Enhancing Customer Experience Through Arabic Aspect-Based Sentiment Analysis of Saudi Reviews

Razan Alrefae, Revan Alqahmi, Munirah Alduraibi, Shatha Almatrafi, Asmaa Alayed
College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia

***Abstract*—Big brands thrive in today's competitive marketplace by focusing on customer experience through product reviews. Manual analysis of these reviews is labor-intensive, necessitating automated solutions. This paper conducts aspect-based sentiment analysis on Saudi dialect product reviews using machine learning and NLP techniques. Addressing the lack of datasets, we create a unique dataset for Aspect-Based Sentiment Analysis (ABSA) in Arabic, focusing on the Saudi dialect, comprising two manually annotated datasets of 2000 reviews each. We experiment with feature extraction techniques such as Part-of-Speech tagging (POS), Term Frequency-Inverse Document Frequency (TF-IDF), and n-grams, applying them to machine learning algorithms including Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). Our results show that for electronics reviews, RF with TF-IDF, POS tagging, and tri-grams achieves 86.26% accuracy, while for clothes reviews, SVM with TF-IDF, POS tagging, and bi-grams achieves 86.51% accuracy.**

Keywords*—Customer experience; Arabic natural language processing; sentiment analysis; Arabic Aspect-Based Sentiment Analysis; online reviews; review analytic; e-commerce; business owners

I. INTRODUCTION

Ensuring customer satisfaction is crucial for businesses of all sizes, with successful companies like Amazon thriving on their customer-centric approaches [1]. Analyzing online reviews helps manufacturers identify areas for improvement, leading to increased customer satisfaction and sales. For example, a major appliance manufacturer saw a 17% sales increase after addressing complaints identified through online reviews. However, manual analysis of these reviews is time-consuming and labor-intensive [2]. This paper addresses the challenge of sentiment analysis in Arabic, specifically focusing on Saudi dialects, by developing a unique dataset and evaluating the performance of various machine learning algorithms. Existing research on Arabic Aspect-Based Sentiment Analysis (ABSA) is limited, particularly in the Saudi context, making it difficult to establish a robust baseline or conduct comprehensive comparisons.

To fill this gap, we created two datasets of 2000 reviews each, manually annotated for sentiment analysis. We employed feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Part-of-Speech (POS) tagging, and tested machine learning algorithms including Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN). This approach allows us to navigate the linguistic diversity of Saudi Arabian dialects effectively.

Recent research in aspect-based sentiment analysis (ABSA) within Arabic contexts has provided valuable insights and methodologies, as evidenced by a number of notable works in the field.

Mir et al. [3] focused on developing a model for aspect-based opinion mining tailored to social reviews, which are typically longer and more complex than product reviews. The model focuses on auto-tagging and data training, defining feature sets, and utilizing a dictionary. Achieving an accuracy of 98.17% with precision, recall, and F1 scores of around 96%, the proposed model outperforms CR and Naïve Bayes classifiers. Future work aims to identify implicit aspects and refine aspect-wise sentiment analysis without relying on dictionaries.

In 2022, researchers aimed to support Saudi government efforts by analyzing user reviews on governmental mobile applications [4]. The labeled dataset underwent preprocessing, and features like supervised lexicon weights, TF-IDF, terms frequency matrix (TFM), and terms document matrix (TDM) were extracted. Using a supervised automatic sentiment lexicon, these features were weighted for each record. Four classifiers were employed and trained, with the highest accuracy recorded for DT (59.92%), KNN (78.46%), NB (54.78%), and SVM (55.38%).

In 2022, a sentiment analysis on customer satisfaction with logistics services in Saudi Arabia's private and public sectors during the COVID-19 pandemic was conducted [5]. Using a lexicon-based approach, 67,124 tweets were classified as positive, negative, or neutral, with preprocessing involving text cleaning and tokenization. The TF-IDF algorithm was employed to adjust word weights, and an SVM classifier achieved an average accuracy of 82% in 3-class classification, along with 81% precision and 80% recall.

Researchers aimed to assess customer opinions on mobile banking applications for updates and maintenance [6]. They manually collected an Arabic dataset and applied four ML techniques (NB, KNN, DT, and SVM). The NB model stood out, achieving 89.65% accuracy, 88.08% recall, 88.25% precision, and an f-score of 88.25%.

In 2023, researchers analyzed over 120,000 Arabic reviews on telecommunications services in Saudi Arabia [7]. They employed a machine learning approach, utilizing the SVM model for sentiment analysis. The study identified many factors influencing customer sentiments and recommendations for improvement were provided.

Research on ABSA in Arabic is advancing, but only a few NLP techniques have been tested on proposed machine learning models. In preprocessing, some studies treat negation words as stop words, which can misrepresent context, and disregard natural polarities, harming model performance. Datasets in literature have limitations like restricted availability, small size, domain specificity, imbalance, and being collected from non-Arabic sources or single regions, affecting generalizability. They often prioritize Modern Standard Arabic over dialects and use labor-intensive techniques. Our study addresses these gaps by applying diverse preprocessing and feature extraction techniques to evaluate their impact on model accuracy, addressing Saudi dialect complexities and resource scarcity in Arabic compared to English. The dataset has been carefully collected and prepared to enhance the training process.

