



Sentiment analysis of reviews written in the Saudi dialect to
classify customer satisfaction levels in e-commerce.

Prepared By

Areej Haddaj Alqarni

252001729

Submitted in fulfillment of the requirements for the degree
Master of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor: Dr. Yasser Abdul Hamid

College of Informatics, Midocean University, Comoros

[2026- Third Trimester]

Abstract

Sentiment analysis is crucial for all businesses and has become an effective method for researching customer satisfaction in e-commerce, given the widespread presence of e-commerce on electronic platforms for purchase decision-making purposes. To address this challenge, in this study, we have developed a sentiment analysis model of Customer Reviews using Arabic, primarily covering the Saudi dialect, to help companies effectively assess the quality of products and services. The second part involves a literature review of Arabic sentiment analysis and six academic papers that have added significant rigor to the theoretical frameworks supporting our study. A dataset of customer reviews in the Saudi dialect corpus was collected and compiled using text preprocessing (cleaning the data and normalizing, filtering, and extracting their features). Then, a sentiment analysis model was developed with different machine learning algorithms, including various classifiers and Aspect-Based Sentiment Analysis (ABSA), that can determine customer satisfaction from positive, negative, and neutral perspectives. In this work, we proposed a model for Arabic text analysis in the Saudi dialect and suggested a robust methodology that allows our language to effectively handle the inclusion of some informal texts. It also offers an annotated dataset available for consideration in future studies. The findings can provide more insight into customer satisfaction trends for Saudi companies in their market, thereby facilitating information-driven decision-making for e-businesses.

Table of Contents

Abstract	1
1 Introduction	6
1.1 Problem Definition.....	6
1.2 Thesis Objectives	6
1.3 Thesis Contribution.....	7
1.4 Thesis Organization	7
2 Background.....	8
2.1 The Role of Sentiment Analysis in E-Commerce	8
2.2 Sentiment analysis in the Arabic language	8
2.3 Machine Learning Approaches	9
2.4 Pre-trained Language Models	9
2.5 Datasets in Arabic Sentiment Analysis	9
3 Related work	10
4 Literature Review Matrix	12
5 Research Gaps	13
6 summary.....	14
7 Proposed Approach	15
7.1 Data collection and Description.....	15
7.2 Dataset Refinement.....	16
7.3 Data preprocessing.....	17
7.4 Input Representation and Tokenization:	18
7.5 Training setup:	19
7.6 Model training:.....	19
7.7 Performance evaluation Metrics:	19
8 Experimental Results and Discussion	20
8.1 Experimental Setup:.....	20
8.2 Feature Relationship Analysis:	20

8.3	Overall Model Performance Evaluation:	21
8.4	Comparative Performance Analysis SaudiBERT vs. CAMELBERT-DA:	22
8.5	Performance Evaluation per Aspect:	23
8.6	illustrative case study:	24
8.7	Discussion:	25
9	Conclusion and Future Work	26
9.1	Conclusion	26
9.2	Future work	26
	References.....	28

List of Abbreviations

- NLP: Natural Language Processing
- SA: Sentiment Analysis
- SVM: Support Vector Machine
- NB: Naive Bayes
- KNN: K-Nearest Neighbors
- DT: Decision Tree
- RF: Random Forest
- ABSA: Aspect-Based Sentiment Analysis
- TF-IDF: Term Frequency – Inverse Document Frequency
- BOW: bag of words
- POS: Part-of-Speech Tagging
- BERT : Bidirectional Encoder Representations from Transformers

List of Figures

FIGURE 1:A SIMPLE FROM CLOTHES DATASET	15
FIGURE 2:DISTRIBUTION OF SENTIMENT IN THE ORIGINAL DATASET	16
FIGURE 3: AFTER MERGING AND ENHANCEMENT	16
FIGURE 4:A SAMPLE FROM DATASET AFTER PREPROCESSING	18
FIGURE 5:CORRELATION BETWEEN ASPECTS	20
FIGURE 6:TRANING AND VALIDATION LOSS CURVE	21
FIGURE 7:OVRALL CONFUSION MATRIX.....	22
FIGURE 8:MODEL EVALUATION FOR EACH ASPECT	24

Chapter 1

1 Introduction

1.1 Problem Definition

E-commerce has experienced significant growth in recent years, and dependence on digital platforms is an integral part of customers' decision-making processes. Nowadays, customer reviews and textual evaluations are regarded as the main sources of understanding user experiences with products and services. The quantity of consumer content created by users on e-commerce sites has made the challenge for organizations to deal with the large and continuously growing volume of data generated today an extremely serious one compared to earlier days when data was analyzed by traditional methods or through manual analysis, making data processing no longer possible. This highlights the need to create more sophisticated methods based on artificial intelligence technologies for customer sentiment analysis and the accurate determination of customers' sentiments so that enterprises can modify the quality of products offered, improve the quality of services, and help with future decision-making. Various Arabic studies have tackled the subject of sentiment analysis, but the proposed approaches still need more development in terms of data, accuracy, and the variety of data sources. In addition, there is still a clear need to contribute better natural language processing methods that can process dialectal or linguistics-specific information, analyze textual content, and make the analysis more accurate and effective.

1.2 Thesis Objectives

This thesis presents the following aims:

1. To collect and prepare a dataset of customer reviews written in the Saudi dialect from e-commerce platforms.
2. To apply Natural Language Processing techniques to preprocess, clean, and analyze informal text written in Saudi dialect.

