https://doi.org/10.22581/muet1982.3449

2025, 44(2) 197-216

Exploring the best fit: A comparative analysis of AFINN, Textblob, VADER, and Pattern on Arabic reviews for optimal dictionary extraction

Shakeel Ahmad [a], Sheikh Muhammad Saqib [b, *], Asif Hassan Syed [a], Nashwan Alromema [a], Ali Kararay [a]

[a] *Department of Computer Science, College of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah, Saudi Arabia*

[b] *Department of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan*

[*] Corresponding author: Shakeel Ahmad, Email: sarahmad@kau.edu.sa

Received: 15 February 2025, Accepted: 27 March 2025, Published: 01 April 2025

KEYWORDS

Natural Language Processing (NLP)

Deep Learning

AFINN

Textblob

ADER

Pattern.en

ABSTRACT

In the realm of natural language processing (NLP), the pivotal task of analysing affective states, including sentiment and emotion, has seen significant advancements in recent years. However, in the context of the Arabic language, studies predominantly resort to machine learning or deep learning algorithms for sentiment and emotion analysis, often neglecting the utilization of current pre-trained language models. While deep learning models tailored for Arabic text have garnered attention, there exists a considerable gap in integrating widely used tools like AFINN, TextBlob, VADER, and Pattern.en for text polarity due to compatibility issues with Arabic text. This study addresses this gap by striving to make Arabic text compatible with these dictionaries, presenting a comprehensive analysis. The findings suggest that AFINN and VADER emerge as the most suitable dictionaries for effective sentiment analysis in Arabic text. Specifically, AFINN achieved 83% accuracy, with a precision of 0.88, recall of 0.80, and an F1-score of 0.84 for negative sentiment, and a precision of 0.77, recall of 0.86, and an F1-score of 0.82 for positive sentiment. VADER demonstrated 83% accuracy, with a precision of 0.88, recall of 0.80, and an F1-score of 0.84 for negative sentiment, and a precision of 0.78, recall of 0.86, and an F1-score of 0.82 for positive sentiment. These results indicate that both AFINN and VADER are effective tools for sentiment analysis in Arabic, providing a reliable solution for text polarity classification.

1. Introduction

Arabic currently ranks fourth in internet popularity among languages [1], holding a crucial position in sentiment analysis that warrants further research. Arabic possesses distinct features setting it apart from other languages, including 28 letters without capitalization and a unique writing style where letters are connected differently when positioned within a word [2]. Additionally, the language exhibits variations based on dialects, particularly noticeable in its usage across social media platforms.

Natural Language Processing (NLP), a subset of Artificial Intelligence, plays a key role in analyzing and interpreting human language, discerning the intentions and sentiments of writers or speakers [3][4]. While NLP's popularity has surged, its applications in Arabic remain significantly fewer compared to English.

Challenges in Arabic NLP stem from limited resources, morphological complexity, and dialectical variations, as well as broader linguistic issues like implicit sentiment, sarcasm, and review quality [5].

Stemming Arabic words, especially in the context of text classification, has been a complex task [6]. Research has delved into sentiment identification challenges in informal Arabic on platforms like Twitter and YouTube [7][8], noting the unstructured nature of informal Arabic.

Existing resources for Arabic texts are constrained, leading to low accuracy in methods due to specific Arabic complexities and general linguistic issues [5]. Limited research on opinion mining using Arabic Twitter has been conducted [9], mainly because NLP tools for Arabic are often tailored for formal language. Although Arabic text is suitable for automatic analysis using algorithms like SVM, Naïve Bayes, Logistic Regression, and Random Forest [10], the majority of NLP tools, such as Affin [11], Textblob [12], Pattern [13], and Vedar [14], have been designed for English text. The objective of this research is to narrow the existing gap by utilizing Googletrans to tailor established dictionaries for Arabic text classification. The study seeks to offer a thorough comparison, identifying the most suitable dictionary for effective analysis of Arabic text.

This research makes a distinctive contribution by pioneering efforts to address the scarcity of Natural Language Processing (NLP) applications in the Arabic language, despite its standing as the fourth most popular language on the internet. Through an exploration of the unique characteristics of Arabic, encompassing its intricate writing style, dialectical variations, and challenges in stemming for text classification, the study sheds light on the specific complexities encountered in sentiment analysis. Notably, the research adopts an innovative approach by utilizing Googletrans to adapt well-established NLP dictionaries originally designed for English, including Affin, Textblob, Pattern, and Vedar. This inventive methodology, integrating the capabilities of Googletrans, seeks to fill the void in Arabic-specific tools. Beyond identifying and navigating challenges related to the linguistic nuances of Arabic, the work offers a comprehensive comparison of these adapted dictionaries, presenting a practical framework to enhance sentiment analysis within the realm of Arabic language processing.

1.1 Novelty of Work

The novelty value of this research emerges from its distinctive implementation of Googletrans for adapting widely used English-based sentiment analysis tools (Affin, Textblob, Pattern, and VADER) so they function effectively on Arabic text. The approach innovates Arabic sentiment analysis specifically by addressing the language distinct challenges of Arabic morphology and dialects along with providing valuable insight into the NLP discipline for the Arabic language.

1.2 Major Contribution

Research in Arabic Natural Language Processing (NLP) gains significant value due to its creation of sentiment analysis tools made for the Arabic language which currently remains scarce. The research implements well-known English-based NLP dictionaries (Affin, Textblob, Pattern, and VADER) for Arabic text classification through Googletrans which addresses problems related to Arabic writing styles and dialectical differences and stemming intricacies. These adapted dictionaries receive an in-depth assessment within the research which establishes a functional approach for Arabic sentiment analysis enhancement.

1.3 Application Area

Sentiment analysis of Arabic text operates as a fundamental tool which multiple industries can use across several domains. Social media monitoring becomes more effective when automated because businesses along with governments benefit from public sentiment tracking and event or product reaction prediction. The evaluation system allows businesses to quickly discover client satisfaction levels and spot problems for service enhancement in customer review analysis. The measurement of public sentiment and political policy assessment benefits greatly from sentiment analysis because Arabic-language social media has substantial importance in political regions. Sentiment analysis enables healthcare monitoring of mental health through digital content evaluation while market research relies on the tool to track product or service consumer sentiments and media monitoring tracks public reactions to news stories. The system helps e-commerce platforms improve their services because it tracks consumer opinions through review analysis and assists businesses in protecting their brand through public perception tracking. Sentiment analysis tools AFINN and VADER that process Arabic text effectively support organizations to enhance their decision processes and develop strategies and optimize their services across all these domains.

This paper follows this structure: Section-2 contains a literature review and Section-3 presents the proposed method followed by Experimental Results and Discussions in Section-4 then Section-5 includes Comprehensive Analysis of Dictionaries Based On Arabic Text, and Section-6 describes benchmark comparison while this paper's conclusion with limitations appears in Sections 7 and 8.

2. Literature Review

Not only has machine learning been applied to image analysis [15][16][17], but it has also been extensively utilized in text analysis [18][19][20].

Utilizing the internet, a research investigation [21] has delved into Arabic and Islamic content containing pertinent information extracted from prophetic narrations' texts employing Artificial Intelligence (AI) methodology. The authors advocate a semantically driven approach for analyzing Arabic discourse within the framework of Segmented Discourse Representation Theory (SDRT). Additionally, discourse analysis has proven effective in generating indicative summaries of Arabic documents [22]. In a related study, a novel summarization model was proposed by the authors of [23], leveraging clusters from document clustering and key phrase extraction. Another approach, termed "Adaboost", was introduced by the authors of [24], utilizing a supervised method for Arabic summary extraction [25]. This study incorporates a set of statistical features, including sentence position, keyword count in a sentence, overlap with the document title, and sentence length. The proposed approach employs a Genetic algorithm and the MapReduce parallel programming model for automatic text summarization of extensive Arabic documents. Notably, this research underscores the significance of scalability, speed, and accuracy in the summarization process, areas that have received limited attention in prior studies [22].

Numerous studies in Text Classification (TC) [26] [27][28] have been conducted and evaluated using languages such as English, French, German, Spanish, Chinese, Greek, and Japanese [29]. Opinion mining [30][31], which involves automatically identifying opinions expressed in Arabic texts on specific subjects [32][33][34], faces challenges in the automatic classification of text documents in Arabic. This is attributed to issues such as diverse spellings of certain words, variations in the writing of specific character combinations, the presence of short (diacritics) and long vowels, and a deficiency of appropriate tools [5][35].

In contrast to the proposed work, which focuses on adapting widely used sentiment analysis tools like AFINN and VADER for Arabic text, achieving 83% accuracy with reliable precision and recall scores, the work of [36] introduces a more complex deep learning-based model, Arb-MCNN-Bi, utilizing AraBERT with MCNN and BiGRU to achieve accuracies as high as 96.92% on specific datasets. Both approaches contribute to Arabic sentiment analysis, but the proposed work offers a simpler, more accessible solution with strong performance across different sentiment categories, making it a valuable alternative for real-world applications where deep learning models may not be feasible due to resource constraints.

The adopted AFINN and VADER approach for Arabic sentiment analysis reached a 83% accuracy level with solid precision along with recall and F1-score metrics. The proposed method provides a dictionary-based solution which outperforms [37] because it utilizes SVM along with other machine learning models yet reaches only 92.4% accuracy in Twitter data analysis. The proposed work presents a practical solution with effective performance suitable for limited resource applications even though [37] targets complex models and feature extraction for Arabic sentiment analysis.

The proposed work demonstrates how sentiment analysis tools AFINN and VADER can become effective for Arabic text analysis by reaching 83% accuracy following precision and recall and F1-score measurements. The proposed method delivers an efficient strategy to perform Arabic sentiment analysis. A complex method which combined feature extractions and ensemble learning to evaluate Twitter reviews in Arabic obtained accuracy rates higher than 95% according to [38]. The proposed work delivers an accessible and reliable alternative to the sophisticated learning solution in [38] thereby better serving applications with limited resources or simpler requirements.

Extensive sentiment analysis research has been conducted on Arabic text though research [39][40][41] shows that popular text analysis tools including AFINN, Textblob, VADER, and pattern.en have not received significant application on Arabic text.

