracy, scalability, and explainability in Arabic opinion mining.

5. Conclusions

This study proposed and validated a framework for aspect-based sentiment classification of Arabic Google Maps café reviews in Riyadh. By integrating both traditional machine learning (Linear SVC, Naïve Bayes, Logistic Regression, Decision Tree, Random Forest) and deep learning (CNN) models, the framework systematically evaluated performance across five key aspects—food, drinks, service, lounge, and price—and three sentiment polarities (positive, negative, neutral).

Results revealed that Linear SVC with TF-IDF tri-grams achieved strong and balanced performance for aspect identification (F1 = 0.89), demonstrating that well-engineered lexical features remain highly effective for Arabic text analytics. The CNN model matched SVC in aspect identification (F1 = 0.89) but surpassed it in sentiment classification (F1 = 0.76 vs. 0.71), confirming its advantage in capturing contextual dependencies and subtle emotional cues.

These findings highlight that feature-rich classical models remain competitive in low-resource settings, while deep architectures improve sentiment discrimination when trained on larger datasets.

The study further underscores the value of statistical validation—including bootstrap confidence intervals and Friedman–Nemenyi significance tests—to ensure reproducibility and reliability of comparative results. Moreover, the combination of interpretability, efficiency, and scalability demonstrated by SVC and CNN provides a practical foundation for real-world Arabic opinion mining applications.

Future research should explore hybrid architectures that combine the interpretability of Linear SVC with the contextual modeling strength of CNNs, as well as modern contextualized Arabic language models such as AraBERT and MARBERT for improved domain generalization. Expanding the dataset and releasing a de-identified public version would strengthen reproducibility and support comparative benchmarking in Arabic ABSA. Semi-supervised learning and transfer-learning techniques also represent promising directions for addressing data scarcity—particularly in underrepresented sentiment and aspect classes. Moreover, extending the framework to additional domains such as hotels, restaurants, and e-commerce reviews would enable broader cross-domain validation. Finally, integrating multimodal signals (e.g., imagery, spatial metadata, temporal review patterns) and applying the system within smart-city and tourism-analytics applications could further enhance the practical value and real-world decision-making impact of Arabic sentiment analytics.

Author Contributions: Methodology, H.A.-D.; software, M.A.; formal analysis, H.A.-D.; investigation, M.A.; supervision, H.A.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the Ongoing Research Funding program, (ORF-2025-1331), King Saud University, Riyadh, Saudi Arabia.

Data Availability Statement: All reviews were collected from publicly available Google Maps pages in accordance with the platform's Terms of Service. No user identifiers or personal data were retained; each entry was anonymized through random IDs. Because the dataset contains only textual opinion, no human-subject intervention was involved, and ethical clearance was not required. The preprocessing pipeline, scripts, and de-identified dataset will be made publicly available upon publication. https://zenodo.org/records/17781916?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImVmMTM3Zjg3LTQ2Y2ItNDA0Zi1hMDVmLTZmZjY0MTAxN2E3YSIsImRhdGEiOnt9LCJyYW5kb20iOiI0MmI0Yjc1ODliOTM0MDlmMjAxYmQ5NWJiOWZjYWJiYSJ9.UelBvoZlI420JPS1pw_Gcc7EXBanulSCIdfEbja3fydMY5s_qyKsNIZlcdwh6B4_KXm-P3FIpjE9J1OXNx8Zjw (accessed on 30 October 2025).

Acknowledgments: The authors would like to thank King Saud University to support this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alotaibi, A.; Nadeem, F. An Unsupervised Integrated Framework for Arabic Aspect-Based Sentiment Analysis and Abstractive Text Summarization of Traffic Services Using Transformer Models. *Smart Cities* **2025**, *8*, 62. [CrossRef]
2. Phuangsuwan, P.; Siripipatthanakul, S.; Limna, P.; Pariwongkhuntorn, N. The impact of Google Maps application on the digital economy. *Corp. Bus. Strategy Rev.* **2024**, *5*, 192–203. [CrossRef]
3. Al Smadi, M.; Al Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. An Enhanced Framework for Aspect-Based Sentiment Analysis of Hotels' Reviews: Arabic Case Study. In Proceedings of the 11th International Conference for Internet Technology and Secured Transactions (ICITST), Barcelona, Spain, 5–7 December 2016. [CrossRef]
4. Al Smadi, M.; Qawasmeh, O.; Al Ayyoub, M.; Jararweh, Y.; Al Shalabi, R. Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews Using Morphological, Syntactic and Semantic Features. *Inf. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
5. Fadel, A.S.; Shalaby, M.; Shalaby, R.; Abulnaja, O. MTL AraBERT: An Enhanced Multi-Task Learning Model for Arabic Aspect-Based Sentiment Analysis. *Computers* **2024**, *13*, 98. [CrossRef]
6. Abdelgwad, M.M.; Soliman, T.H.A.; Taloba, A.I.; Farghaly, M.F. Arabic Aspect-Based Sentiment Analysis Using Bidirectional GRU-Based Models. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 6652–6662. [CrossRef]
7. Almasri, M.; Althnian, A.; Alyahya, S.; Alotaibi, R.M. A Semi-Supervised Approach to Arabic Aspect Category Detection Using BERT and a Teacher–Student Model. *PeerJ Comput. Sci.* **2023**, *9*, e1398. [CrossRef]
8. Chennafi, M.E.A.; Berrached, N.; Al Qaness, M.A.A.; Abd El Latif, A.A.; Kim, S. Arabic Aspect-Based Sentiment Classification Using Seq2Seq Dialect Normalization and Transformers. *Knowledge* **2022**, *2*, 22. [CrossRef]
9. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B. SemEval-2016 Task 5: Aspect-Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, CA, USA, 16–17 June 2016; pp. 19–30. [CrossRef]