3. To develop a sentiment analysis model, including Aspect-Based Sentiment Analysis (ABSA), capable of classifying customer satisfaction levels (positive, negative, and neutral).
4. To evaluate and compare the performance of various machine learning and deep learning algorithms in sentiment classification.
5. To identify the most influential linguistic features that contribute to determining sentiment in Saudi dialect customer reviews

1.3 Thesis Contribution

This thesis also makes several scientific contributions, including:

1. A sentiment analysis model tailored for customer reviews, the Saudi dialect (a relatively new area of research).
2. Defining clear processes for processing informal Arabic text in the e-commerce context (e.g., preprocessing, analysis and feature extraction).
3. An annotated dataset, which would be useful to conduct future research on sentiment analysis in Arabic.
4. Making strides in the study of customer satisfaction in the Saudi market, data-driven decision-making for e-commerce platforms.

1.4 Thesis Organization

The thesis is organized into a number of chapters as follows:

- Chapter 1: Introduction introduces the research topic, objectives, research problem, and study significance.
- Chapter 2: Background establishes the theory behind the study and literature review presents previous studies and compares them to one another.
- Chapter 3: Proposed Approach describes the research design, data collection methods, preprocessing steps, and the machine learning techniques involved.
- Chapter 4: Experimental Results and Discussion we reported the results of experimental results and an evaluation of model performance.
- Chapter 5: Conclusions and Future Work give an indication of the major findings and their contributions and shortcomings with recommendations for future research.

Chapter 2

2 Background

2.1 The Role of Sentiment Analysis in E-Commerce

E-commerce is at a period of phenomenal development in the past several years and has become a critical component in consumers' purchasing behavior and is now making a significant contribution towards an overall change in the purchasing practices of buyers. This growth has led organizations to understand customers through the digital channel and by reading opinions and reviews on every digital medium, including every website and social network. Today's market cannot be limited to the traditional feedback methods alone, markets must have smart tools to understand customer digital impressions in order to improve service quality and lead to data-driven decisions and improved levels of customer satisfaction. Thus, the need to develop digital tools in which companies can find out what consumers really think of their products, especially since there is this huge volume of reviews that cannot be done manually, can only give rise to an impression of a good situation. Sentiment analysis provides a good insight of customer beliefs by analyzing the reviews of customers in text with text content processing and showing whether they are positive/negative/neutral.

2.2 Sentiment analysis in the Arabic language

In Arabic, sentiment analysis plays a crucial role, especially in analyzing customer reviews and social media. Language analysts struggle with dialectal variation, morphological complexity, and ambiguity in informal or abbreviated texts. In the face of these challenges, numerous studies have emphasized the importance of constructing corpora and lexicons for various dialects such as Jordanian, Egyptian, and Saudi Arabic in order to create more robust and efficient models of sentiment analysis in Arabic. A blend of traditional and modern algorithms is being used to analyse types of text and to classify sentiment, each with a characteristic value and operation approach. Among the most widely used algorithms in previous studies are the following:

2.3 Machine Learning Approaches

Much work has been done on Arabic sentiment analysis using machine learning techniques, and the studies mainly focus on specific stages starting from text preprocessing to ensure the quality of data before the classification. This part consists of noise removal, text normalization, and dealing with the many language features in the Arabic language. Then, feature extraction techniques were used to make numerical representation of textual data, allowing machine learning algorithms to process this effectively. [1] used six different feature extraction combinations to discover the most robust way to improve model accuracy. Several machine learning classifiers were evaluated, including SVM, NB, RF, and KNN. The main result of this study is that the SVM model combined with TF-IDF, POS tagging, and bigram features attained the greatest rate of performance among the models evaluated. Additionally, the research conducted by [9] proved that the performance of machine learning algorithms was improved significantly when they convert local dialects texts into Modern Standard Arabic (MSA). Specifically, the SVM classifier achieved the best performance after this conversion, indicating the significance of linguistic normalization in Arabic sentiment analysis. [12] proved that the SVM algorithm shows strong performance very close to the deep learning model CAMELBERT, which is further proof indicating that the SVM algorithm can perform very fast and effectively in Arabic sentiment analysis tasks.

2.4 Pre-trained Language Models

These are artificial intelligence models that were first trained with a large set of texts to learn the linguistic structures, context, and relationships of words. They can later be fine-tuned for particular tasks, such as text classification, sentiment analysis, question answering, or information extraction, without need to train them from scratch: freeing up time and resources and increasing accuracy. There are clear gaps in previous investigations since the amounts of training data used to fit models are small, and many were collected by hand from various platforms. This lack of sufficient data directly affected the models' performance, and it might lead to some inaccuracies for the models' evaluation results due to the limited dataset.