II. Background

Natural language processing (NLP) is a branch of Artificial Intelligence (AI) that enables computers to understand, generate, and manipulate human language [8].

The linguistic challenges of Arabic in NLP are significant due to its dynamic nature, including lexical changes, regional variations, and context-dependent interpretations [9]. With over 400 million Arabic speakers [10], the diversity in dialects complicates the development of universally effective NLP applications. Arabic's rich morphology, where prefixes and suffixes alter meanings, and its orthographic connectedness, where a letter's form changes with its position, further complicate NLP tasks [11]. Additionally, diacritical marks, or "harakat," create orthographic ambiguity. A shortage of large-scale, high-quality labeled data for Arabic, crucial for training accurate supervised learning models, combined with the need for diverse labeled data to account for dialectal variations, hinders the advancement of Arabic NLP [12]. Sentiment Analysis (SA) uses NLP techniques to classify the polarity of text such as documents or review as positive, negative, or neutral. SA operates at four levels: document-level analysis determines overall polarity, sentence-level analysis classifies sentences as subjective or objective, and aspect-based analysis evaluates an object's features to determine the polarity of each, considering the context of opinion words. Aspect-based summarization aggregates opinions and attributes to summarize feedback and derive insights [13]. ABSA includes tasks such as Aspect Term Extraction, Aspect Term Polarity, Aspect Category Identification, and Aspect Category Polarity, as demonstrated in the sentence: " حلو الستايل والقماش حلو وخفيف، بس اللون مو نفس الصورة" ("The style is nice and the fabric is nice and light, but the color is not the same as the picture"). Machine learning approaches for SA include supervised, unsupervised, and semi-supervised learning, lexicon-based methods, rule-based approaches, and deep learning [14–16]. These techniques, particularly deep learning, which uses neural networks, offer superior performance in NLP tasks.

III. Methodology

In this work, we aim to enhance customer experience by conducting aspect-based sentiment analysis on reviews written in the Saudi Arabian dialect. As depicted in Fig. 1, the methodology involves data collection and annotation, cleaning and preprocessing, feature extraction, algorithm selection, and model training.

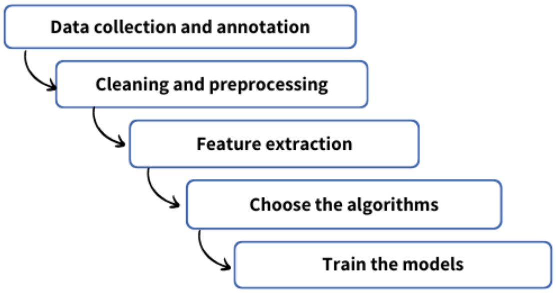


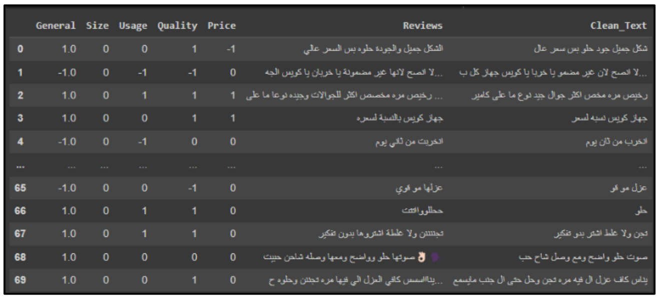
Fig. 1. Schematic representation of the research methodology.

A. Data Collection and Annotation

Due to the lack of established datasets, we manually collected 4000 product reviews (2000 electronics, 2000 clothing) in Saudi dialects [17], ensuring relevance and accuracy. Reviews were sourced from platforms like Amazon, Jarir Bookstore, Shein, Noon, and Twitter (X). Annotation involved labeling aspects such as quality, battery, and price for electronics, and size, color, and fabric for clothing, assigning sentiment scores of '1' (positive), '-1' (negative), or '0' (neutral). We focused on the four most recurring features for each category to streamline model input, maintaining a balanced sentiment distribution. During the training phase, we manually conducted term extraction and polarity assignment for the highest accuracy. Each review was carefully annotated by identifying specific aspects and marking their presence with binary indicators, as well as recording the overall polarity of each review as positive, negative, or neutral. To enhance reliability and objectivity, we split the researchers into two groups who cross-reviewed each other's annotations, ensuring thorough validation. This rigorous process ensured the accuracy and reliability of the annotations, thus enhancing the overall performance of our sentiment analysis models.