The limited integration of these tools with Arabic stems from inherent incompatibility issues, hindering their seamless implementation in the analysis of sentiment in this linguistic context. This unexplored terrain presents an opportunity for research to bridge the gap and adapt these widely used tools to better cater to the unique characteristics of Arabic language sentiment analysis.

3. Proposed Methodology

The Arabic dataset will undergo translation using Googletrans, converting all content into English text. Subsequently, this English text will be subjected to each dictionary, determining the class of each segment. These identified classes will then be compared with the actual class of the original Arabic text, facilitating the identification of the most effective dictionary for Arabic text, as illustrated in Fig. 1.

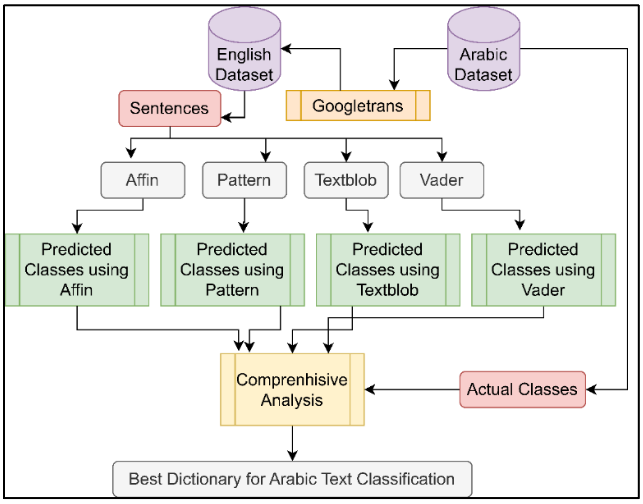

Fig. 1. Proposed Framework

3.1 Googletrans

A freely accessible and unrestricted Python library, Googletrans, has been developed to leverage the Google Translate API. This library utilizes the Google Translate Ajax API, making calls to methods like detect and translate. The efficacy of Googletrans is attributed to its ability to navigate recent updates in Google's translation service, which now incorporates a ticket mechanism to thwart numerous crawler programs. The library's success stems from reverse engineering the obfuscated and minified code employed by Google to generate tokens. This implementation, built on Python, provides a workaround for generating tickets, although it is important to note that this approach may face potential blocking in the future [42]. Table-1 outlines the procedure for translating all Arabic reviews into English using the googletrans package, utilizing the designated keyword 'ar' to indicate the Arabic language. The translated dataset serves to bridge the gap between state-of-the-art automatic dictionaries and Arabic text.

Table 1

Algorithm for Arabic to English Dataset Conversion using 'googletrans'

Step 1: Import required libraries
- Import pandas as pd
- Import Translator from googletrans

Step 2: Load data from CSV into a pandas DataFrame
- df = pd.read_csv('ArabicText.csv')

Step 3: Extract Arabic reviews from the DataFrame
- ArabicDataSet = df['ArabicText'].tolist()

Step 4: Initialize Translator
- translator = Translator()

Step 5: Create a list to store translated English reviews
- TranslatedEnglishReviews = []

Step 6: Translate each Arabic review to English
- For each ArabicReview in ArabicDataSet:
  - EnglishReview = translator.translate(ArabicReview, src="ar").text
  - Append EnglishReview to TranslatedEnglishReviews

Step 7: Print or use TranslatedEnglishReviews as needed
- print(TranslatedEnglishReviews)

3.2 AFINN Dictionary for Arabic Text Classification

The AFINN lexicon consists of English terms manually rated for valence by Finn Årup Nielsen between 2009 and 2011, assigning integer values between -5 (negative) and +5 (positive). Excluding multi-word phrases, the original lexicon is subject to the Open Database License (ODbL) v1.0 [43], allowing freedom to share, create derivative works, and adapt the lexicon with the condition of proper attribution [44]. Table-2 presents the algorithm for categorizing Arabic translated reviews through the utilization of the AFINN dictionary.

Table 2

Arabic Text Classification using Translated Text and AFINN Dictionary

Step 1: Import Libraries
a. Import Afinn library
b. Import pandas library
Step 2: Read Data
a. Read data from 'D:\AffinArabic\DataSet\AabicToEnglishDataset.xlsx' using pandas
b. Store the DataFrame in a variable 'df'
Step 3: Extract Columns
a. Extract 'English' column from 'df' and store it in 'sentences'
b. Extract 'Class' column from 'df' and store it in 'actual_class'
c. Initialize an empty list 'predicted_class'
Step 4: Initialize Afinn
a. Create an Afinn object and store it in 'afinn'
Step 5: Sentiment Analysis Loop
a. For each sentence in 'sentences':
- Step 5a: Calculate Sentiment Score
  a. Use 'afinn.score(sentence)' to get sentiment score and store it in 'sentiment_score'
- Step 5b: Determine Sentiment Label
  a. If 'sentiment_score > 0', then set 'sentiment' to "positive"
  b. If 'sentiment_score <= 0', then set 'sentiment' to "Negative"
- Step 5c: Append to 'predicted_class'

a. If 'sentiment_score > 0', then append "Positive" to 'predicted_class'
b. If 'sentiment_score <= 0', then append "Negative" to 'predicted_class'
- Step 5d: Output Results
a. Print "Sentence: '" + sentence + "'"
b. Print 'Sentiment: ' + sentiment
c. Print 'Score: ' + sentiment_score

The program starts its operation by importing necessary libraries which include Afinn for sentiment analysis and pandas for data manipulation. The program then reads content from the Excel file named 'AabicToEnglishDataset.xlsx' which gets processed into a DataFrame named 'df'. The algorithm uses pandas to extract specific column data from the DataFrame because it needs the text in the 'English' column and the corresponding sentiment labels in the 'Class' column. The algorithm initializes an empty list named 'predicted_class' which will contain the sentiment labels that the algorithm predicts.

The Afinn library provides sentiment scoring for all English sentences found in the 'English' column. When sentiment scores exceed zero the algorithm assigns the sentiment as "positive" while adding "Positive" to the 'predicted_class' list. The algorithm labels the sentiment "Negative" when the score falls below or matches zero because the list receives the "Negative" label at this point. The algorithm prints information about each sentence, its determined sentiment, and the corresponding sentiment score.

3.3 TextBlobs Dictionary for Arabic Text Classification

TextBlob operates as a flexible Python library which operates across Python 2 and 3 to process whole textual information. Its straightforward API facilitates seamless engagement in various natural language processing (NLP) endeavors. The TextBlob Python module performs multiple NLP functions which include both part-of-speech tagging and noun phrase extraction together with sentiment analysis and classificaiton and translation capabilities and other tasks. Developers alongside data scientists find TextBlob library highly beneficial when they need an effective Python solution for executing different NLP tasks across their programming work. This user-friendly platform stands out because of its wide range of features which attracts development projects that need text analysis capabilities [45]. Table-3 outlines the algorithm for classifying Arabic translated reviews by employing the TextBlob dictionary.

Table 3

Arabic Text Classification using Translated Text and Textblob Dictionary

Step 1: Import Libraries
a. Import TextBlob library
b. Import pandas library
Step 2: Read Data
a. Read data from 'D:\AffinArabic\DataSet\AabicToEnglishDataset.xlsx' using pandas
b. Store the DataFrame in a variable 'df'
Step 3: Extract Columns
a. Extract 'English' column from 'df' and store it in 'sentences'
b. Extract 'Class' column from 'df' and store it in 'actual_class'
c. Initialize an empty list 'predicted_class'
Step 4: Sentiment Analysis Loop
a. For each text in 'sentences':
- Step 4a: Create TextBlob object
a. Create a TextBlob object using the current text and store it in 'blob'
- Step 4b: Get Sentiment Polarity
a. Get the sentiment polarity from the TextBlob object and store it in 'sentiment_polarity'
- Step 4c: Determine Sentiment Label
a. If 'sentiment_polarity > 0', set 'sentiment' to 'Positive' and append 'Positive' to 'predicted_class'
b. If 'sentiment_polarity <= 0', set 'sentiment' to 'Negative' and append 'Negative' to 'predicted_class'
- Step 4d: Output Results
a. Print "Text: " + current text
b. Print "Sentiment Polarity: " + sentiment_polarity
c. Print "Sentiment: " + sentiment

A sentiment analysis process using TextBlob library operates on a dataset according to the pseudocode specification. The program first imports libraries in Step 1 while Step 2 performs Excel data reading and DataFrame storage. At step three the code selects data from two essential columns which comprise 'English' for sentence data and 'Class' for original sentiment measurements. A new empty list named predicted_class receives storage for future sentiment predictions. The loop from Step 4 runs through every element contained in the 'sentences' variable. Each text generates a TextBlob object during the loop execution and the algorithm processes its sentiment polarity calculation. Within the list a Positive or Negative sentiment label gets assigned according to polarity measurement before insertion into the 'predicted_class' list. The algorithm prints the text, sentiment polarity, and sentiment label for each iteration.

3.4 VADER Dictionary for Arabic Text Classification

VADER, short for Valence Aware Dictionary and Sentiment Reasoner, stands out as a sentiment analysis tool that relies on a lexicon and rule-based approach. It is finely tuned to decipher sentiments articulated in the realm of social media. Operating on a sentiment lexicon – a compilation of lexical features like words, annotated based on their semantic orientation as positive or negative – VADER not only provides information about positivity and negativity scores but also offers insights into the degree of positivity or negativity expressed in a sentiment [46]. Table-4 delineates the algorithm for categorizing Arabic translated reviews using the VADER dictionary.