2.5 Datasets in Arabic Sentiment Analysis

In previous studies on sentiment analysis, datasets tend to be sourced from websites like social media or review websites such as Twitter or e-commerce platforms. The

latter dataset type includes user-generated texts in Modern Standard Arabic or various Arabic dialects. Texts are typically annotated into positive, negative, neutral, and otherwise, using preprocessing such as the removal of noise, normalization, and stemming to give high quality data and to improve the processing performance. However, compiling high-quality Arabic sentiment data is still difficult due to the range of dialects, the rise of informal language practices, and a shortage of publicly annotated resources. In fact, in the study [1], researchers developed aspect-based sentiment analysis (ABSA) compared to conventional full-text sentiment analysis. For this, a carefully annotated dataset on aspects can be needed which can include factors such as price, quality, service, and so forth, to correctly identify sentiment of the other aspects. Our results showed that ABSA delivers a much more detailed and in-depth analysis than general sentiment analysis, offering actionable and accurate insights into customer opinions and a richer grasp of user experience

3 Related work

In study [11], the researchers built a dataset of Arabic tweets in the Jordanian dialect. The data were collected using keywords related to general Jordanian topics. Data was preprocessed: data cleaning, removal of URLs, unwanted characters, repeated words, and hashtags. The tweets were subsequently annotated manually and categorized as positive, negative, or neutral. The dataset was trained with conventional machine learning algorithms like Support Vector Machine (SVM) and Naïve Bayes. TF-IDF representation along with stemming and bigram features in the SVM model led to the best accuracy, F1-score, and recall scores, according to the findings.

Similarly, in [9] in particular, focused on the citizens' opinions towards the traffic conditions in three Moroccan cities. Responses were gathered from social media posts on Twitter and Facebook in the Moroccan dialect. The authors used standard machine learning algorithms, and it was observed that the SVM model performed better than other algorithms after translating the Moroccan dialect into Modern Standard Arabic (MSA) with an accuracy of 92%. These findings also indicated a predominance of negative sentiments, particularly in the city of Rabat.

In 2024, study [1] applied Aspect-Based Sentiment Analysis (ABSA) to analyze customer reviews written in the Saudi dialect using Natural Language Processing techniques and machine learning algorithms. Two separate datasets were extracted from e-commerce platforms such as Amazon and Shein, one for the clothing category and the other for electronics. The data were then annotated by recognizing relevant aspects, such as quality and price in the electronics category, as well as size and color in the clothing category, and all values for sentiment levels were set as 1 (positive), -1 (negative), or 0 (neutral). The results of the evaluation showed that the Support Vector Machine (SVM) method outperformed other models with an accuracy of 86.5%, an F1-score of 85.82%, and a recall of 86.50% when using a combination of TF-IDF, POS, and bigram features.

In study[13], the sentiment of comments on the YouTube platform to enable content creators to assess viewers satisfaction with their content. This study addressed the challenge resulting from the removal of the dislike count feature on YouTube by relying on user comments as an alternative indicator of audience perception. The data were collected using a Python-based script connected to the YouTube Application Programming Interface (API), through which 4,760 comments were gathered from various video categories. The sentiments were manually labeled by five annotators into two classes: positive and negative. After removing irrelevant comments, a balanced dataset consisting of 4,212 comments was obtained, equally divided into 2,106 positive and 2,106 negative comments. Following text preprocessing, several machine learning classifiers were evaluated. The results demonstrated that the Naïve Bayes classifier outperformed the other classifiers, achieving an accuracy of 94.62%, a precision of 94.64%, a recall of 94.64%, and an F1-score of 94.62%.

In study [12], The researchers sought to design an opinion-based sentiment system for Arabic language opinions of higher education students in order to enhance and improve the learning process. The paper advanced this SVM-SAA-SCR model which prioritizes text preprocessing prior to classifying opinions under Positive, Negative and Neutral sentiment using SVM (Support Vector Machine). The actual dataset obtained from Prince Sattam bin Abdulaziz University (PSAU) which comprises students perspectives on several educational topics. The performance of the system was assessed by comparing it to the advanced CAMELBERT model. The

difference between the two models was small, which confirmed the robustness and effectiveness of the proposed SVM–SAA–SCR algorithm.

4 Literature Review Matrix

This table provides a summary of the most relevant previous studies related to sentiment analysis in the context of e-commerce, highlighting the methodologies used and the key findings of each study.

Study	Algorithm	Dataset	Key Findings
Alrefae, et al., (2024)	SVM, Random Forest (RF), Naïve Bayes (NB), KNN	Amazon, Jarir, Shein, Noon, and Twitter	1. For the electronics dataset, the RF algorithm worked well with TF-IDF, POS tagging, and trigrams, achieving an average accuracy of 86.26% per aspect. 2. For the clothing dataset, using TF-IDF along with POS tagging and bigrams, an average accuracy of 86.51% per aspect was achieved.
Al-Thubaity, et al.,(2018)	Lexicon-based approach	Collect data in twitter	SauDiSenti performed better for neutral tweets and showed promising results despite being smaller than AraSenTi.
Alomari, et al.,(2017)	SVM Naïve Bayes	collected from the X platform(Twitter).	The results showed that SVM outperformed Naïve Bayes in terms of accuracy and F-score.

Dahbi,et al.,(2019)	Naive Bayes, SVM,KNN, Decision Tree	using a Python script from Twitter and Facebook	SVM algorithm achieved the highest performance after converting the texts from the Moroccan dialect to Modern Standard Arabic (MSA).
Louati, et al., (2023)	SVM	real student opinions from the Department of Computer Science and Information Systems at Prince Sattam University.	The SVM-SAA-SCR model achieved strong performance with an accuracy of 84.7% ,Precision of 86% and Recall 85%.
Musleh, et al.,(2023)	NB, SVM,KNN, DT,RF	YouTube comments	Achieved an accuracy of 94.62%, precision of 94.64%, recall of 94.64%, and F1-score of 94.62%, making it the best-performing classifier among the six evaluated classifiers.