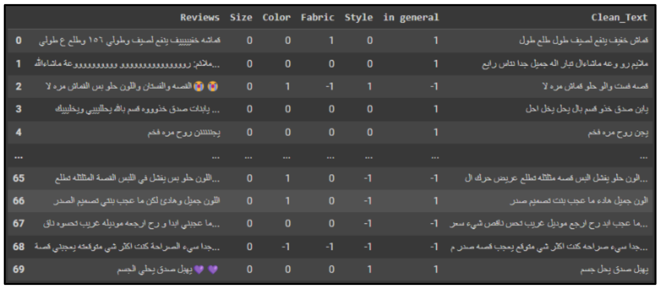
B. Cleaning and Preprocessing

Cleaning and preprocessing are crucial for enhancing data quality in complex languages like Arabic. We removed null values, emojis, punctuation, symbols, English letters, Arabic diacritics, and stop words (while preserving those essential for opinion comprehension). Tokenization and stemming were applied, and specific Arabic characters were normalized. This thorough approach ensures dataset robustness and reliability, aligning with scientific rigor for meaningful analysis. Fig. 2 and Fig. 3 illustrate the datasets before and after cleaning.



	General	Size	Usage	Quality	Price	Reviews	Clean_Text
0	1.0	0	0	1	-1	الشكل جميل والجودة حلوه بس السعر عالي	شكل جميل جود حلو بس سعر عال
1	-1.0	0	-1	-1	0	لا انصح لانها غير مضمونة يا خربان يا كويس الجه...	لا انصح لان غير مضمو يا خربا يا كويس جهاز كل ب...
2	1.0	0	1	1	1	رخيص مره مخصص اكثر للجوالات وجيده نوعا ما على ...	رخيص مره مخص اكثر جوال جيد نوع ما على كامير
3	1.0	0	0	1	1	جهاز كويس بالنسبة لسعره	جهاز كويس نسبه لسعر
4	-1.0	0	-1	0	0	انخربت من ثاني يوم	انخرب من ثان يوم
...	...	...	...	...	...	...	...
65	-1.0	0	0	-1	0	عزلها مو قوي	عزل مو قو
66	1.0	0	1	1	0	حلللوواااتت	حلو
67	1.0	0	1	1	0	تجننن ولا غلطة اشتروها بدون تفكير	تجن ولا غلط اشتر بدو تفكير
68	1.0	0	0	0	0	💜 👌 صوتها حلو وواضح ومعها وصله شاحن حبيت	صوت حلو واضح ومع وصل شاح حب
69	1.0	0	0	1	0	يااااسسس كافي العزل الي فيها مره تجنن وحلوه ح...	يئاس كاف عزل ال فيه مره تجن وحل حتى ال جنب مايسمع

Fig. 2. Electronics dataset before and after cleaning.



	Reviews	Size	Color	Fabric	Style	in general	Clean_Text
0	قماشه خفيييييف ينفع لصيف وطولي ١٥٦ وطلع ع طولي	0	0	1	0	1	قماش خفيف ينفع لصيف طول طلع طول
1	ملائم: روووووووووووووو ووووووووووعة ماشاءالله...	0	0	0	0	1	ملايم رو وعه ماشاءال تبار اله جميل جدا تناس رايع
2	😭 😭 القصه والفستان واللون حلو بس القماش مره لا	0	1	-1	1	-1	قصه فست والو حلو قماش مره لا
3	يابنات صدق خذوووه قسم بالله يحلليييي ويخليييك ...	0	0	0	0	1	يابن صدق خذو قسم بال يحل يخل احل
4	يجنننننن روح مره فخم	0	0	0	0	1	يجن روح مره فخم
...	...	...	...	...	...	...	...
65	اللون حلو بس يفشل في اللبس القصة المثلثله تطلع...	0	1	0	-1	-1	الون حلو يفشل البس قصه مثلثله تطلع عريض حرك ال...
66	اللون جميل وهادئ لكن ما عجب بنتي تصميم الصدر	0	1	0	-1	1	الون جميل هادء ما عجب بنت تصميم صدر
67	ما عجبني ابدا و رح ارجعه موديله غريب تحسوه ناق...	0	0	0	-1	-1	ما عجب ابد رح ارجع موديل غريب تحس ناقص شيء سحر...
68	جدا سيء الصراحة كنت اكثر شي متوقعته يعجبني قصة...	0	-1	-1	-1	-1	جدا سيء صراحه كنت اكثر شي متوقع يعجب قصه صدر م...
69	💜 💜 يهبل صدق يحلي الجسم	0	0	0	1	1	يهبل صدق يحل جسم

Fig. 3. Clothes dataset before and after cleaning.

