Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.3130

2024, 43 (2) 183-194

CNN and LSTM based hybrid deep learning model for sentiment analysis on Arabic text reviews

Shakeel Ahmad [a], Sheikh Muhammad Saqib [b, *], Asif Hassan Syed [a]

[a] *Department of Computer Science, Faculty 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: Sheikh Muhammad Saqib, Email: saqibsheikh4@gu.edu.pk

Received: 12 December 2023, Accepted: 28 March 2024, Published: 01 April 2024

K E Y W O R D S

Sentiment Analysis

Deep Learning

Long Short-Term Memory (LSTM)

Convolutional Neural Network (CNN)

Globalmaxpooling

A B S T R A C T

Companies with diverse product offerings rely on customer reviews to gauge product reception. Following a purchase, customers often share their opinions on the website. Prospective buyers, prior to deciding, typically peruse these reviews to inform their choices. Analysing such feedback, whether positive or negative, holds paramount importance for companies seeking to improve product quality. Researchers are actively exploring methods to categorize comments based on sentiment scores. Notably, customers may express their reviews in Arabic text. Despite challenges such as the structure and morphology of Arabic text, a scarcity of machine-readable Arabic dictionaries, and limited tools for handling Arabic text, minimal progress has been made in the analysis of Arabic reviews. While some attempts have been undertaken, they have achieved suboptimal accuracy. In response, the authors propose a hybrid deep learning model comprising Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) with GlobalMaxPooling. Through multiple iterations, the authors fine-tuned the proposed model and applied it to publicly available Arabic Reviews dataset, achieving a notable 95% accuracy, precision, recall, and F1 score. The results indicate that, when compared to alternative models, the proposed model exhibits superior accuracy.

1. Introduction

The widespread use of the Internet today, especially through smartphones, has completely changed many facets of our lives. People now use this digital medium for a variety of tasks, including making hotel reservations, purchasing tickets, and gathering crucial data before making important travel arrangements. As people use the opinions and experiences of others to make informed decisions, online reviews are crucial in forming consumer decisions. The importance of these reviews is especially clear in terms of quality, where up to 80% of customers 'purchasing decisions can be significantly influenced by online negative sentiments [1][2].

Given the wide range of reviews available, how can a business improve the calibre of its goods, locations, and other offerings? Sentiment analysis, a field devoted to identifying the sentiment conveyed in reviews or comments, has been the subject of numerous studies [1][3]. In this context, sentiment orientation refers to the precise categorization of opinions as clearly positive or negative [4]. The term "sentiment" is used to describe a person's sentiment, which includes their point of view, assessment, or emotional reaction to [5], feature [6], or service [7]. A

scoring system [8][9] can be used to quantify this sentiment, whether it is positive or negative.

There is a noticeable gap in the literature regarding the analysis of sentiments in Arabic, even though most sentiment analysis research has focused on English-written reviews and comments. Further development is required, particularly in improving deep learning models designed for Arabic reviews, even though some recent studies have delved into sentiment analysis of Arab reviews. Notably, the work presented in [10] has created a model for Arabic reviews using semantic orientation approaches. However, 79.20% and 78.75%, respectively, are reported as the overall accuracy and the f-measure. A support vector machine was used in another study [11] that focused on comparable reviews to determine polarity, and the results showed an 84% accuracy.

The linear Support Vector Classifier(Linear SVC) model was used to achieve the ideal F1 measure, which reached 91.70%, while the supported vector classifier's employed model produced a commendable result of 91.41% [12]. It's important to note that these achievements were made using a machine learning methodology, and that deep learning models have the potential to be explored.

Convolutional neural networks (CNN) and bi-directional long-term memory (BiLSTM) deep learning algorithms were used to classify the sentiment of Arabic tweets. According to the experiment's findings, BiLSTM performed with performance accuracy of 91.99%, while CNN achieved 92.80% accuracy [13]. Notably, combining CNN and BiLSTM (Bi-Directional Long Short-Term Memory) has the potential to be improved even more.