5 Research Gaps

Most previous Arabic sentiment analysis studies have scarce resources and datasets on Arabic language, while limited datasets that do not include all dialects and domains were considered. In addition, the linguistic lexicons and dictionaries are still not well utilized, and have not been integrated with modern techniques, such as Transfer Learning, which limits the model's ability to understand nuanced and figurative expressions and accurately analyze sentiments, delaying the construction of proper tools for Arabic text analysis.

6 summary

The studies highlight the importance of sentiment analysis in business development and in leveraging information extracted from social media texts to enhance and improve business practices. They demonstrate that Arabic text analysis is feasible and effective, whether in Modern Standard Arabic or in various dialects, with the necessity of performing text preprocessing, feature extraction, and considering dialectal differences to improve accuracy. The results also showed the effectiveness of machine learning algorithms, particularly the SVM algorithm, which achieved the best performance and was considered the most efficient scenario among the evaluated classifiers.

7.2 Dataset Refinement

After reviewing the original dataset, it was refined by reducing and merging aspects for reasons aimed at increasing the model's training efficiency, Including addressing the Addressing Data Sparsity. It was observed that the aspects in the original dataset suffered from an imbalance, as the neutral category dominated most of the data, which led to the model's SaudiBert, and the prediction of the neutral category consistently instead of the other categories. To reduce dispersion, detailed features were merged and made into holistic features. For example, transparency was merged with fabric, and sleeve and length were merged with size. This step ensures data condensation in one aspect so that the model can be trained on intensive samples of categories, and also reduces overlap and ambiguity, as detailed aspects lead to model confusion. For example, when testing, the model predicts a length of a dress that is classified as -1 in the length aspect, and a different review classifies the size as -1. Therefore, reducing the aspects contributes to removing overlap and increasing the model's accuracy , Figure (2) shows the distribution of feelings across aspects after merging and refining, and clearly demonstrates the intensification of positive and negative data compared to the original data set in Figure (1).

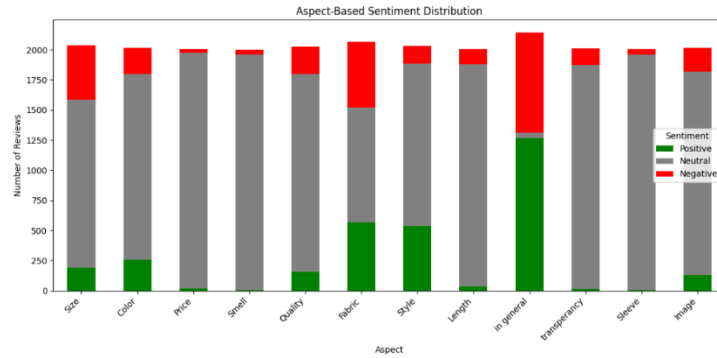


Figure 2: Distribution of sentiment in the original dataset

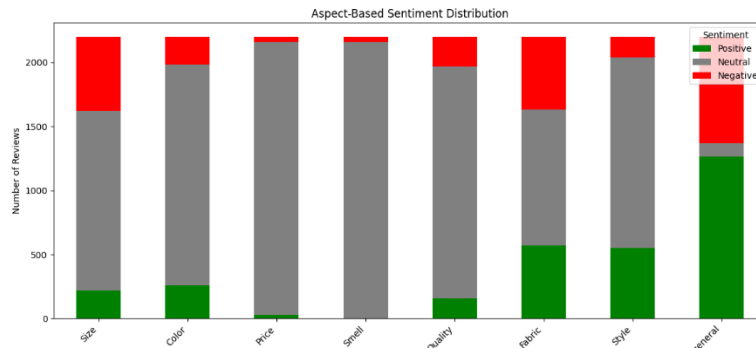


Figure 3: Distributing sentiment across eight aspects after merging and enhancement

7.3 Data preprocessing

The data processing stage is a crucial phase for ensuring the accuracy and quality of the model. Cleaning and formatting processes are applied to reduce noise before passing the data to the model, including:

- **Handling missing value** : after examining the data and detecting the missing values, they were processed. The cleaning process was also applied to Empty Strings/Whitespace, which does not contain meaningful review but contain spaces. They were dealt with using regular expression to convert the spaces into missing values and then processed.
- **Data cleaning** : Delete URLs , emojis, hashtags, special characters and punctuation that does not convey emotional meaning.
- **Arabic text normalization**: To simplify the Arabic language and ensure its consistency, several processes must be applied, namely:
 1. letter normalization : standardizing various forms of specific characters to single uniform shape , this including converting Alif (أ) to (ا) and replacing (ة) to (ه) also (ى) to (ي).
 2. Removal of diacritics (tashkeel): eliminating Arabic vowels such as َ ُ ِ ّ ّ ّ This step is necessary for matching because the model considers it different due to the diacritics.
 3. Removal of Tatweel such as (جميل) to (جميل) .
 4. Removal whitespace normalization: all redundant whitespaces including tabs and multiple consecutive spaces and leading and trailing spaces were removed using strip() function.
 5. Removal of repeated letters in the word such as (مررررره) to (مره).