C. *Feature Extraction*

Feature extraction converts raw text into numerical vectors, which are crucial for machine learning models in aspect-based sentiment analysis. We implemented Part-of-Speech (POS) tagging using the NLTK library to assign grammatical categories to words, aiding in contextual analysis [18]. The POS tagging addresses the complexity of Arabic morphology and contextual forms by understanding the grammatical structure of sentences and identifying sentiment-carrying parts of speech such as adjectives and adverbs. N-gram models, which group contiguous words, capture the sequence of terms, providing features like unigrams and bigrams [19]. These models handle dialectal variation and morphological richness by recognizing and interpreting various expressions and phrases that convey sentiment. Additionally, Term Frequency-Inverse Document Frequency (TF-IDF) assesses term importance in the corpus, transforming words into numerical values for better model input [20]. The TF-IDF technique highlights the importance of words within a document relative to their frequency across a corpus, identifying key sentiment-bearing terms while reducing the influence of common, less informative words. These combined techniques—TF-IDF, POS tagging, and unigrams and bigrams—significantly contribute to overcoming the linguistic and morphological challenges of Arabic sentiment analysis, resulting in a more robust and accurate model.

It is crucial to emphasize that establishing a strong foundational model is essential for advancing research and practical applications in the field of sentiment analysis. Our primary goal was to create a robust yet straightforward model that could be easily understood and implemented by researchers and practitioners in the field. Traditional machine learning techniques and simpler models provide more transparency and interpretability compared to the complex architectures of Word2Vec and BERT. Additionally, this study aimed to establish a baseline using traditional methods and ensure they were thoroughly evaluated before moving on to more complex models. Future work can build upon this baseline by incorporating Word2Vec, BERT, or other advanced techniques to further enhance performance.

D. *Algorithm Selection*

For the analysis, we chose algorithms that, based on our comprehensive research, have shown potential in handling the linguistic and structural complexities of the Arabic language, which can differ from those encountered in English text analysis. Our aim was to experiment with various algorithms and NLP techniques to identify those that would perform aspect-based sentiment analysis on Arabic texts with high accuracy. The selection of algorithms was guided by the results of our literature review (LR), which highlighted the top-performing algorithms in existing research. Consequently, we selected the top four algorithms from the LR results and tested them with our new benchmark datasets. In the following subsections, we provide a detailed description of each algorithm, outlining the rationale for their selection and why they are well-suited for our research.

1) Support Vector Machine: Support Vector Machine (SVM) is a popular supervised learning algorithm used for both classification and regression. It categorizes data points by mapping them to a high-dimensional feature space and aims to find a hyperplane that serves as the optimal decision boundary, maximizing the margin between classes. The data points closest to this boundary are known as support vectors, which play a crucial role in defining the hyperplane's position [21].

The choice of the kernel function is critical for the SVM algorithm's performance. The kernel is a mathematical function used to transform data for finding the hyperplane. It is beneficial to experiment with different kernel functions—namely, linear, polynomial, radial basis function (RBF), and sigmoid—to determine the best model for each case, as each function involves different algorithms and parameters.

2) Random Forest: The Random Forest algorithm is a popular machine learning technique for classification and regression tasks. It is an ensemble learning method that combines the predictions of multiple decision trees to enhance accuracy. In a random forest, an ensemble of decision trees is created, where each tree is trained on a random subset of the training data and considers only a subset of the features at each split. During prediction, each tree provides its predictions independently, and the final prediction is determined by a majority vote (for classification) or an average (for regression) of all individual tree predictions [22]. We chose the Random Forest algorithm because it is particularly effective with high-dimensional datasets, which is important for our project due to the multiple aspects we consider within each category.

3) Naïve Bayes: The Naïve Bayes classifier is a supervised learning algorithm that belongs to the generative learning algorithms group. It is a probabilistic classifier based on Bayes' Theorem [23], making probability assignments based on prior knowledge of conditions related to the predicted event. The "Naïve" aspect of the algorithm refers to its assumption that features are independent, meaning the presence of one feature does not affect the presence of another. This assumption, while simplifying computation and allowing for fast predictions, may not reflect real-world complexities. Despite its simplicity, Naïve Bayes is effective for text classification problems, commonly used in Natural Language Processing (NLP). It calculates the probability of each tag for a given piece of text and selects the tag with the highest probability. This makes it well-suited for categorizing sentiments related to specific aspects within reviews.

4) K-nearest Neighbors: The K-nearest neighbors (KNN) algorithm is a technique for supervised machine learning applicable to both classification and regression problems. It operates on the principle that similar objects or instances are located near each other. To be effective, the KNN algorithm relies on the assumption that the notion of similarity holds true. The algorithm uses mathematical principles, such as calculating distances between points on a graph, to measure similarity. The Euclidean distance, or straight-line distance, is a popular method for this measurement. Selecting the appropriate K value for KNN requires an iterative process to minimize errors and ensure accurate predictions with new data. A higher K value is often beneficial for data with outliers or significant noise. It is advisable to choose an odd number for K to avoid classification ties [24].

In NLP and sentiment analysis, KNN is a straightforward yet powerful tool. Given that textual data often exists in high-dimensional space due to a large vocabulary, measuring distances between text vectors can help classify sentiments or topics. Texts with similar word patterns or sentiments are closer in this space. For sentiment analysis, if the majority of the 'k' nearest text instances have a positive sentiment, the new text is likely classified as positive, and vice versa. However, due to high-dimensionality, considerations such as dimensionality reduction and careful feature extraction are crucial for effective KNN performance in NLP tasks [25].