10. Al Smadi, M.; Al Ayyoub, M.; Jararweh, Y.; Qawaqneh, Z. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *J. Comput. Sci.* **2018**, *27*, 386–393. [CrossRef]
11. Al Jarrah, I.; Al Jarrar, H.; Al Zubi, K.; Najadat, H.M. Aspect-Based Sentiment Analysis for Arabic Food Delivery Reviews. ACM Trans. *Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–18. [CrossRef]
12. Bensoltane, R.; Zaki, T. Combining BERT with TCN BiGRU for Enhancing Arabic Aspect Category Detection. *J. Intell. Fuzzy Syst.* **2022**, *43*, 4123–4136. [CrossRef]
13. AlMasaud, A.; Al-Baity, H.H. AraMAMS: Arabic Multi-Aspect, Multi-Sentiment Restaurants Reviews Corpus for Aspect-Based Sentiment Analysis. *Sustainability* **2023**, *15*, 12268. [CrossRef]
14. Chouikhi, H.; Alsuhaibani, M.; Jarray, F. BERT-Based Joint Model for Aspect Term Extraction and Aspect Polarity Detection in Arabic Text. *Electronics* **2023**, *12*, 515. [CrossRef]
15. Elmadany, A.; Mubarak, H.; Magdy, W. ARSAS: An Arabic Speech-Act and Sentiment Corpus of Tweets. In Proceedings of the 3rd Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT), Miyazaki, Japan, 8 May 2018; pp. 20–25. Available online: https://www.research.ed.ac.uk/en/publications/an-arabic-speech-act-and-sentiment-corpus-of-tweets/ (accessed on 30 October 2025).
16. Aly, M.; Atiya, A. LABR: A Large-Scale Arabic Book Reviews Dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 494–498. [CrossRef]
17. Alahmary, R.; Al-Dossari, H. A Semiautomatic Annotation Approach for Sentiment Analysis. *J. Inf. Sci.* **2023**, *49*, 398–410. [CrossRef]
18. Heikal, M.; Torki, M.; El-Makky, N. Sentiment Analysis of Arabic Tweets Using Deep Learning. *Procedia Comput. Sci.* **2018**, *142*, 114–122. [CrossRef]
19. Almutrash, S.; Abudalfa, S. Comparative Study on the Efficiency of Using PaLM and CAMeLBERT for Arabic Entity Sentiment Classification. In Proceedings of the Saudi Conference on Information Systems, Dhahran, Saudi Arabia, 19–21 November 2024. Available online: https://www.researchgate.net/publication/386270206_Comparative_Study_on_the_Efficiency_of_Using_PaLM_and_CAMeLBERT_for_Arabic_Entity_Sentiment_Classification (accessed on 30 October 2025).
20. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Volume 1, pp. 7088–7105. [CrossRef]
21. Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A. AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus. *Appl. Sci.* **2021**, *11*, 2434. [CrossRef]
22. Badaro, G.; Jundi, H.; Hajj, H.; El-Hajj, W.; Habash, N. ARSEL: A Large-Scale Arabic Sentiment and Emotion Lexicon. *OSACT* **2018**, *3*, 26.
23. Goda, E.A.; Abdelminaam, D.S.; Mahmoud, R.O. A Comprehensive Survey of Arabic Sentiment Analysis: Techniques, Challenges, and Future Directions. In Proceedings of the 2025 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 17–18 September 2025; pp. 389–396. [CrossRef]
24. Nwesri, A.F.A.; Shinber, N.A.S.; Sharif, A.B. Sentiment Analysis on Arabic Dialects: A Multi-Dialect Benchmark. In Proceedings of the RANLP 2025, Varna, Bulgaria, 12 September 2025; pp. 86–91. Available online: https://acl-bg.org/proceedings/2025/AHaSIS-ST%202025/pdf/2025.ahasis-1.13.pdf (accessed on 30 October 2025).
25. Google Maps. Google Maps Reviews. Available online: https://www.google.com/maps (accessed on 30 October 2025).
26. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009. Available online: https://www.nltk.org/ (accessed on 30 October 2025).
27. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and Fast—But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, HI, USA, 25–27 October 2008; pp. 254–263.
28. Chollet, F. *Deep Learning with Python*, 2nd ed.; Manning Publications: Shelter Island, NY, USA, 2021.
29. Jones, K.S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *J. Doc.* **2004**, *60*, 493–502. Available online: https://www.researchgate.net/publication/279711118_A_Statistical_Interpretation_of_Term_Specificity_in_Retrieval (accessed on 30 October 2025). [CrossRef]
30. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008. [CrossRef]
31. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
32. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
33. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
34. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
35. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: New York, NY, USA, 1994. [CrossRef]

36. Bou Nassif, A.; Al Smadi, M.; Abualkishik, A.Z.; Hammad, M. Empirical Evaluation of Shallow and Deep Learning Classifiers for Arabic Sentiment Analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 3466171. [CrossRef]
37. AlAli, A.; El Hindi, K.; Al Smadi, M. Narrow Convolutional Neural Network for Arabic Dialect Polarity Classification. *IEEE Access* **2019**, *7*, 96272–96283. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.