Table 4

Arabic Text Classification using Translated Text and VADER Dictionary

```
Step 1: Import Libraries
a. Import nltk library
b. Import SentimentIntensityAnalyzer from nltk.sentiment
c. Import pandas library
d. Import numpy library
Step 2: Read Data
a. Read data from 'D:\AffinArabic\DataSet\AabicToEnglishDataset.xlsx' using pandas
b. Store the DataFrame in a variable 'df'
Step 3: Extract Columns
a. Extract 'English' column from 'df' and store it in 'sentences'
b. Extract 'Class' column from 'df' and store it in 'actual_class'
Step 4: Download Resource for NLTK
a. Download the 'vader_lexicon' resource using nltk
Step 5: Initialize Sentiment Intensity Analyzer
a. Create a SentimentIntensityAnalyzer object and store it in 'sid'
Step 6: Initialize Predicted Class List
a. Initialize an empty list named 'predicted_class'
Step 7: Sentiment Analysis Loop
a. For each sentence in 'sentences':
   - Step 7a: Calculate Sentiment Scores
     a. Use 'sid.polarity_scores(sentence)' to get sentiment scores and store them in 'sentiment_scores'
   - Step 7b: Output Results
     a. Print "Sentence: " + current sentence
     b. Print "Sentiment Scores: " + sentiment_scores
   - Step 7c: Determine Sentiment based on Compound Score
     a. If 'sentiment_scores['pos'] > sentiment_scores['neg']', then print "Sentiment: Positive" and append 'Positive' to 'predicted_class'
     b. Else, print "Sentiment: Negative" and append 'Negative' to 'predicted_class'
```

The above pseudocode outlines the steps for sentiment analysis using the NLTK library's SentimentIntensityAnalyzer. It starts by importing necessary libraries, including nltk for natural language processing and pandas for data manipulation. The algorithm then reads a dataset from an Excel file, extracting the 'English' column for sentences and the 'Class' column for actual sentiment labels. The next step involves downloading a resource ('vader_lexicon') needed for sentiment analysis using NLTK. Subsequently, it initializes a SentimentIntensityAnalyzer. A list named 'predicted_class' is initialized to store predicted sentiment labels. The algorithm then enters a loop, where for each sentence, it calculates sentiment scores, prints the sentence and its scores, and determines the sentiment (Positive or Negative) based on the compound score. The predicted sentiment labels are appended to the 'predicted_class' list. This pseudocode provides a structured representation of the sentiment analysis process using NLTK.

3.5 Pattern Dictionary for Arabic Text Classification

Inside the pattern.en module one finds a comprehensive set of linguistic tools that strengthen the capabilities of processing English text. The module contains a speedy part-of-speech tagger that effectively recognizes the grammatical elements such as nouns and verbs within text. The tool contains an English sentiment evaluation system that assesses written text sentiments within the pattern.en module. The module enhances linguistic functionality by offering tools which allow English verb form manipulation along with noun singularization/pluralization features which enable better control over noun variations. The module combines perfectly with WordNet which provides students with direct access to lexical databases so they can develop detailed semantic insights. [47]. Table-5 outlines the algorithm for classifying Arabic translated reviews by employing the Pattern dictionary.

Table 5

Arabic Text Classification using Translated Text and Pattern Dictionary

```
Step 1: Import Libraries
a. Import sentiment function from pattern.en
b. Import pandas library
c. Import numpy library
```

Step 2: Read Data
a. Read data from 'D:\AffinArabic\DataSet\AabicToEnglishDataset.xlsx' using pandas
b. Store the DataFrame in a variable 'df'
Step 3: Extract Columns
a. Extract 'English' column from 'df' and store it in 'sentences'
b. Extract 'Class' column from 'df' and store it in 'actual_class'
Step 4: Initialize Predicted Class List
a. Initialize an empty list named 'predicted_class'
Step 5: Sentiment Analysis Loop
a. For each text in 'sentences':
- Step 5a: Perform Sentiment Analysis
a. Use 'sentiment(text)' from pattern.en to analyze the sentiment of the current text and store the result in 'analysis_result'
- Step 5b: Extract Sentiment Polarity and Subjectivity
a. Extract sentiment polarity and subjectivity from 'analysis_result' and store them in 'sentiment_polarity' and 'sentiment_subjectivity' respectively
- Step 5c: Determine Sentiment Label
a. If 'sentiment_polarity >= 0', set 'sentiment_label' to 'Positive' and append 'Positive' to 'predicted_class'
b. Else, set 'sentiment_label' to 'Negative' and append 'Negative' to 'predicted_class'
- Step 5d: Output Results
a. Print "Text: " + current text
b. Print "Sentiment Polarity: " + sentiment_polarity
c. Print "Sentiment Subjectivity: " + sentiment_subjectivity
d. Print "Sentiment Label: " + sentiment_label

The above pseudocode outlines the process of sentiment analysis using the Pattern library's sentiment function on a dataset. Initially, it imports necessary libraries, including Pattern for sentiment analysis and pandas for data handling. The algorithm then reads a dataset from an Excel file, extracting the 'English' column for sentences and the 'Class' column for actual sentiment labels. It initializes an empty list named 'predicted_class' to store the predicted sentiment labels. The algorithm enters a loop, where for each text in the 'sentences' variable, it performs sentiment analysis using Pattern's sentiment function. It extracts sentiment polarity and subjectivity from the analysis result and determines the sentiment label based on the polarity. The predicted sentiment labels are appended to the 'predicted_class' list. The algorithm prints information about each text, including the sentiment polarity, subjectivity, and label. This pseudocode provides a structured representation of sentiment analysis using Pattern.

4. Experimental Results and Discussions

The dataset titled "Arabic_Reviews_Sentiment_analysis.csv" was sourced from Kaggle [48], comprising 449 Arabic reviews, where 229 are categorized as negative (0) and 220 as positive (1). The dataset underwent processing through the algorithm detailed in Table-1, involving the translation of data into English. Table-6 provides a glimpse of sample reviews from both the original Arabic dataset and the translated English dataset, showcasing instances of both negative and positive sentiments.

Table 6

Sample of Dataset after passing Arabic Dataset to Algorithm in Table-1

Arabic	English	Class
من الصباح الي الأن التطبيق ماهو شغال يقولك صيانة عطلو مصالح الناس	From morning until now, the application is not working, it tells you that maintenance is disrupting people's interests	Negative
التطبيق يعلق مره كثير	The app hangs a lot	Negative
سي سي جداً	CC very much	Negative
تطبيق سيء جدا وكثير التعليق والعطل .. غرررريبه	Very bad application and a lot of comments and malfunctions	Negative
مو راضي يدخلني يكتب (لم نتمكن من تنفيذ طلبك، الرجاء المحاولة مرة أخرى) ارجو حل المشكلة.	You are not satisfied with me, write (we were unable to fulfill your request, please try again) I hope the problem is resolved.	Negative
التطبيق أصبح فاشل بعد التحديث الأخير.. ولازال فاشل حتى بعد 4/4	The application has become a failure after the last update.. and it is still failing even after 4/4	Negative

للاسف التطبيق لا يعمل ارجو تصحيح الاخطاء والعمل على حلها	Unfortunately, the application does not work. Please correct the errors and work to solve them	Negative
جيد ولكن اتمنى إضافة المزيد من المطاعم والماركت	Good, but I hope to add more restaurants and markets	Positive
التطبيق جيد جدا بعد أن حلو المشكلة التي كنت اوجهها #التطبيق_سريع_التوصيل	The app is very good after it solved the problem I was facing #fast_delivery_app	Positive
التطبيق راءع	Awesome app	Positive
اخب طلبات للامانه احب اهتمامهم وسرعه حل الإشكالية، بس ف مشكله تخليني ابحث عن برنامج بديل،،، والسبب انه مره الجفع بالبطاقع يروح ومره لا واتصل بالبنك يقولو الإشكالية من البرنامج، من امس ابا اطلب من سن مارت بالبطاقه لا فائده	I tell Talabat to be honest. I like their interest and the speed of solving the problem, but I have a problem that makes me look for an alternative program. The reason is that sometimes the card goes away and sometimes it doesn't. I call the bank and they say the problem is from the program. Since yesterday, I asked Sun Mart with the card, it's useless.	Positive
اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98 ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم	Today I ordered the same order more than 10 times, and each time they deducted 98 riyals from me, and the order did not arrive.	Positive

4.1 AFINN Based Sentiment Analysis of Arabic Reviews

The complete dataset, a sample of which is presented in Table-6, underwent processing through the Affin-Based Algorithm outlined in Table-2. The results for the entire dataset, as well as a subset, are displayed in Table-7. The predicted score is calculated using the AFINN dictionary, where a score greater than 0 indicates a positive sentiment, while a score less than or equal to 0 signifies a negative sentiment. The predicted class is determined based on the predicted score, as illustrated in the sample data from the dataset showcased in Table-7.

Table 7

Results Based on AFINN Based Sentiment Analysis

Sentence	Actual Class	Predicted Score	Predicted Class
جيد ولكن اتمنى إضافة المزيد من المطاعم والماركت	Positive	5.0	positive
التطبيق جيد جدا بعد أن حلو المشكلة التي كنت اوجهها #التطبيق_سريع_التوصيل	Positive	2.0	positive
التطبيق راءع	Positive	4.0	positive
اخب طلبات للامانه احب اهتمامهم وسرعه حل الإشكالية، بس ف مشكله تخليني ابحث عن برنامج بديل،،، والسبب انه مره الجفع بالبطاقع يروح ومره لا واتصل بالبنك يقولو الإشكالية من البرنامج، من امس ابا اطلب من سن مارت بالبطاقه لا فائده	Positive	-2.0	Negative
اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98 ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم	Positive	0.0	Negative
من الصباح الي الأن التطبيق ماهو شغال يقولك صيانة عطلو مصالح الناس	Negative	-4.0	Negative
التطبيق يعلق مره كثير	Negative	0.0	Negative

سي سي جداً	Negative	0.0	Negative
تطبيق سيء جدا وكثير التعليق والعطل .. غرررريه	Negative	-3.0	Negative
مو راضي يدخلني يكتب (لم نتمكن من تنفيذ طلبك، الرجاء المحاولة مرة أخرى) ارجو حل المشكلة.	Negative	5.0	positive
التطبيق أصبح فاشل بعد التحديث الأخير.. ولازال فاشل حتى بعد 4/4	Negative	-4.0	Negative
للاسف التطبيق لا يعمل ارجو تصحيح الاخطاء والعمل على حلها	Negative	-3.0	Negative

4.1.1 Analysis of AFINN

Applying the AFINN dictionary to Arabic text, the evaluation revealed 202 instances of true negatives and 50 instances of false negatives. Additionally, there were 170 occurrences of true positives and 27 instances of false positives, as depicted in the accompanying Fig. 2.