In addition to the insights gained from the literature review, we performed a 10-fold cross-validation to ensure the reliability of our experimental findings and to avoid overfitting. Cross-validation is a technique that involves dividing the dataset into distinct parts or folds, using each fold once as a validation set while the rest serve as the training set. This thorough evaluation method enables us to assess the model's performance across various subsets of the data, providing a more accurate estimate of its ability to generalize. By using 10-fold cross-validation, we aimed to confirm that our models not only fit the training data well but also perform effectively on unseen data. This rigorous validation approach enhances the reliability of our results and highlights the robustness of our conclusions. The results of the cross-validation are detailed in Table I, providing a clear overview of the model performances across different folds.

TABLE I. SUMMARY OF 10-FOLD CROSS VALIDATION

Combination	Algorithm	Electronics Dataset	Clothes Dataset
TF-IDF	SVM	83.03%	84.92%
TF-IDF+POS	SVM	82.96%	85.03%
TF-IDF+Bigrams	SVM	81.71%	84.04%
TF-IDF+Trigrams	SVM	80.50%	82.91%
TF-IDF + POS + Bi-grams	SVM	81.79%	83.81%
TF-IDF + POS + Tri-grams	SVM	80.58%	82.79%
TF-IDF	RF	81.86%	84.82%
TF-IDF+POS	RF	82.13%	83.48%
TF-IDF+Bi-grams	RF	82.36%	85.02%
TF-IDF + Tri-grams	RF	81.82%	84.79%
TF-IDF + POS + Bi-grams	RF	81.99%	83.61%
TF-IDF + POS + Tri-grams	RF	80.91%	83.25%
TF-IDF	NB	74.69%	72.90%
TF-IDF+POS	NB	74.69%	71.11%
TF-IDF+Bi-grams	NB	75.04%	73.29%
TF-IDF + Tri-grams	NB	74.55%	73.01%
TF-IDF + POS + Bi-grams	NB	74.08%	71.41%
TF-IDF + POS+ Tri-grams	NB	73.74%	71.35%
TF-IDF	KNN	76.81%	73.11%
TF-IDF+POS	KNN	74.37%	71.79%
TF-IDF+Bi-grams	KNN	68.92%	68.94%
TF-IDF + Tri-grams	KNN	53.91%	68.60%
TF-IDF + POS + Bi-grams	KNN	69.49%	71.43%
TF-IDF + POS + Tri-grams	KNN	57.10%	71.61%

E. *Training Phase*

We used a linear kernel for SVM after testing both linear and polynomial kernels to ensure optimal performance. The models were trained on two datasets: electronic product reviews and clothing product reviews. We experimented with six feature extraction combinations, including TF-IDF, Part-of-Speech tagging, and bi-grams and tri-grams, to determine the best approach for enhancing model accuracy.

F. *Implementation of Aspect-Based Sentiment Analysis*

We deployed the models with the highest accuracy. For electronics, the Random Forest model predicted overall sentiment and four specific aspects: quality, usage, price, and size, leveraging its multi-output capability. For clothing, five SVM models were deployed to predict overall sentiment and four aspects: fabric, color, style, and size, since SVM is a single-output classifier. The models and TF-IDF vectorizer were serialized using Python's pickle library to enable predictions on new data without retraining. Applied to a preprocessed dataset, the models predicted product aspects with high accuracy, as detailed in the Experiments section.

All preprocessing and feature extraction combinations are summarized in Table II, Fig. 4, and Fig. 5.

TABLE II. SUMMARY OF PREPROCESSING AND FEATURE EXTRACTION COMBINATIONS

Combination	Algorithm	Electronics Dataset	Clothes Dataset
TF-IDF	SVM	70.9375%	85.8075%
TF-IDF+POS	SVM	83.25%	85.3975%
TF-IDF+Bigrams	SVM	71.0625%	77.685%
TF-IDF+Trigrams	SVM	81.0625%	69.45%
TF-IDF + POS + Bi-grams	SVM	84.0625%	86.5075%

TF-IDF + POS + Tri-grams	SVM	83.625%	86.39%
TF-IDF	RF	82.3791%	85.1075%
TF-IDF+POS	RF	82.1246%	83.4112%
TF-IDF+Bi-grams	RF	83.2675%	84.3457%
TF-IDF + Tri-grams	RF	82.6335%	84.1705%
TF-IDF + POS + Bi-grams	RF	72.9643%	84.1705%
TF-IDF + POS + Tri-grams	RF	86.2595%	82.5934%
TF-IDF	NB	84.3125%	82.651%
TF-IDF+POS	NB	83.1875%	82.418%
TF-IDF+Bi-grams	NB	83.875%	82.9425%
TF-IDF + Tri-grams	NB	83.125%	82.359%
TF-IDF + POS + Bi-grams	NB	82.9375%	82.593%
TF-IDF + POS+ Tri-grams	NB	83.062%	82.593%
TF-IDF	KNN	77.5575%	74.4775%
TF-IDF+POS	KNN	76.405%	73.715%
TF-IDF+Bi-grams	KNN	64.64%	72.255%
TF-IDF + Tri-grams	KNN	71.8025%	67.875%
TF-IDF + POS + Bi-grams	KNN	76.4075%	74.53%
TF-IDF + POS + Tri-grams	KNN	64.5775%	71.67%