	Review	Size	Color	Price	Smell	Quality	Fabric	Style	in general	Review_after_clean
0	قماشه خفيف ينقع لصيف وطولي ١٥٦ وطلع ع طولي	1	0	0	0	0	1	0	1	قماشه خفيف ينقع لصيف وطولي ١٥٦ وطلع ع طولي
1	ملاتم: روووووووووووووووووو ... ماشاء الله	0	0	0	0	0	0	0	1	ملاتم روعه ماشاء الله تبارك الله ...جميل جددا فن
2	القصة والفستان واللون حلو بس القماش مره لا 🙄🙄	0	1	0	0	0	-1	1	-1	القصة والفستان واللون حلو بس القماش مره لا
3	يابنات صدق خذوووه قسم بالله يحلليبي ... ويخليبك	0	0	0	0	0	0	0	1	يابنات صدق خذوووه قسم بالله يحلليبي ...ويخليبك احلا
4	يجننننننننننن روح مره فخم	0	0	0	0	0	0	0	1	يجنن روح مره فخم
5	الفساتين حقه شي مره حلوه ولا غلطه انصح ... فيها	0	0	0	0	1	0	0	1	الفساتين حقه شي مره حلوه ولا غلطه ...انصح فيها ب
6	خيال بس طلعت كبيره عليي من عند الصدر ...بييلها وح	0	0	0	0	1	0	1	1	خيال بس طلعت كبيره عليي من عند ...الصدر بييلها وح

Figure 4:A sample from Dataset after preprocessing

- **Label encoding :** This includes the fact that the targets to represent emotions are 1 for positive, -1 for negative, and 0 for neutral. Label encoding was performed to convert the values to 0 for negative, 1 for neutral, and 2 for positive in order to avoid programming errors during training.
- **Data splitting:** The dataset of 2142 reviews was divided into three sets: the training set 70% to train the model and teach it how to link words and different emotions for each side the validation set 15% used during training to adjust hyperparameters and monitor the model to avoid the problem of overfitting; and the test set 15% used to measure accuracy and F1 score.

7.4 Input Representation and Tokenization:

At this stage, the texts were converted from raw format to a digital format that the SaudiBERT model could process. The model's tokenizer was used to perform operations such as tokenization and padding and Truncation to ensure the stability of operations within the arrays. A length of 128 tokens was set for the sentences, and revisions that are shorter than the length are padded with zeros, while those that exceed the length are cut. Also, from the operations Attention masks binary arrays are generated, the benefit of which is to direct the model to focus on the actual tokens and ignore the padding.

7.5 Training setup:

To ensure efficient training, the PyTorch framework was used. A custom data class ABSADataset was built to link text to an 8 aspect label matrix . A Dataloader was used to organize the data into batches of 16 revisions per batch, which helps streamline its processing in the GPU.

7.6 Model training:

At this stage, the model SaudiBERT was adjusted to suit the multi aspect-based sentiment analysis task. The outputs of the final classification layer were modified to include 24 neurons representing 8 aspects *3 classes, as the outputs (logits) were then reshaped to ensure that each aspect was classified separately. The training process was based on calculating joint loss function using cross-Entropy loss for each aspect, then summing them to update the weights based on the total error, with the use of an optimizer (AdmW) with a learning rate of 2e-5 and a Linear Scheduler with Warmup to ensure that the updates were stable and to avoid divergence in the early stages and to avoid overfitting. The process continued for 9 epochs.

7.7 Performance evaluation Metrics:

At this stage, the model SaudiBERT ability to accurately classify emotions for each of the eight aspects is evaluated, using standardized measures that provide a clear indication of the model's performance:

1. **Accuracy:** To measure the overall percentage of correct predictions made by the model.
2. **Precision:** To measure the model's reliability in predicting a specific category and to ensure there aren't too many incorrect classifications.
3. **Recall:** To measure the model's ability to detect all actual cases belonging to a specific category and to ensure that no important review is missed.
4. **F1-Score:** This is the most important metric because it represents the harmonious average between (precision) and (recall). It was used to confirm the strong performance of the model even in the event of an imbalance between the number of positive, negative, and neutral reviews.

Chapter 4

8 Experimental Results and Discussion

This chapter reviews the results obtained after training and evaluating the model, focusing on its ability to classify emotions across several aspects of the Saudi dialect.

8.1 Experimental Setup:

The experiment was carried out on the Kaggle platform, relying on a GPU (GPU P100) to accelerate the training of the model SaudiBERT . Python was used as the main development environment, relying on the Pytorch library to build the model structure and the Hugging Face Transformers library to call the pre-trained weights and the model's tokenizer.

8.2 Feature Relationship Analysis:

To understand the relationship between different aspects of the data set, a correlation matrix was created. Figure 4 illustrates the correlation between variables, showing that most aspects exhibit weak to moderate correlations, indicating that each aspect contributes relatively independent information, this is suitable for Aspect based sentiment analysis.