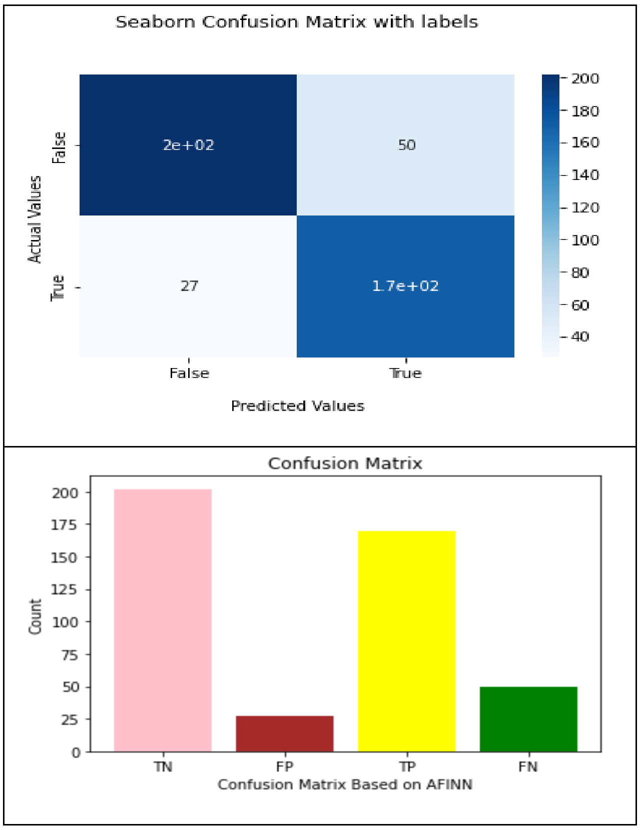

Fig. 2. AFINN Based Confusion Matrix

Leveraging the presented confusion matrix, the calculated average values for precision, recall, F1 score, and accuracy stand at 83%, 83%, 83%, and 83%, respectively, as indicated in Table-8.

Table 8

Confusion Matrix Measures using AFINN

Class	Precision	Recall	F1-Score
Negative	0.88	0.80	0.84
Positive	0.77	0.86	0.82
Average	0.83	0.83	0.83
Accuracy	0.83		

4.2 Textblob Based Sentiment Analysis of Arabic Reviews

The entire dataset, with a sample shown in Table-6, was processed using the Textblob-Based Algorithm described in Table-3. The outcomes for both the entire dataset and a subset are shown in Table-9. The predicted score is figured out using the Textblob dictionary – a score above 0 means a positive sentiment, while a score less than or equal to 0 means a negative sentiment. The predicted class is then determined based on this score, as demonstrated in the sample data from Table-9.

Table 9

Results Based on Textblob Based Sentiment Analysis

Sentence	Actual Class	Predicted Score	Predicted Class
جيد ولكن اتمنى إضافة المزيد من المطاعم والماركت	Positive	0.6	Positive
التطبيق جيد جدا بعد أن حلو المشكلة التي كنت اوجهها #التطبيق_سريع_ال توصيل	Positive	0.909999999 9999999	Positive
التطبيق راءع	Positive	1.0	Positive
اخب طلبات للامانه احب اهتمامهم وسرعه حل الإشكالية، بس ف مشكله تخليني ابحث عن برنامج بديل،،، والسبب انه مره الجفع بالبطاقع يروح ومره لا واتصل بالبنك يقولو الإشكالية من البرنامج، من امس ابا اطلب من سن مارت بالبطاقه لا فائده	Positive	0.049999999 99999999	Positive
اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98	Positive	0.25	Positive

ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم			
من الصباح الي الأن التطبيق ماهو شغال يقولك صيانة عطلو مصالح الناس	Negative	0.0	Negative
التطبيق يعلق مره كثير	Negative	0.0	Negative
سي سي جداً	Negative	0.26	Positive
تطبيق سيء جدا وكثير التعليق والعطل .. غرررريبه	Negative	-0.9099999999999998	Negative
مو راضي يدخلني يكتب (لم نتمكن من تنفيذ طلبك، الرجاء المحاولة مرة أخرى) ارجو حل المشكلة.	Negative	-0.375	Negative
التطبيق أصبح فاشل بعد التحديث الأخير.. ولازال فاشل حتى بعد 4/4	Negative	-0.15833333333333335	Negative
للاسف التطبيق لا يعمل ارجو تصحيح الاخطاء والعمل على حلها	Negative	-0.5	Negative

4.2.1 Analysis of Textblob

Using the Textblob dictionary on Arabic text, the assessment showed 188 cases of correct negatives and 50 cases of incorrect negatives. Moreover, there were 170 instances of correct positives and 41 instances of incorrect positives, as illustrated in the accompanying Fig. 3.

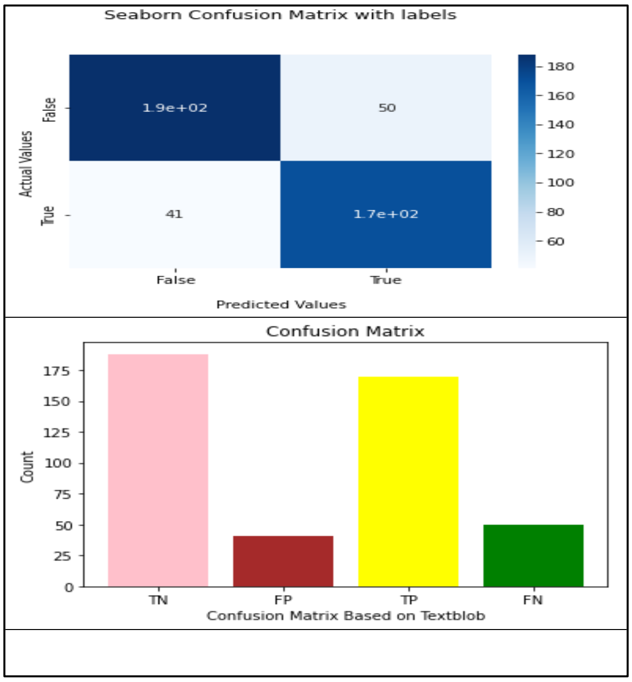

Figure 3. Textblob Based Confusion Matrix

Leveraging the presented confusion matrix, the calculated average values for precision, recall, F1 score, and accuracy stand at 80%, 80%, 80%, and 80%, respectively, as indicated in Table-10.

Table 10

Confusion Matrix Measures using Textblob

Class	Precision	Recall	F1-Score
Negative	0.82	0.79	0.81
Positive	0.77	0.81	0.79
Average	0.80	0.80	0.80
Accuracy	0.79		

4.3 VADER Based Sentiment Analysis of Arabic Reviews

The complete dataset, exemplified in Table-6, underwent analysis through the VADER-Based Algorithm delineated in Table-4. The results for both the entire dataset and a subset are detailed in Table-11. The predicted score is determined using the VADER dictionary, considering it positive if the positive score surpasses the negative score, and negative otherwise. The predicted class is subsequently assigned based on this score, as exemplified in the sample data from Table-11.

Table 11

Results Based on VADER Based Sentiment Analysis

Arabic	Actual Class	Pos Score	Neg Score	Predicted Class
جيد ولكن اتمنى إضافة المزيد من المطاعم والماركت	Positive	0.453	0.0	Positive
التطبيق جيد جدا بعد أن حلو المشكلة التي كنت اوجهها #التطبيق_سريع_التوصيل	Positive	0.294	0.15	Positive
التطبيق راءع	Positive	0.804	0.0	Positive
اخب طلبات للامانه احب اهتمامهم وسرعه حل الإشكالية، بس ف مشكله تخليني ابحث عن برنامج بديل،،، والسبب انه مره الجفع بالبطاقع يروح ومره لا واتصل	Positive	0.104	0.173	Negative

بالبنك يقولو الإشكالية من البرنامج، من امس ابا اطلب من سن مارت بالبطاقه لا فائده				
اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98 ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم	Positive	0.0	0.0	Negative
من الصباح الي الأن التطبيق ماهو شغال يقولك صيانة عطلو مصالح الناس	Negative	0.105	0.0	Positive
التطبيق يعلق مره كثير	Negative	0.0	0.0	Negative
سي سي جداً	Negative	0.0	0.0	Negative
تطبيق سيء جدا وكثير التعليق والعطل .. غرررربيه	Negative	0.0	0.322	Negative
مو راضي يدخلني يكتب (لم نتمكن من تنفيذ طلبك، الرجاء المحاولة مرة أخرى) ارجو حل المشكلة.	Negative	0.318	0.163	Positive
التطبيق أصبح فاشل بعد التحديث الأخير.. ولازال فاشل حتى بعد 4/4	Negative	0.0	0.306	Negative
للاسف التطبيق لا يعمل ارجو تصحيح الاخطاء والعمل على حلها	Negative	0.092	0.346	Negative

4.3.1 Analysis of VADER

Using the VADER dictionary on Arabic text, the assessment showed 202 cases of correct negatives and 49 cases of incorrect negatives. Moreover, there were 171 instances of correct positives and 27 instances of incorrect positives, as illustrated in the accompanying Fig. 4.