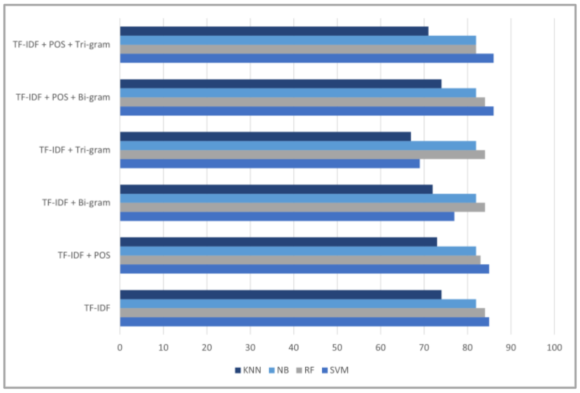


Fig. 4. Preprocessing and feature extraction combinations for clothes dataset.

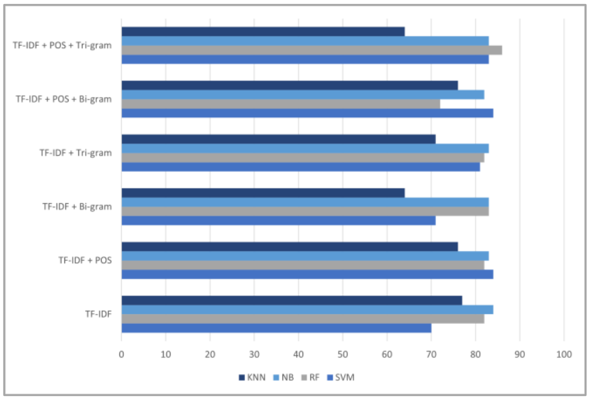


Fig. 5. Preprocessing and feature extraction combinations for electronic dataset.

IV. Experiments

In this experiment, the primary objective is to establish a benchmark for the main contribution of this work, which is the new datasets in the Saudi dialect. It is important to emphasize that while this study provides a foundational benchmark, future research can conduct extensive hyperparameter tuning or explore different approaches to enhance the performance of the algorithms with these datasets.

A. *Models Evaluation*

Researchers evaluated the trained models and reviewed each algorithm with the best feature extraction combinations for each dataset. Accuracy was used as the performance metric. Six combinations of feature extraction techniques were applied: TF-IDF alone, TF-IDF with POS, with bi-grams, with tri-grams, TF-IDF with POS and bi-grams, and TF-IDF with POS and tri-grams. Table II shows that for electronics, Random Forest using TF-IDF, POS, and tri-grams achieved the highest accuracy of 86.2595%. For clothes, SVM with TF-IDF, POS, and bi-grams scored the highest accuracy at 86.5075%. SVM achieved 84.0625% for electronics and 86.5075% for clothes. Random Forest had an accuracy of 86.2595% for electronics and 85.1075% for clothes. Naïve Bayes scored 84.3125% for electronics and 82.9425% for clothes. KNN registered 78.43% for electronics and 75.87% for clothes. These results highlight the importance of feature extraction combinations and domain-specific challenges in sentiment analysis.

B. *SVM Performance*

Table III shows the performance of the SVM classifier for aspect-based sentiment analysis on Electronics and Clothes datasets. For electronics reviews, SVM using TF-IDF, POS, and bi-grams showed varied performance. Quality aspect had moderate results with 67.50% accuracy and recall, and a 66.97% F1-score. Usage aspect improved with 78.75% accuracy and recall, and a 76.70% F1-score. Size and Price aspects excelled with over 95% accuracy and recall, and F1-scores of 94.72% and 94.31%, respectively. Average metrics for the SVM electronics model were 84.0625% accuracy and recall, and 83.175% F1-score.

For clothes reviews, using the same extraction techniques, SVM performed consistently well. Style aspect achieved 84.35% accuracy and recall, and an 82.87% F1-score. Fabric aspect followed with 84.11% accuracy and recall, and an 84.04% F1-score. Size aspect showed 87.62% accuracy and recall, and an 87.07% F1-score, while Color aspect excelled with 89.95% accuracy and recall, and an 89.32% F1-score. The average metrics were 86.5075% accuracy and recall, and an 85.825% F1-score.