The authors of this study have started creating a hybrid deep learning model in response to the identified research gap in the field of sentiment analysis for Arabic reviews. To optimize computation for dense layers, this model uses Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks in addition to GlobalMaxPooling1D. The proposed work's main contributions can be summed up as follows:

- How is the Hybrid Model created to improve sentiment analysis for Arabic reviews by incorporating CNN, LSTM, and GlobalMaxPooling1D?
- What metrics are used to assess the proposed hybrid model's effectiveness in accurately categorizing sentiment in Arabic reviews?
- What standards are used to evaluate the proposed hybrid model's superiority or find areas where it could be improved, and how is it compared to accepted benchmarks?

The authors advance and improve computational models that are suited to the touches of the Arabic language by addressing these key aspects, adding valuable insights and methodologies for the field of sentiment analysis for Arabic reviews.

The rest of this paper are organized as follows: Section-2 contains literature review, Section-3 dedicates to describe the proposed method, Model Implementation is done in Section-4, Results and discussion are interpreted in Section-5, Comparison with benchmarks is describe in Section-6 and finally this paper is concluded in Section-7.

2. Literature Review

Numerous studies are dedicated to computing-based classification, particularly within the domain of Sentiment Analysis on different language such as Urdu [14], Chinese [15], Arabic [16], etc. Many works focus on word polarity as a basis for sentiment analysis [1][3][17]. Additionally, research has delved into sentiment orientation, product aspects, individual emotions, and the identification of emoticons [4][5]. Beyond traditional machine learning approaches, sentiment analysis has witnessed substantial exploration through various dimensions of deep learning [18][19][20][21]. In [22], the utilization of word2vec is notable for reducing the number of parameters, adopting a bag-of-words approach within deep learning. Authors in [23] scrutinized the impact on performance through multiple runs by altering hyperparameters in convolutional neural networks. The OpCNN model, incorporating k-max pooling, was introduced in [24] to address the word order problem in Chinese. Furthermore, sentiment classification on tweets, discerning tweets as positive or negative, was accomplished through the implementation of LSTM neural networks in [25] [26].

Leveraging the internet, a research study [27] explored Arabic and Islamic content, specifically focusing on information extracted from prophetic narrations, utilizing the Artificial Intelligence (AI) approach. The authors introduced a semantically driven method for analysing Arabic discourse within the framework of Segmented Discourse Representation Theory (SDRT). It was discovered that discourse analysis can effectively generate indicative summaries of Arabic documents [28] In another approach, the authors of [29] proposed a novel summarization model based on document clustering and key phrase extraction clusters. Additionally, [30] introduced the "Adaboost" approach, a supervised

method for Arabic summary extraction. This method incorporates statistical features such as sentence position, the number of keywords in a sentence, overlap with the word title, and sentence length.

The task of Arabic text summarization remains relatively novel, primarily due to the inherent complexity of the Arabic language. Evaluating Arabic summarization faces challenges, particularly as there are no gold standard summaries available in machine-readable dictionaries [31].

While numerous Text Classification (TC) research studies have been conducted and tested in languages such as English, French, German, Spanish, Chinese, Greek, and Japanese [32]. opinion mining in Arabic texts regarding specific subjects [33] is an emerging area. However, current research on the automatic classification of Arabic text documents is limited. This limitation is attributed to various factors, including diverse spellings of certain words, distinct combinations of characters, the presence of short (diacritics) and long vowels, and the prevalence of affixes in most Arabic words [34].

NLP is applied in sentiment analysis (SA), which extracts opinions from text. Analysing sentiments in Arabic poses challenges due to ambiguity, dialects, and limited resources. Using convolutional neural networks (CNN) and hybrid models, this research achieved high accuracy in predicting Arabic sentiment using various deep learning models. The proposed approach, leveraging Arabert model features and combining FastText and GLOVE embeddings, outperformed standard deep learning methods with a margin of 0.9112 [35].

A hybrid approach [36] integrates aspect extraction, association rule mining, and Bidirectional Encoder Representations from Transformers (BERT) for improved sentiment analysis with accuracy 89%.