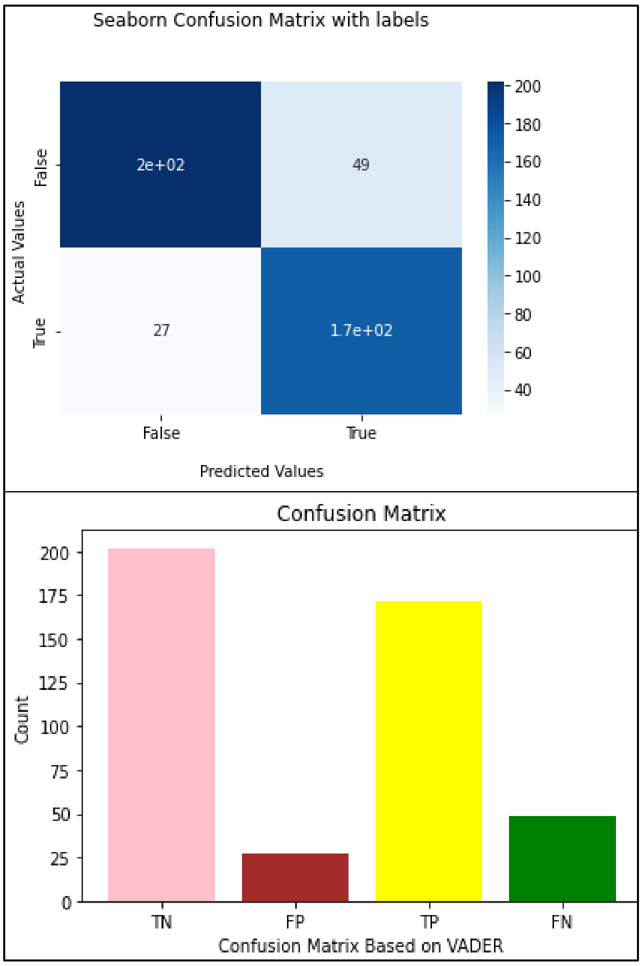

Fig. 4. VADER Based Confusion Matrix

Leveraging the presented confusion matrix, the calculated average values for precision, recall, F1 score, and accuracy stand at 83%, 83%, 83%, and 83%, respectively, as indicated in Table-12.

Table 12

Confusion Matrix Measures using Vader

Class	Precision	Recall	F1-Score
Negative	0.88	0.80	0.84
Positive	0.78	0.86	0.82
Average	0.84	0.83	0.83
Accuracy	0.83		

4.4 Pattern.en Based Sentiment Analysis of Arabic Reviews

The entire dataset, with a sample shown in Table-6, was processed using the pattern.en-Based Algorithm described in Table-5. The outcomes for both the entire dataset and a subset are shown in Table-13. The predicted score is figured out using the pattern dictionary – a score above 0 means a positive sentiment, while a score less than or equal to 0 means a negative sentiment. The predicted class is then determined based on this score, as demonstrated in the sample data from Table-13.

Table 13

Results Based on pattern.en Based Sentiment Analysis

Sentence	Actual Class	Predicted Score	Predicted Class
جيد ولكن اتمنى إضافة المزيد من المطاعم والماركت	Positive	0.6	Positive
التطبيق جيد جدا بعد أن حلو المشكلة التي كنت اوجهها #التطبيق_سريع_التوصيل	Positive	0.9099999999999999	Positive
التطبيق راءع	Positive	1.0	Positive
اخب طلبات للامانه احب اهتمامهم وسرعه حل الإشكالية، بس ف مشكله تخليني ابحث عن برنامج بديل،،، والسبب انه مره الجفع بالبطاقع يروح ومره لا واتصل بالبنك يقولو الإشكالية من البرنامج، من امس ابا اطلب من سن مارت بالبطاقه لا فائده	Positive	0.04999999999999999	Positive
اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98 ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم	Positive	0.25	Positive
من الصباح الي الأن التطبيق ماهو شغال يقولك صيانة عطلو مصالح الناس	Negative	0.0	Positive
التطبيق يعلق مره كثير	Negative	0.0	Positive
سي سي جداً	Negative	0.26	Positive
تطبيق سيء جدا وكثير التعليق والعطل .. غرررربيه	Negative	-0.9099999999999998	Negative
مو راضي يدخلني يكتب) لم نتمكن من تنفيذ طلبك، الرجاء المحاولة مرة أخرى (ارجو حل المشكلة.	Negative	-0.375	Negative
التطبيق أصبح فاشل بعد التحديث الأخير.. ولازال فاشل حتى بعد 4/4	Negative	-0.1583333333333335	Negative
للاسف التطبيق لا يعمل ارجو تصحيح الاخطاء والعمل على حلها	Negative	-0.5	Negative

4.4.1. Analysis of Pattern.en

Using the pattern.en dictionary on Arabic text, the assessment showed 110 cases of correct negatives and 119 cases of incorrect negatives. Moreover, there were 210 instances of correct positives and 10 instances of incorrect positives, as illustrated in the accompanying Fig. 5.

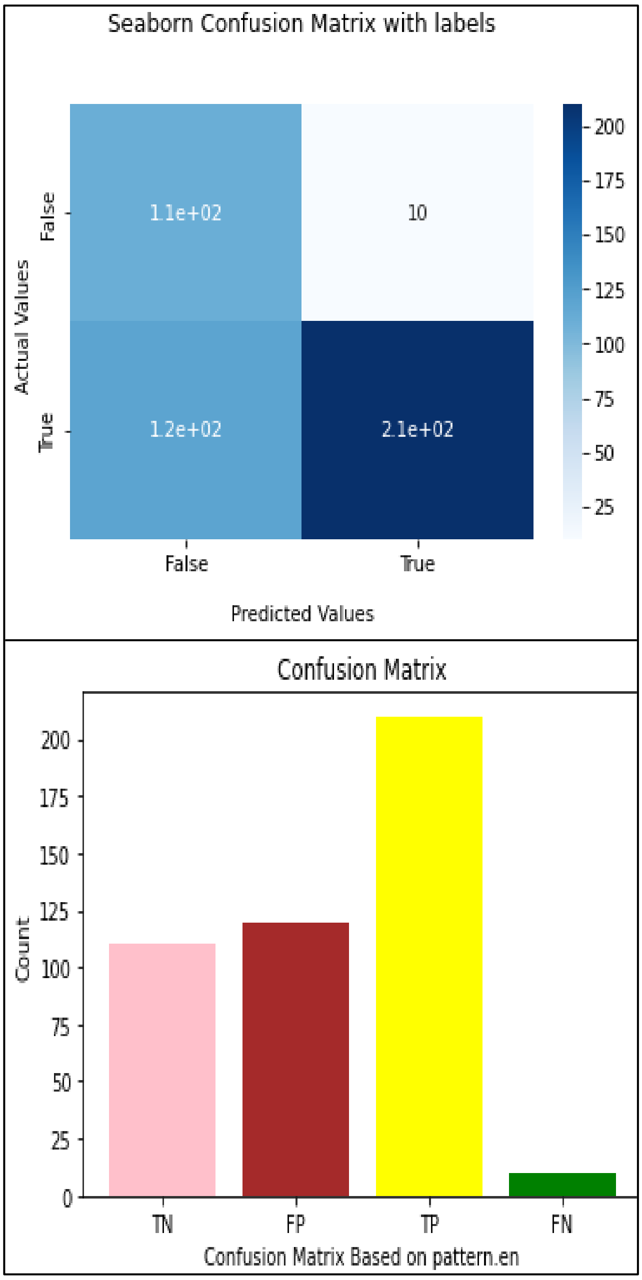

Fig. 5. pattern.en Based Confusion Matrix

Leveraging the presented confusion matrix, the calculated average values for precision, recall, F1 score, and accuracy stand at 83%, 71%, 73%, and 71%, respectively, as indicated in Table-14.

Table 14

Confusion Matrix Measures using pattern.en

Class	Precision	Recall	F1-Score
Negative	0.48	0.92	0.63
Positive	0.95	0.64	0.77
Average	0.83	0.71	0.73
Accuracy	0.71		

5. Comprehensive Analysis of Dictionaries Based On Arabic Text

Dictionaries is meticulously detailed in Section-5 across various dimensions.

For instance, the review in Arabic that reads, "اليوم طلبت نفس الطلب اكثر من ١٠ مرات وكل مره يخصمو مني 98 ريال وما وصل الطلب ارجو تصليح العطل+ لا اريد المال حلال عليكم." is inherently positive. However, there are disparities in how various dictionaries interpret it. Affin deems it negative, Textblob as positive, VADER as negative, and pattern.en as positive. Such discrepancies highlight that certain reviews are accurately predicted by one dictionary, while others are predicted differently, resulting in a varied performance range of 70% to 85% across all dictionaries on Arabic reviews. A detailed technical comparison is expounded upon in the subsequent sections.

5.1 Analysis of Precision for all Dictionaries

Precision in a confusion matrix is a metric that assesses the accuracy of positive predictions made by a classification model. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives. In other words, precision helps measure the model's ability to avoid falsely labeling instances as positive. A higher precision indicates fewer false positives, demonstrating a more accurate identification of positive cases by the model. Precisions on negative, positive and average reviews for all dictionaries are shown in Table-15.

Table 15

Precision Based on Negative Arabic Reviews

Dictionaries	Precision on Negative Reviews	Precision on Positive Reviews	Average Precision
AFINN	88%	77%	83%
Textblob	82%	77%	80%
Vedar	88%	78%	83%
Patter.en	48%	95%	83%

Fig.-6 illustrates that AFINN and VADER exhibit high precision for negative reviews, whereas pattern.en excels in precision for positive reviews. When considering average precision across both positive and negative sentiments, AFINN, VADER, and pattern.en consistently demonstrate the highest levels.

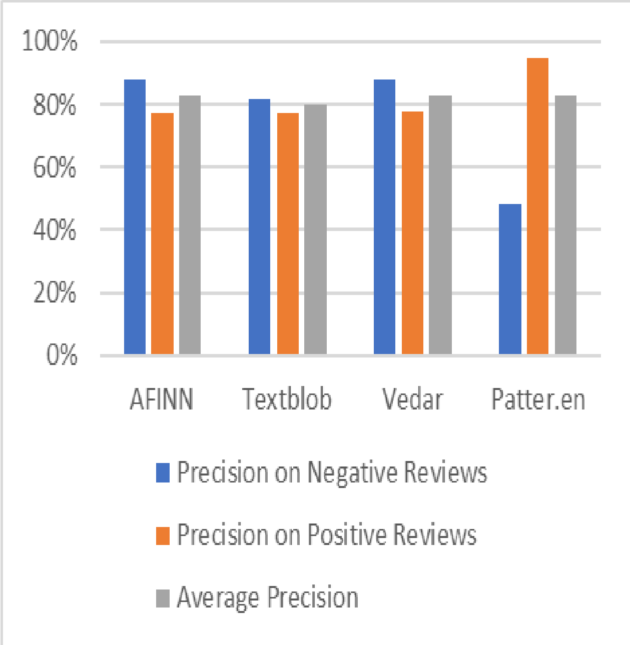

Fig. 6. Visual Representation of Precision on All Dictionaries

5.2 Analysis of Recall for all Dictionaries

In a confusion matrix, recall (also known as sensitivity or true positive rate) is a metric that gauges the ability of a classification model to correctly identify all relevant instances. It is calculated as the ratio of true positive predictions to the sum of true positives and false negatives. In essence, recall helps measure the model's capability to capture all instances of a particular class. A higher recall indicates that the model is proficient at identifying the relevant instances of the positive class, minimizing the number of false negatives. Recall on negative, positive and average Arabic reviews for all dictionaries are shown in Table-16.