TABLE III. The Best Features Combination for each Dataset for SVM

Dataset	Combination	Accuracy	F1-score	Recall
Electronics	TF-IDF + POS + Bi-grams	84.0625%	83.175%	84.0625%
Clothes	TF-IDF + POS + Bi-grams	86.5075%	85.825%	86.5075%

C. RF Performance

Table IV illustrates the performance of the RF classifier on Electronics and Clothes datasets. In electronics reviews, RF using TF-IDF, POS, and tri-grams excelled: Quality (60.55% accuracy and recall, 58.70% F1-score), Usage (94.91% accuracy and recall, 92.11% F1-score), Size (94.65% accuracy and recall, 93.76% F1-score), and Price (94.75% accuracy and recall, 93.34% F1-score). Average metrics were 86.2595% accuracy and recall, and 84.4775% F1-score. For clothes reviews, RF performed well: Style (83.18% accuracy and recall, 80.17% F1-score), Fabric (82.01% accuracy and recall, 81.82% F1-score), Size (87.62% accuracy and recall, 87.27% F1-score), and Color (87.62% accuracy and recall, 86.21% F1-score), with average metrics of 85.1075% accuracy and recall, and 83.8675% F1-score.

TABLE IV. THE BEST FEATURES COMBINATION FOR EACH DATASET FOR RF

Dataset	Combination	Accuracy	F1-score	Recall
Electronics	TF-IDF + POS + Tri-grams	86.2595%	84.4775%	86.2595%
Clothes	TF-IDF	85.1075%	83.8675%	85.1075%

D. NB Performance

The performance of the NB classifier on Electronics and Clothes datasets is illustrated in Table V. In electronics reviews, NB using TF-IDF showed: Quality (74.00% accuracy and recall, 68.19% F1-score), Usage (80.75% accuracy and recall, 72.85% F1-score), Size (95.50% accuracy and recall, 93.30% F1-score), and Price (87.00% accuracy and recall, 81.42% F1-score). Average metrics were 84.3125% accuracy and recall, and 78.94% F1-score. For clothes reviews, NB with TF-IDF and bi-grams performed: Style (81.31% accuracy and recall, 72.93% F1-score), Fabric (73.13% accuracy and recall, 64.17% F1-score), Size (89.95% accuracy and recall, 85.20% F1-score), and Color (87.38% accuracy and recall, 81.50% F1-score). Average metrics were 82.9425% accuracy and recall, and 75.95% F1-score.

TABLE V. THE BEST FEATURES COMBINATION FOR EACH DATASET FOR NB

Dataset	Combination	Accuracy	F1-score	Recall
Electronics	TF-IDF	84.3125%	78.94%	84.3125%
Clothes	TF-IDF + Bi grams	82.9425%	75.95%	82.9425%

E. KNN Performance

Table VI shows the performance of the KNN classifier for aspect-based sentiment analysis on Electronics and Clothes datasets, using accuracy, F1-score, and recall as metrics. The classifier achieved reasonable accuracy: 77.56% for Electronics with TF-IDF and 74.53% for Clothes with POS + bi-gram + TF-IDF. However, lower F1-scores and recall rates suggest an imbalance between precision and recall, indicating many false negatives. Optimization may require retraining with different parameters, alternative feature sets, or exploring other classification algorithms.

TABLE VI. THE BEST FEATURES COMBINATION FOR EACH DATASET FOR KNN

Dataset	Combination	Accuracy	F1-score	Recall
Electronics	TF-IDF	77.56%	50.25%	49.83%
Clothes	POS+Bi-gram + TF-IDF	74.53%	55.66%	51.33%

V. DISCUSSION AND FUTURE WORK

As discussed earlier in the previous section, the models were implemented using four machine learning algorithms, three feature extraction techniques were applied with six different combinations, and one metric was used for performance measurement. To maintain a reliable performance for the models, two categories of products were analyzed: electronics and clothes. Despite this selection being devoted to a certain scope of data, it implies that there is potential to expand the research scope to other product categories, e.g., food, offering a broader understanding of sentiment patterns across various domains. While the evaluation process of the models is meticulous, it is not without its limitations. Further enhancements in model accuracy could potentially be achieved if additional feature extraction techniques were explored. Future research endeavors should consider exploring different algorithms and additional sophisticated techniques, and consider the exploration of other performance metrics, aiming to develop a more subtle and holistic understanding of model efficacy in aspect-based sentiment analysis.

VI. CONCLUSION

In an era dominated by data, effectively analyzing customer reviews is essential for product success. Our research focused on product reviews in Saudi Arabian dialects, a relatively less explored area in sentiment analysis. We performed aspect-based sentiment analysis on online product reviews utilizing machine learning algorithms with NLP techniques and provided a thorough experimentation to highlight their effectiveness. We collected and preprocessed two datasets from the clothing and electronics sectors, each with 2000 reviews, implementing techniques like tokenization, stemming, and normalization to prepare the data for training the model for the analysis. Our experiments tested four machine learning algorithms (SVM, RF, NB, KNN) across six combinations of feature extraction methods (TF-IDF, POS tagging, n-grams), finding the RF algorithm with TF-IDF, POS tagging, and tri-grams combination the most effective for electronics reviews at an average accuracy of 86.2595% per aspect, and the SVM for clothing reviews using TF-IDF, POS tagging, and bi-grams at 86.5075% average accuracy for every aspect.