In the study by Mesleh [37], three machine-learning algorithms (SVM, KNN, and Naïve Bayes) were employed to classify Arabic data collected from Arabic newspapers. Additional studies [38] [39]utilized Text Classification and analyses of Arabic texts to automatically categorize texts into predefined categories based on linguistic features. Classifiers such as Naïve Bayes (NB), K-Nearest Neighbour (KNN), Multinomial Logistic Regression (MLR), and Maximum Weight (MW) were employed, with the NB classifier serving as the primary classifier and the others as auxiliary classifiers [40].

The prior efforts have primarily centred on machine learning, with additional attempts using deep learning showing less-than-optimal accuracy [10][11][12][13][41], as detailed in the introduction section. The principal contribution of the proposed work lies in the design of a deep learning model aimed at enhancing accuracy specifically for Arabic data.

The existing literature lacks sufficient exploration of sentiment analysis in Arabic, despite the predominant focus on English reviews in sentiment analysis research. There is a need for further advancements, particularly in refining deep learning models tailored for Arabic reviews. While recent studies have initiated sentiment analysis in Arab reviews, there is room for improvement. In conclusion, more research is warranted to enhance sentiment analysis models for Arabic content.

3. Hybrid Deep Learning Model for Arabic Dataset

Proposed model consists of converting the Arabic text reviews into matrix for machine readable form as shown in Fig. 1.

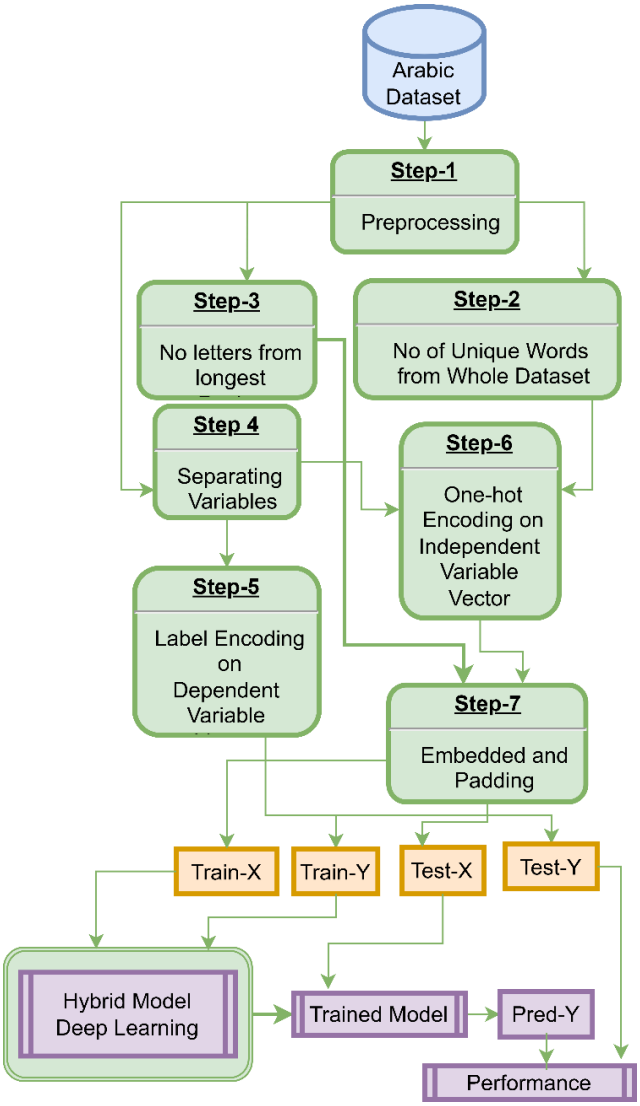


Fig. 1. Arabic Data Conversion into Matrix

Fig. 1. depicts the conversion of an Arabic dataset for model compatibility. In Step-1, outliers in dependent variables (from Step-4) are removed. Dependent variables are then turned into integers using LabelEncoding in Step-5, and independent variables undergo one-hot encoding in Step-6. Step-2

calculates the total word count as vocabulary, and Step-3 finds the maximum sentence length. These values guide Step-7 in converting the independent variable matrix into an embedded vector. The final outputs are Train-X, Train-Y, Test-X, and Test-Y. Train-X and Train-Y train the 'Deep Learning Hybrid Model.' The model's performance is evaluated by comparing predicted values from Test-X to Test-Y using a confusion matrix for accuracy assessment.