Table 16

Recall Based on Negative Arabic Reviews

Dictionaries	Recall on Negative Reviews	Recall on Positive Reviews	Average Recall
AFINN	80%	86%	83%
Textblob	79%	81%	80%
Vedar	80%	86%	83%
Patter.en	92%	64%	71%

In Fig. 7, it is evident that pattern.en showcases a noteworthy level of recall for negative reviews. Conversely, AFINN and VADER demonstrate commendable recall performance for positive reviews, and this observation extends to the overall average score across sentiments. The figure effectively highlights the distinctive recall strengths exhibited by each model, emphasizing their ability to accurately capture instances of both negative and positive sentiments.

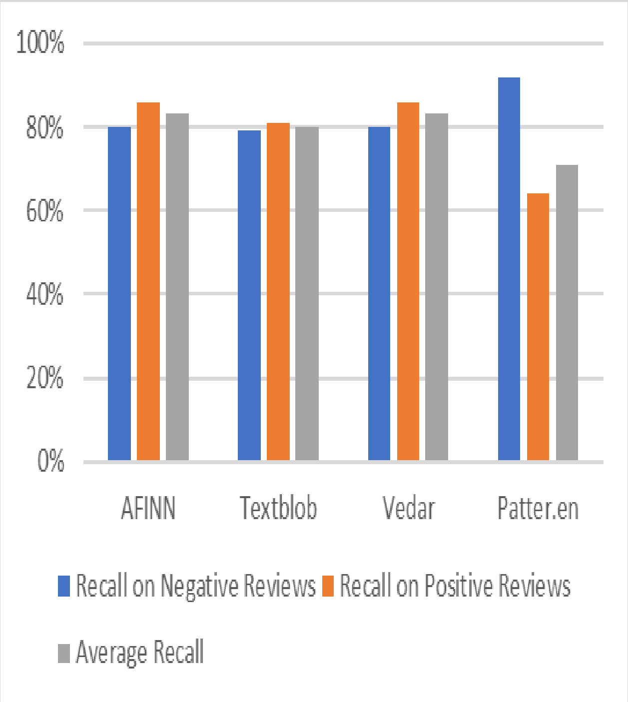

Fig. 7. Visual Representation of Recall on All Dictionaries

5.3 Analysis of F1-Score for all Dictionaries

The F1-score, also known as the F1 measure or F1-value, is a metric derived from the combination of precision and recall in a confusion matrix. It provides a balanced assessment of a classification model's performance by considering both false positives and false negatives. It is the harmonic mean of precision and recall, offering a single metric that balances the trade-off between precision and recall. A higher F1-score indicates a model that achieves a good balance between precision and recall, demonstrating robust performance in both positive and negative predictions.

F1-Score on negative, positive and average Arabic reviews for all dictionaries are shown in Table-17.

Table 17

Recall Based on Negative Arabic Reviews

Dictionaries	F1-Score on Negative Reviews	F1-Score on Positive Reviews	Average F1-Score
AFINN	84%	82%	83%
Textblob	81%	79%	80%
Vedar	84%	82%	83%
Patter.en	63%	77%	73%

As depicted in Fig. 8, AFINN and VADER distinctly demonstrate a commendable F1-Score across negative, positive, and average Arabic reviews.

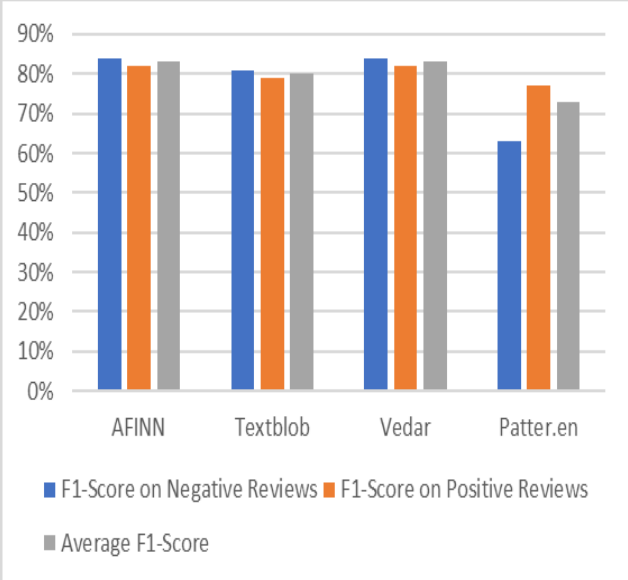

Fig. 8. Visual Representation of F1-Score on All Dictionaries

5.4 Analysis of Accuracy for all Dictionaries

Accuracy in a confusion matrix is a metric that assesses the overall correctness of a classification model by considering both true positives and true negatives. It is calculated as the ratio of the sum of true positives and true negatives to the total number of instances. Accuracy provides a general measure of how well a model performs across all classes. However, it might not be the most suitable metric in situations where the classes are imbalanced. In such cases, additional metrics like precision, recall, or F1-score might offer a more comprehensive evaluation of the model's performance.

Table 18

Recall Based on Negative Arabic Reviews

Dictionaries	Accuracy
AFINN	83%
Textblob	79%
Vedar	83%
Patter.en	71%

As depicted in Fig. 9, it is evident that both AFINN and VADER exhibit identical and the highest levels of accuracy, while pattern.en demonstrates the lowest accuracy on Arabic reviews. This observation suggests that AFINN and VADER present robust and comparable performance, making them potentially straightforward choices for implementation in research endeavors following this thorough analysis.

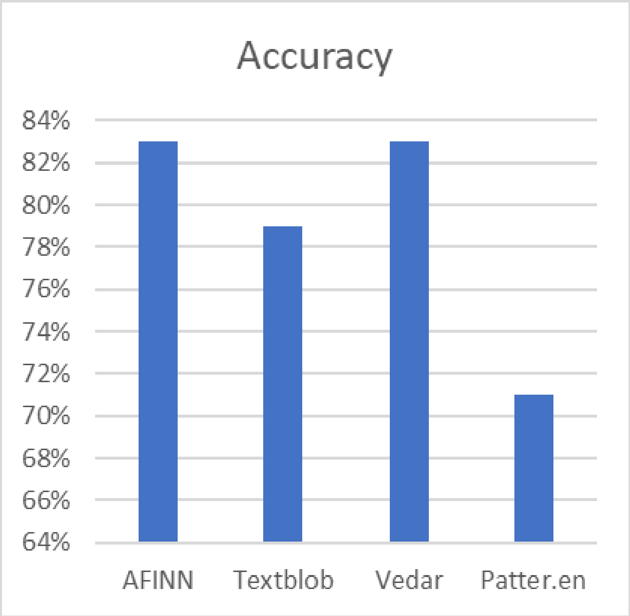

Fig. 9. Visual Representation of Accuracy on All Dictionaries

5.5 ROC Analysis for All Models

The evaluation of machine learning models for sentiment analysis requires robust performance metrics to ensure accuracy and reliability. Among these, the Receiver Operating Characteristic (ROC) curve is a powerful visualization tool that illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate. One key advantage of the ROC curve is that it provides a comprehensive view of a model's performance across different threshold values, making it highly useful in scenarios where class distributions are imbalanced. The area under the ROC curve (AUC-ROC) quantifies this performance, with higher values indicating superior discriminatory ability.

In Fig. 10, the ROC curves for Affin, TextBlob, Vader, and Pattern.en sentiment analysis models are presented, showing respective AUC values of 83%, 80%, 83%, and 72%. These values suggest that Affin and Vader exhibit the highest performance, both achieving an AUC of 83%, followed closely by TextBlob at 80%. In contrast, Pattern.en has the lowest performance with an AUC of 72%, indicating weaker classification capabilities.

Given these results, Affin and Vader are the most reliable choices for sentiment classification in this particular dataset. While TextBlob remains a viable option with competitive performance, Pattern.en appears to be the least effective, as it shows a lower ability to distinguish between sentiment classes. Thus, for applications that require precise sentiment analysis, Affin or Vader should be preferred due to their superior AUC scores and enhanced classification performance.