REFERENCES

[1] “Leadership Principles.” 2023, [Online]. Available: https://www.amazon.jobs/content/en/our-workplace/leadership-principles.

[2] A. A. Shad, “Feedback Analysis: How To Analyze Customer Feedback?” May 2023, [Online]. Available: https://userpilot.com/blog/feedback-analysis/.

[3] J. Mir, M. Azhar, and S. Khatoon, “Aspect based classification model for social reviews,” Engineering, Technology and Applied Science Research, vol. 7, pp. 2296–2302, June 2017, https://doi.org/10.48084/etasr.1578.

[4] M. Hadwan, M. Al-Hagery, M. Al-Sarem, and F. Saeed, “Arabic sentiment analysis of users’ opinions of governmental mobile applications,” Computers, Materials, and Continua, vol.72, no.3, pp. 4675-4689, 2022.

[5] A. Bahamdain, Z. H. Alharbi, M. M. Alhammad, and T. Alqurashi, “Analysis of logistics service quality and customer satisfaction during Covid-19 pandemic in Saudi Arabia,” International Journal of Advanced Computer Science and Applications, , vol. 13, no. 1, pp. 174-180, 2022.

[6] S. Al-Hagree and G. Al-Gaphari, “Arabic sentiment analysis based machine learning for measuring user satisfaction with banking services’ mobile applications: comparative study,” in 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), 2022.

[7] N. Almuhanna and Z. H. Alharbi, “Factors affecting customer satisfaction with the telecommunication industry in Saudi Arabia,” TEM Journal, May 2023.

[8] J. Holdsworth, “What is Natural Language Processing? | IBM.” [Online]. Available: https://www.ibm.com/sa-en/topics/natural-language-processing.

[9] “2. Dealing with Linguistic Variation.” 2023, [Online]. Available: http://www.aviarampatzis.com/Avi_Arampatzis/publications/HTMLized/encyclop/node2.html.

[10] UNESCO, “World Arabic Language Day”, December 2023, [Online]. Available: https://www.unesco.org/en/world-arabic-language-day

[11] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. Monem, “Challenges in Arabic natural language processing,” In Computational linguistics, speech and image processing for arabic language, pp. 59–83. 2019.

[12] R. Badawi, “Data Annotation for Arabic NLP- Deceptively Easy - Globitel.” May 2021, [Online]. Available: https://www.globitel.com/data-annotation-for-arabic-nlp-deceptively-easy/.

[13] R. Duwairi and M. El-Orfali, “A study of the effects of preprocessing strategies on sentiment analysis for Arabic text,” Journal of Information Science, vol. 40, no. 4, pp. 501–513, 2014.

[14] V. L. S. Lee, K. H. Gan, T. P. Tan, and R. Abdullah, “Semi-supervised learning for sentiment classification using small number of labeled data,” Procedia Computer Science, vol. 161, pp. 577–584, 2019.

[15] J. Holdsworth, and M. Scapicchio, “What is Deep Learning?” 2015, [Online]. Available: https://www.ibm.com/topics/deep-learning.

[16] G. Belani, “6 Interesting Deep Learning Applications for NLP.” Paperspace Blog, May 2019, [Online]. Available: https://blog.paperspace.com/6-interesting-deep-learning-applications-for-nlp.

[17] M. Alduraibi, R. Alrefaey, R. Alqahmi, S. Almatrafi, and A. Alayed, “SaudiShopInsights Dataset: Saudi Customer Reviews in Clothes and Electronics.” IEEE Dataport, 2023, https://doi.org/10.21227/6e56-4e15.

[18] J. Daniel and J. Martin, “Speech and Language Processing Part-of-Speech Tagging.” Stanford University, May 2019, [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/old_oct19/8.pdf.

[19] N. V, “What Are N-Grams and How to Implement them in Python?” May 2021, [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/what-are-n-grams-and-how-to-implement-them-in-python/.

[20] F. Karabiber, “TF-IDF — Term Frequency-Inverse Document Frequency – LearnDataSci.” 2023, [Online]. Available: https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/#:~:text=Term.

[21] I, Javaid, “How to use SVM for sentiment analysis.” [Online]. Available: https://www.educative.io/answers/how-to-use-svm-for-sentiment-analysis.

[22] N. Donges, “A Complete Guide to the Random Forest Algorithm.” May 2021, [Online]. Available: https://builtin.com/data-science/random-forest-algorithm.

[23] “What are Naive Bayes classifiers? | IBM.” 2023, [Online]. Available: https://www.ibm.com/topics/naive-bayes#:~:text=The%20Na%C3%AFve%20Bayes%20classifier%20is,a%20given%20class%20or%20category.

[24] IBM, “What is the k-nearest neighbors algorithm? | IBM,” 2023, [Online]. Available: https://www.ibm.com/topics/knn.

[25] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for text classification," *Augmented Human Research*, vol. 15, no. 1, pp. 12-24, Mar. 2020.