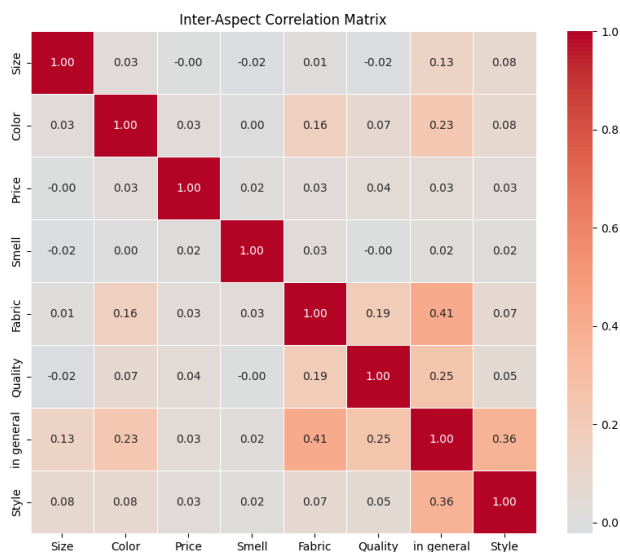


Figure 5: Correlation between Aspects

8.3 Overall Model Performance Evaluation:

To evaluate the performance of the SaudiBERT model at aspect based sentiment analysis (ABSA) two visual metrics were used : the curve loss which illustrates the loss for the model trained over 9 epochs for both training loss and validation loss , as well as a confusion matrix which provides detailed analysis of the final classification results for positive , negative and neutral classes and error analysis.

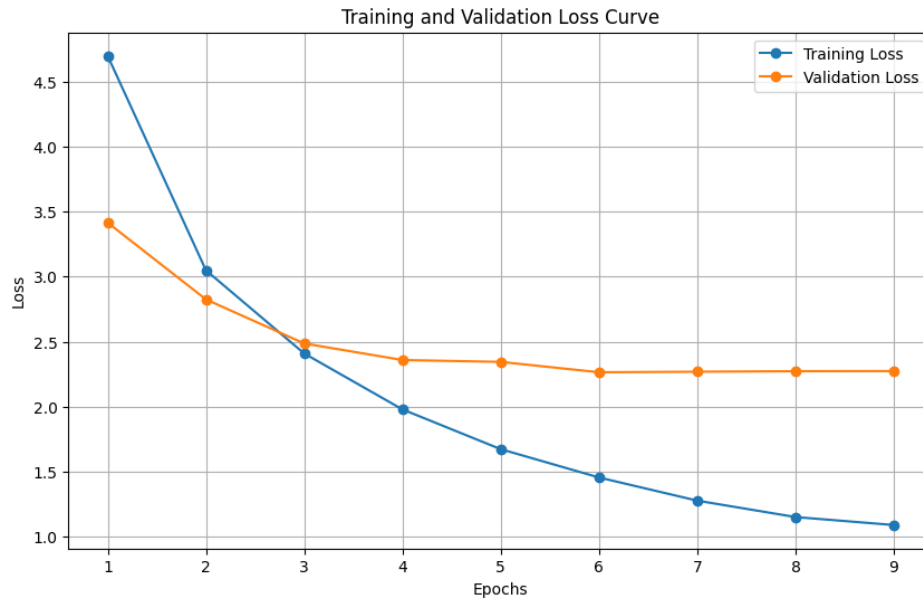


Figure 6: Training and validation loss curve

In figure, it can be observed that the loss curve shows a continuous downward trend, reflecting a gradual decrease in the model's loss as it learns from the data and progressively improves its predictions as training progresses. Figure 4 also shows that the validation curve began to converge with the training curve by the 9 epochs, which means that the model has reached a sufficient stage of learning and has avoided overfitting during training.

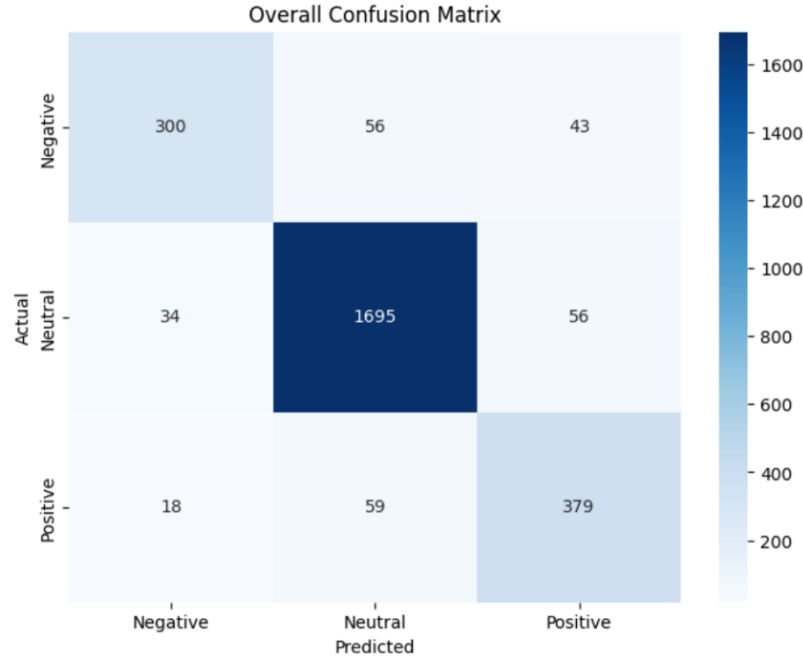


Figure 7: overall confusion Matrix

The confusion matrix in figure 5 illustrates the models final performance in classifying sentiment into three categories (positive, negative and neutral) analysis of the data shows that the model achieved the highest classification accuracy for neutral category with 1695 correct predictions , the positive category was correctly classified in 379 cases with a slight overlap of 59 cases with the neutral category and the negative category was correctly classified in 300 cases .