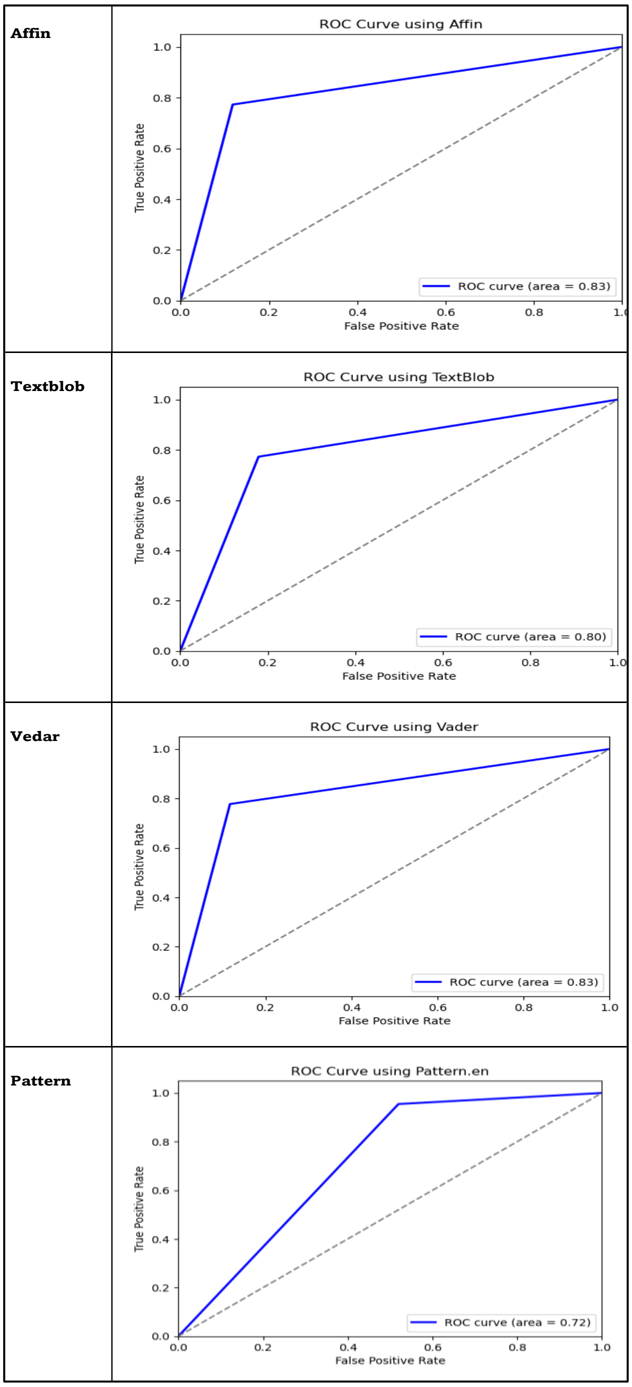

Fig. 10. ROC-Curve for all Models

5.6 Model Evaluation Using Statistical Scores

When comparing the four models (Affin, TextBlob, VADER, and Pattern.en) from Table-19, Table-20, Table-21 and Table-22, we can observe notable differences in their performance metrics, which give insights into their predictive accuracy and ability to explain variance in the target variable.

During assessment VADER proves better than other models in all evaluation criteria. The predictions of VADER match actual values more accurately with the smallest MSE value at 0.1693. Under the metrics evaluation VADER obtained the maximum R-Squared score at 0.3227 while reaching an Explained Variance Score of 0.3323 to explain the highest amount of target variable variance. The results indicate VADER's effective competence in analyzing underlying patterns present in the dataset. The implementation of VADER achieves the best Matthews Correlation Coefficient (MCC) at 0.6639 because it produces the most optimal binary classification with reduced errors.

Affin delivers predictive accuracy at a level similar to VADER by achieving an MSE value of 0.1715 but surpasses VADER by a minimal margin. The 0.3138 R^2 plays well in these tests as it means the model successfully predicts around 31% of the underlying data variability. The classification performance of Affin ranks highly because its MCC value stands at 0.6597 yet it trails behind VADER in accuracy measures. The 0.3243 value of EVS places the model in the group with higher performance capabilities as it shows how well the model predicts the target variable.

TextBlob shows lower predictive accuracy compared to VADER and Affin when applied to a text. The predictive errors of TextBlob indicate lower accuracy because it yields MSE and MAE at 0.2027. The 0.1890 R^2 value of this model ranks it as the second lowest among all models for explaining data variations. The EVS of 0.1906 indicates that TextBlob has a less effective ability to detect underlying data patterns. TextBlob produces more misclassifications than VADER and Affin since its MCC score stands at 0.5946 despite being positive.

Pattern.en shows the most inaccurate predictions based on the MSE of 0.2873 which stands as the maximum value between the four models. Pattern.en generates the least accurate predictions in comparison to the competing models. The negative R^2 score value of -0.1497 demonstrates that the model functions worse than making predictions based on the mean target variable value. Data variability that the model should explain remains unaccounted for by its EVS value of 0.0861. Within the group of four models Pattern.en demonstrates the least capability for classification work since its MCC score stands at 0.4913.

Finally VADER stands as the most dependable model that demonstrates both superior prediction accuracy and better variance explanation power when compared to Affin. Prediction and classification accuracy stands lower for TextBlob compared to Pattern.en which experiences performance challenges in all tasks. This evaluation demonstrates how the different models handle both their strongest and weakest points to help understand which model works best for sentiment and emotion analysis work.

Table 19

Statistical Measures using Affin

Metric	Value
Mean Squared Error (MSE)	0.1715
R-Squared (R^2 Score)	0.3138
Mean Absolute Error (MAE)	0.1715
Explained Variance Score (EVS)	0.3243
Matthews Correlation Coefficient (MCC)	0.6597

Table 20: Statistical Measures using Textblob

Metric	Value
Mean Squared Error (MSE)	0.2027
R-Squared (R^2 Score)	0.1890
Mean Absolute Error (MAE)	0.2027
Explained Variance Score (EVS)	0.1906
Matthews Correlation Coefficient (MCC)	0.5946

Table 21: Statistical Measures using Vader

Metric	Value
Mean Squared Error (MSE)	0.1693
R-Squared (R^2 Score)	0.3227
Mean Absolute Error (MAE)	0.1693
Explained Variance Score (EVS)	0.3323
Matthews Correlation Coefficient (MCC)	0.6639

Table 22: Statistical Measures using Pattern.en

Metric	Value
Mean Squared Error (MSE)	0.2873
R-Squared (R^2 Score)	-0.1497
Mean Absolute Error (MAE)	0.2873
Explained Variance Score (EVS)	0.0861
Matthews Correlation Coefficient (MCC)	0.4913

6. Comparison with Similar Studies

The proposed work in this study focuses on enhancing sentiment and emotion analysis for Arabic text by integrating widely-used sentiment analysis tools like AFINN, TextBlob, VADER, and pattern.en, which traditionally faced compatibility issues with Arabic text. By addressing these challenges, the study finds that AFINN and VADER are the most effective tools for Arabic sentiment analysis, achieving high performance metrics (Precision: 83, Recall: 83, F1-Score: 83, Accuracy: 83 for AFINN and Precision: 84, Recall: 83, F1-Score: 83, Accuracy: 83 for VADER). This approach is computationally efficient and accessible, avoiding the need for complex machine learning or deep learning models, making it more suitable for real-world applications where resource constraints are a concern.

In contrast, the work in [49] explores the use of deep learning models, specifically CNN and LSTM, for sentiment analysis. While the bidirectional deep learning model in [49] achieves high accuracy (99%

on training data and 78% on test data), it requires substantial processing power and memory, making it less practical for resource-constrained environments. Moreover, the performance of these models depends on fine-tuning hyperparameters like epochs and batch size. While deep learning models can provide high accuracy, the proposed work offers a more practical and efficient solution for Arabic sentiment analysis, with minimal computational requirements and strong results using established tools. Therefore, the proposed work is better suited for broader, more accessible sentiment analysis applications.

7. Conclusion

The study concludes by addressing the critical role of sentiment analysis in the Arabic language, which presently ranks fourth in internet popularity. Despite its significant online presence, there exists a noteworthy scarcity of Natural Language Processing (NLP) applications tailored specifically for Arabic. In response to this gap, the research introduces an innovative methodology, leveraging Googletrans to adapt established NLP dictionaries originally designed for English—such as Affin, Textblob, Pattern, and Vedar—to facilitate effective sentiment analysis in Arabic text.

The research in arear of Arabic linguistic uniqueness and intricate writing and dialectical variations together with stemming challenges shows how difficult sentiment analysis becomes. AFINN and VADER proved themselves to be effective sentiment analysis tools for Arabic text through their highly precise classification of both positive and negative reviews.

The research demonstrates how Googletrans functions as a bridge that connects state-of-the-art automatic dictionaries with Arabic linguistic content. The practical framework presented functions as a research guide for scientists conducting future studies about sentiment analysis implementation.

In the specific evaluation of dictionaries, AFINN and VADER consistently exhibit robust performance, demonstrating comparable levels of accuracy and precision. This suggests that these dictionaries, coupled with the introduced innovative approach, hold promise for further research and practical application in sentiment analysis for the Arabic language. The comprehensive nature of this research contributes to the advancement of NLP applications tailored for Arabic, acknowledging the language's unique characteristics and challenges in the digital landscape.

8. Limitations and Future Work

Despite the commendable results obtained in this study, it is essential to acknowledge certain limitations that may guide future investigations. First and foremost, this research is limited to reviews written exclusively in Arabic, which, while valuable, restricts the generalizability of the findings to other languages. Future studies could explore expanding the scope to include reviews in other languages, particularly those with rich linguistic diversity or similar challenges to Arabic, to evaluate the effectiveness of the proposed approach across a broader spectrum.

Additionally, this study utilized Googletrans for translation from Arabic to English, which proved effective in bridging the gap between Arabic and English-based sentiment analysis dictionaries. However, there are several other translation packages and techniques that could be explored to potentially enhance translation accuracy and better capture the nuances of Arabic text. Alternative tools, such as DeepL, Microsoft Translator, or domain-specific models, could offer improvements in translation quality, leading to better alignment with the sentiment analysis tools.

Furthermore, the study relied on four specific sentiment analysis dictionaries—AFINN, TextBlob, VADER, and Pattern.en—which were adapted for Arabic text. While these tools were suitable for the scope of this research, future work could incorporate additional or alternative sentiment analysis dictionaries, particularly those designed for non-English languages, to compare performance and explore whether other tools yield improved results for Arabic sentiment analysis.

Lastly, the research used a dataset of 450 reviews, which, although sufficient for the study, is relatively small in the context of modern NLP research. Expanding the dataset to include a larger number of reviews or incorporating data from different domains (e.g., social media posts, news articles) could provide a more comprehensive understanding of sentiment analysis performance and help validate the robustness of the findings. The inclusion of a wider variety of text sources could also help assess the adaptability of the proposed methodology to various types of language and writing styles, further improving the model's generalization ability.

9. References

[1] S. Tiwari, M. C. Trivedi, M. L. Kolhe, K. K. Mishra, and B. K. Singh, “Advances in Data and Information Sciences”, Proceedings of ICDIS 2021, 2022, vol. 2, pp. 293–293, [Online].

Available: http://dx.doi.org/10.1007/978-981-10-8360-0_8%0Ahttp://link.springer.com/10.1007/978-981-10-8360-0.

[2] L. Rao, "Sentiment Analysis of English Text with Multilevel Features", Sci. Program., vol. 2022, 2022, doi: 10.1155/2022/7605125.