3.1 Model Construction

Fig. 2. illustrates the overall structure of the 'Hybrid Deep Learning Model.' It comprises input, hidden, and output layers. The input layer includes an embedding layer that receives an embedded vector. Each vector has a uniform maximum length determined by padding and a shape aligned with the independent variable vector. This layer has 100 neurons as computational units.

The hidden layers consist of a Convolutional layer and an LSTM layer. The Convolutional layer employs a filter of size 128 with a ReLU function suitable for a hidden layer. The LSTM layer is utilized to retain previous context using 64 cells. Subsequently, a Global Max Pooling layer is added for feature mapping to extract maximum values and reduce the data dimensionality.

To set weights and biases, the next dense layer is introduced, followed by a final dense layer with a sigmoid function to generate a single output representing the percentage of detection as positive or negative.

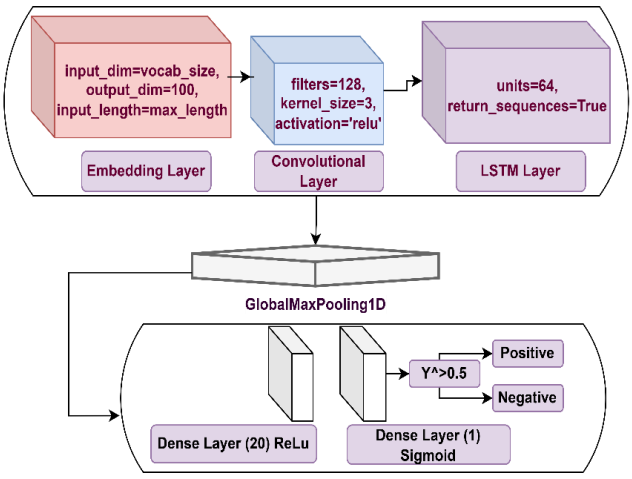


Fig. 2. Hybrid Deep Learning Model

3.2 Why Did We Choose the Sigmoid Function?

In a standard neural network architecture, the output typically consists of two numbers. A value exceeding 0.5 signifies a positive result, while a value below 0.5 indicates a negative outcome. For instance, if the network outputs [0.9550], it signifies a positive review because 0.95 surpasses the 0.5 threshold. The use of the sigmoid function allows the network to provide probabilistic responses to classification queries, offering more nuanced information than simple binary answers like yes or no.

3.3 Dataset Collection

The dataset comprises Arabic reviews labelled with name "Arabic_Reviews_Sentiment_analysis" containing data from many companies like talabat, kabiter, nasla, swifil, alsiwidiu, kilubatra, dumat, etc [42] organized into two columns: 'reviews' and 'class.' The 'reviews' column contains the actual Arabic reviews, while the 'class' column includes labels as either 'positive' or 'negative.' In total, there are 4000 reviews, evenly divided into 2000 positive reviews and 2000 negative reviews.

3.3.1 Train set

Approximately 60% of the training set data was utilized for training [43]. The training dataset encompasses both the dependent variable (result identifier) and the input factors (predictor variables). From the entire dataset, 2400 reviews were randomly selected for training, encompassing both 'positive' and 'negative' classes. Table 1 displays a sample of 10 reviews from this dataset, with 0 indicating negative and 1 indicating positive sentiment.

Table 1

Sample Data from Training Dataset

S. No	Reviews	Class
1	التطبيق لا يعمل ليش كان خليتوه زي ما كان	0
2	لماذا برنامج لا يعمل	0
3	اسواء بنك بالسعودية حسبي الله عليكم ونعم الوكيل	0
4	لا استطيع الدخول على التطبيق لماذا ؟	0
5	اسعار معقوله ومضمونه والقائمين عليها ممتازين	1
6	مشاء الله تحت التجربه	1
7	احلي	1
8	البرنامج دة ضعف جدن جدن مبيفتحش على الباقة علشان يفتح لذم شبكة واى فاى	1
9	واكثر من راااائع . . استمروووو 👑🏮	1
10	تطبيق فوق الممتاز يستحق التحميل رائع	1

3.3.2 Testing set

The proposed technique tackles efficiency concerns such as overfitting and underfitting by incorporating validation data into the model. For model evaluation, a 40% subset is designated for testing [44]. This testing dataset comprises 1600 reviews randomly selected from the entire dataset, encompassing both 'positive' and 'negative' classes. Table 2 presents a sample of 10 reviews from this testing data, with 0 denoting negative sentiment and 1 denoting positive sentiment.