8.4 Comparative Performance Analysis SaudiBERT vs. CAMELBERT-DA:

A benchmark comparison was conducted between the SaudiBERT model trained on Saudi dialect data and the CAMELBERT model trained on a wide and diverse range of Arabic dialects, such as Gulf, Egyptian, and Levantine dialects. The purpose of this comparison is to test the hypothesis of whether allocating a model to a specific dialect provides accuracy in analysis or whether comprehensive models for more than one dialect are capable of delivering similar performance.

Model	Accuracy	Precision	F1-Score
SaudiBERT	90%	87%	79%
CAMeLBERT	88%	83%	74%

8.5 Performance Evaluation per Aspect:

The evaluation metrics for the individual product aspects, as detailed in table demonstrate the model robust capability in classifying diverse sentiment categories, the average accuracy reached 90% with a consolidated F1-score of 79%

Aspect	Accuracy	Precision	F1-Score
Size	90%	80%	78%
Color	93%	86%	86%
Price	99%	88%	85%
Smell	99%	99%	92%
Quality	87%	71%	58%
Fabric	83%	83%	83%
Style	80%	80%	74%
In general	88%	92%	81%

A closer look at the results reveals that objective aspects like ‘Smell’ and ‘Price’ achieved exceptional performance, with F1-score of 92% and 85% respectively , this is primarily due to the Saudi dialect , on the other hand , the ‘Quality’ aspect recorded the lowest F1-score 58% , which can be attributed to its high semantic subjectivity and overlap with other descriptors, despite these challenges, the consistently high precision across most categories confirms that the model effectively minimizes false positives classifications, providing a reliable foundation for automated sentiment analysis .

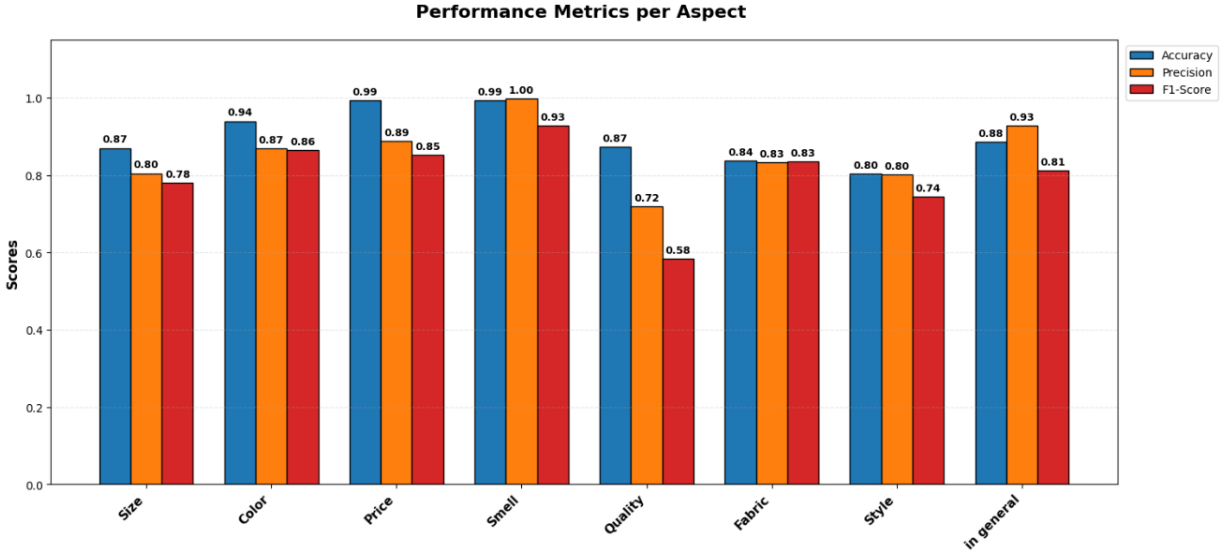


Figure 8: Model evaluation for each aspect

8.6 illustrative case study:

To verify the model's ability to analyze sentiment by aspect in complex contexts, a qualitative analysis was conducted on a review written in the Saudi dialect, the text contains "الفستان مرره بطل والخامة ممتازة بس السعر مره غالي المقاس جاء ضيق", the table shows the model's outputs for a text containing multifaceted opinions Price, Fabric and Size.

Review Text: الفستان مرره بطل والخامة ممتازة بس السعر مره غالي المقاس جاء ضيق

	Aspect	Predicted Sentiment
0	Size	-1
1	Color	0
2	Price	-1
3	Smell	0
4	Quality	0
5	Fabric	1
6	Style	0
7	in general	1

The analysis above demonstrates the model's effectiveness in distinguishing between sentiments within a single review it has proven its ability to identify praise directed at the fabric while simultaneously identifying criticism regarding Price and Size.

8.7 Discussion:

At the conclusion of this chapter, the study's central hypothesis—namely, the ability of specialized language models, specifically the SaudiBERT model, to overcome the linguistic diversity challenges posed by the Saudi dialect—has been confirmed. Achieving an overall accuracy of nearly 90% is not merely a numerical success; rather, it is an indicator of the success of the fine-tuning process in deriving the precise semantic features that distinguish Saudi reviews from other Arabic dialects.