[3] M. R. R. Rana, A. Nawaz, T. Ali, A. M. El-Sherbeeny, and W. Ali, "A BiLSTM-CF and BiGRU-based Deep Sentiment Analysis Model to Explore Customer Reviews for Effective Recommendations", Eng. Technol. Appl. Sci. Res., vol. 13, no. 5, pp. 11739–11746, 2023, doi: 10.48084/etasr.6278.

[4] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews", Eng. Technol. Appl. Sci. Res., vol. 13, no. 3, pp. 10849–10855, 2023, doi: 10.48084/etasr.5854.

[5] A. E. A. Hamouda and F. E. El-taher, "Sentiment Analyzer for Arabic Comments System", Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 3, pp. 99–103, 2013, doi: 10.1055/s-2005-870135.

[6] R. Elhassan and M. Ahmed, "Arabic text classification: review", Int. J. Comput. Sci. Softw. Eng., vol. 4, no. 1, pp. 1–5, 2015.

[7] S. Al-Osaimi and M. Badruddin, "Sentiment Analysis Challenges of Informal Arabic Language", Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 2, 2017, doi: 10.14569/IJACSA.2017.080237.

[8] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media", Comput. Speech Lang., vol. 28, no. 1, pp. 20–37, 2014, doi: 10.1016/j.csl.2013.03.001.

[9] G. Alwakid, T. Osman, and T. Hughes-Roberts, "Challenges in Sentiment Analysis for Arabic Social Networks", Procedia Comput. Sci., vol. 117, pp. 89–100, 2017, doi: 10.1016/j.procs.2017.10.097.

[10] D. A. Musleh et al., "Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation", Big Data Cogn. Comput., vol. 7, no. 3, 2023, doi: 10.3390/bdcc7030127.

[11] D. Conti, C. E. Gomez, J. G. Jaramillo, and V. E. Ospina, "Monitoring the Quality and Perception of Service in Colombian Public Service Companies with Twitter and Descriptive Temporal Analysis", Appl. Sci., vol. 13, no. 18, 2023, doi: 10.3390/app131810338.

[12] O. Abiola, A. Abayomi-Alli, O. A. Tale, S. Misra, and O. Abayomi-Alli, "Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser", J. Electr. Syst. Inf. Technol., vol. 10, no. 1, 2023, doi: 10.1186/s43067-023-00070-9.

[13] M. J. Domínguez Vázquez and R. H. Gouws, "The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography", Int. J. Lexicogr., vol. 36, no. 3, pp. 233–259, 2023, doi: 10.1093/ijl/ecac020.

[14] N. Thakur, "Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox", Big Data Cogn. Comput., vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020116.

[15] S. M. Saqib, M. Z. Asghar, A. Al-rasheed, M. A. Khan, and Y. Ghadi, "DenseHillNet : a lightweight CNN for accurate classification of natural images", PeerJ Comput. Sci., pp. 1–21, 2024, doi: 10.7717/peerj-cs.1995.

[16] S. M. Saqib et al., "Cataract and glaucoma detection based on Transfer Learning using MobileNet", Heliyon, vol. 10, no. 17, 2024, doi: 10.1016/j.heliyon.2024.e36759.

[17] S. Muhammad Saqib et al., "MobVGG: Ensemble technique for birds and drones prediction", Heliyon, vol. 10, no. 21, 2024, doi: 10.1016/j.heliyon.2024.e39537.

[18] A. O. Shakeel Ahmad, Sheikh Muhammad Saqib, "LSI Based Search Technique: Using Extracted Keywords and Key Sentences", VAWKUM Trans. Comput. Sci., vol. 14, no. 2, pp. 1–8, 2017.

[19] S. M. Saqib et al., "Integrating YOLO and WordNet for automated image object summarization", Signal, Image Video Process., 2024, doi: 10.1007/s11760-024-03560-z.

[20] S. Muhammad Saqib, S. Ahmad, A. Hassan Syed, T. Naeem, and F. Mazaed Alotaibi, "Grouping of Aspects into Relevant Category Based on WordNet Definitions", IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 19, no. 2, p. 113, 2019.

[21] E. Atwell, C. Brierley, K. Dukes, M. Sawalha, and A. Sharaf, "An Artificial Intelligence

approach to Arabic and Islamic content on the internet", NITS 3rd Natl. Inf. Technol. Symp., pp. 1–13, 2011, doi: 10.13140/2.1.2425.9528.

[22] S. Lagrini and M. Redjimi, "Automatic Arabic Text Summarization Approaches", Int. J. Comput. Appl., vol. 164, no. 5, pp. 31–37, 2017, doi: 10.5120/ijca2017913628.

[23] H. N. Fejer and N. Omar, "Automatic multi-document Arabic text summarization using clustering and keyphrase extraction", J. Artif. Intell., vol. 8, no. 1, pp. 1–9, 2015, doi: 10.3923/jai.2015.1.9.

[24] R. Belkebir and A. Guessoum, "A supervised approach to arabic text summarization using adaboost", Adv. Intell. Syst. Comput., vol. 353, pp. 227–236, 2015, doi: 10.1007/978-3-319-16486-1_23.

[25] S. A. Sheikh Muhammad Saqib, Fazal Masud Kundi, "Automatic classification of product reviews into interrogative and noninterrogative: Generating real time answer", Int. J. Adv. Appl. Sci., vol. 6, no. 8, pp. 23–31, 2019, doi: 10.21833/ijaas.2019.08.004.

[26] S. M. Saqib, F. M. Kundi, and S. Ahmad, "Unsupervised Learning Method for Sorting Positive and Negative Reviews Using LSI (Latent Semantic Indexing) with Automatic Generated Queries", IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 1, pp. 56–62, 2018.

[27] D. Elangovan and V. Subedha, "Adaptive Particle Grey Wolf Optimizer with Deep Learning-based Sentiment Analysis on Online Product Reviews", Eng. Technol. Appl. Sci. Res., vol. 13, no. 3, pp. 10989–10993, 2023, doi: 10.48084/etasr.5787.

[28] H. J. Alyamani, S. Ahmad, A. H. Syed, S. M. Saqib, and Y. D. Al-Otaibi, "LSI Authentication-Based Arabic to English Text Converter", Int. J. Fuzzy Log. Intell. Syst., vol. 21, no. 4, pp. 409–422, 2021, doi: 10.5391/IJFIS.2021.21.4.409.

[29] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[30] S. Muhammad and F. Masud, "MMO: Multiply-Minus-One Rule for Detecting & Ranking Positive and Negative Opinion", Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 122–127, 2016, doi: 10.14569/IJACSA.2016.070519.

[31] S. M. Saqib and F. M. Kundi, "Semi Supervised Method for Detection of Ambiguous Word and Creation of Sense : Using WordNet", Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 11, pp. 353–359, 2018.

[32] G. Badaro, R. Baly, and H. Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining", Arab. Nat. Lang. Process. Work. co-located with EMNLP 2014, pp. 165–173, 2014.

[33] A. Masri and M. Al-Jabi, "A novel approach for Arabic business email classification based on deep learning machines", PeerJ Comput. Sci., vol. 9, no. 2017, p. e1221, 2023, doi: 10.7717/peerj-cs.1221.

[34] T. Alqurashi, "Arabic Sentiment Analysis for Twitter Data: A Systematic Literature Review", Eng. Technol. Appl. Sci. Res., vol. 13, no. 2, pp. 10292–10300, 2023, doi: 10.48084/etasr.5662.

[35] H. Hastang and R. Ahmad, "Analysis of Arabic language learning difficulties among students in the Qawaid Al-Lughah Al- Arabiyyah materials", Didakt. J. Kependidikan, vol. 17, no. 1, pp. 31–42, 2023.

[36] H. Almaqtari, F. Zeng, and A. Mohammed, "Enhancing Arabic Sentiment Analysis of Consumer Reviews: Machine Learning and Deep Learning Methods Based on NLP", Algorithms, vol. 17, no. 11, 2024, doi: 10.3390/a17110495.

[37] A. Alharbi, N. Sharma, and F. Hussain, "Arabic Sentiment Analysis with Social Network Data: A Comparative Study", Proc. - 2024 IEEE Int. Conf. E-bus. Eng. ICEBE 2024, pp. 87–94, 2024, doi: 10.1109/ICEBE62490.2024.00022.

[38] O. Alsemaree, A. S. Alam, S. S. Gill, and S. Uhlig, "Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions", Heliyon, vol. 10, no. 9, 2024, doi: 10.1016/j.heliyon.2024.e27863.

[39] M. J. Al-khazaleh, M. Alian, and M. A. Jaradat, "Sentiment analysis of imbalanced Arabic data using sampling techniques and classification algorithms", Bull. Electr. Eng. Informatics, vol. 13, no. 1, pp. 607–618, 2024, doi: 10.11591/eei.v13i1.5886.

[40] A. Munshi et al., "Arabic Sentiment Analysis for Student Evaluation using Machine Learning and the AraBERT Transformer", Eng. Technol. Appl. Sci. Res., vol. 13, no. 5, pp. 11945–11952, 2023, doi: 10.48084/etasr.6347.

[41] W. M. S. Yafooz, E. A. Hezzam, and W. Alromema, “Arabic Sentiment Analysis on Chewing Khat Leaves using Machine Learning and Ensemble Methods”, Eng. Technol. Appl. Sci. Res., vol. 11, no. 2, pp. 6845–6848, 2021, doi: 10.48084/etasr.4026.

[42] “https://pypi.org/project/googletrans/.” .

[43] “https://opendatacommons.org/licenses/odbl/1-0/.” .

[44] “http://corpustext.com/reference/sentiment_afinn.html.” .

[45] “https://textblob.readthedocs.io/en/dev/.” .

[46] “https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.” .

[47] “https://digiasset.org/html/pattern-en.html#article.” .

[48] “https://www.kaggle.com/datasets/abanoubsamir004/arabic-reviews-sentiment-analysis?resource=download.” .

[49] H. J. Alyamani, “Sentiment Orientation Using Deep Learning Sequential and Bidirectional Models”, Int. J. Comput. Sci. Netw. Secur., vol. 11, no. 21, 2021.