Through analysis of the confusion matrix and qualitative results, it is clear that the model has developed 'contextual awareness' that enabled it to distinguish between overlapping sentiments toward different aspects of the product; its ability to identify a negative sentiment toward price and a positive one toward Fabric in a single sentence reflects high proficiency in handling complex ABSA tasks. Although there are some challenges posed by words with double meanings or sarcastic styles, which were observed in some classification errors, the overall results position this model as a reliable and applicable tool in Customer Experience Analytics systems. Thus, this chapter paves the way for the transition to Chapter 5, where the final recommendations and future prospects for developing this technology to include other Arabic dialects or more complex linguistic tasks will be reviewed.

Chapter 5

9 Conclusion and Future Work

9.1 Conclusion

This study concluded its research journey into exploring the prospects of natural language processing for the Saudi dialect, having successfully developed an advanced technical model based on the SaudiBERT architecture for aspect-based sentiment analysis (ABSA). Quantitative and qualitative results have demonstrated that adapting deep learning techniques to serve local dialects is not merely a technical challenge, but a strategic necessity for accurately and reliably understanding the voice of the Saudi consumer.

Achieving a high classification accuracy exceeding 90% demonstrates that specialized language models are capable of capturing the semantic and cultural complexities inherent in local texts. Despite the challenges associated with the scarcity of available linguistic resources for dialects, this research offers a practical and methodological solution that can be built upon in e-commerce applications and big data analysis. In conclusion, we hope this study will serve as a cornerstone in the edifice of national research on artificial intelligence, thereby contributing to supporting digital transformation and improving the customer experience in the Kingdom of Saudi Arabia in line with Vision 2030.

9.2 Future work

Based on the results obtained several future research directions can be identified for the development and improvement of sentiment analysis systems based on aspects of the Saudi dialect, namely:

1. **Inclusion of local dialects:** we propose that future studies expand the scope of data collection to include reviews representing all regional dialects in Saudi Arabic such as Najdi, Hijazi and others in order to train a model capable of

capturing region-specific vocabulary thereby ensuring higher classification accuracy.

2. **Expanding the Dataset:** we propose building a larger and more diverse dataset to significantly improve the model's ability to generalize results and reduce errors in rare categories.
3. **Developing an API:** to enable companies to view and utilize the analyses to improve quality.
4. **Sentiment analysis using emojis:** in future studies we propose incorporating emojis as key semantic variables in ABSA rather than simply removing them from the text.

References

- [1] Alrefae, R., Alqahmi, R., Alduraibi, M., Almatrafi, S., & Alayed, A. (2024). Enhancing Customer Experience Through Arabic Aspect-Based Sentiment Analysis of Saudi Reviews. *International Journal of Advanced Computer Science & Applications*, 15(7).
- [2] Hicham, N., Karim, S., & Habbat, N. (2023). Enhancing Arabic Sentiment Analysis in E-Commerce Reviews on Social Media Through a Stacked Ensemble Deep Learning Approach. *Mathematical Modelling of Engineering Problems*, 10(3).
- [3] Al-Thubaity, A., Alqahtani, Q., & Aljandal, A. (2018). Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia computer science*, 142, 301-307.
- [4] Alotaibi, S., Mehmood, R., & Katib, I. (2019, June). Sentiment analysis of Arabic tweets in smart cities: A review of Saudi dialect. In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)* (pp. 330-335). IEEE.
- [5] Abo, M. E. M., Qazi, A., Saad, A. A. A., Elsayib, H. A., & Abdelaziz, A. (2024, August). Sentiment Analysis in Saudi Arabic Dialect for HAJJ Season Services Using Twitter Data. In *2024 1st International Conference on Logistics (ICL)* (pp. 1-5). IEEE.
- [6] Alahmary, R. M., Al-Dossari, H. Z., & Emam, A. Z. (2019, January). Sentiment analysis of Saudi dialect using deep learning techniques. In *2019 International conference on electronics, information, and communication (ICEIC)* (pp. 1-6). IEEE.
- [7] Kaseb, A., & Farouk, M. (2022, December). SAIDS: A novel approach for sentiment analysis informed of dialect and sarcasm. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (pp. 22-30).
- [8] AlFutamani, A. A., & Al-Baity, H. H. (2021). Emotional Analysis of Arabic Saudi Dialect Tweets Using a Supervised Learning Approach. *Intelligent Automation & Soft Computing*, 29(1).
- [9] Dahbi, M., Saadane, R., & Mbarki, S. (2019, October). Social media sentiment monitoring in smart cities: an application to Moroccan dialects. In *Proceedings of the 4th International Conference on Smart City Applications* (pp. 1-6).

- [10] AlFutamani, A. A., & Al-Baity, H. H. (2021). Emotional Analysis of Arabic Saudi Dialect Tweets Using a Supervised Learning Approach. *Intelligent Automation & Soft Computing*, 29(1).
- [11] Alomari, K. M., ElSherif, H. M., & Shaalan, K. (2017, June). Arabic tweets sentimental analysis using machine learning. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 602-610). Cham: Springer International Publishing.
- [12] Louati, A., Louati, H., Kariri, E., Alaskar, F., & Alotaibi, A. (2023). Sentiment analysis of Arabic course reviews of a Saudi university using support vector machine. *Applied Sciences*, 13(23), 12539.
- [13] Musleh, D. A., Alkhwaja, I., Alkhwaja, A., Alghamdi, M., Abahussain, H., Alfawaz, F., ... & Abdulqader, M. M. (2023). Arabic sentiment analysis of youtube comments: Nlp-based machine learning approaches for content evaluation. *big data and cognitive computing*, 7(3), 127.