Table 2

Sample Data from Test Data

S. No	Review	Class
1	دبستوني بطاقة اتمانية مابغاها وكل قرش اخذتوه مني ظلم اخذه منكم يوم القيامة	0
2	ربنا يبارك فيكو	1
3	فالاول كان البرنامج جيد ويفتح بسرعه وأما الان فلا يفتح	0
4	كيف يتم إنشاء حساب	1
5	جميل رائع يوسهل في توصيل الطعام انصح بي تجربته وانصح من الشركه بان تطوره للاجمل	1
6	بعد اخر تحديث التطبيق لايعمل مطلقا انا كل شغلي واقف بسبب عدم عمل التطبيق	0
7	تحديثكم الاخير هذا زفت ماعاد صار يدخلني للتطبيق..بسرعه حدثوه وصلحوا الخطأ .. بعد تحديثكم زفت الاخير حدثوه وصلحوا الخطأ ..مايدخلني للتطبيق اليوم ..	0
8	زفت برنامج لا يعمل ولا يتحدث ايش نقول نشكركم غصب	0
9	بنك لايقدر قيمة عملائه،والدليل انك تروح وماتحصل زحمة عند اي فرع لان الناس غسلو يدهم منه،وعطيتهم نجمة بس عشان التعليق	0
10	بعد التحديث لا يعمل بالمره نرجو حل المشكله وشكرا	0

3.4 Convert Dataset into Model Readable Form

Fig. 1. illustrates the process of converting the text dataset into a format readable by the model. In Step-1, each review is segmented before the removal of stop words and punctuation. Step-2 involves determining the vocabulary size, a crucial step for subsequent One-hot encoding. The vocabulary size represents the overall count of unique terms in the dataset. Step-3 utilizes long sentences from the entire dataset to compute the embedding vector. The obtained vocabulary size and the length of a long phrase are presented in Table 3.

Table 3

Total words and Maximum Length of Each Vector for Embedded Vector

Name	Values
Total Words	6937
Maximum Length of each Vector	104

Next, in Step-4, dependent and independent variables are identified for vectors. Step-5 utilizes label encoding, assigning 0 to negative values and 1 to positive values, to convert the dependent vector from text to integer values. With a specified vocabulary size of '6937' in Step-6, both Test and Training Data from Table 1 and Table 2 are transformed into integer values using one-hot encoding. The resulting integer vectors from the Training and Test data are detailed in Table 4 and Table 5, respectively. For a sentence with 50 words, 50 integer values correspond to each word. It is crucial to maintain the same length for all vectors when reading into the deep learning model. In Process-7, padding (from Step-3) is implemented by constructing an embedded vector with 104 columns, equivalent to the length of the longest sentence. A sample of Training and Test data with the first five and last five columns is provided in Table 4 and Table 5, where columns 0-103 signify a vector length of 104.

Table 4

Embedded Vector on Training Data

	0	1	2	3	4	99	100	101	102	103
1	1038	6	25	96	26	0	0	0	0	0
2	137	17	6	25	0	0	0	0	0	0
3	538	19	2282	341	20	0	0	0	0	0
4	6	103	8	13	2	0	0	0	0	0
5	2283	2284	2285	1039	821	0	0	0	0	0
6	822	20	468	823	0	0	0	0	0	0
7	824	0	0	0	0	0	0	0	0	0
8	10	1406	2287	825	825	0	0	0	0	0
9	370	1	2291	2292	2293	0	0	0	0	0
10	7	263	312	255	411	0	0	0	0	0

Table 5

Embedded Vector on Test Data

	0	1	2	3	4	99	100	101	102	103
1	362	269	324	271	100	0	0	0	0	0
2	775	1106	0	0	0	0	0	0	0	0
3	26	10	107	1661	248	0	0	0	0	0
4	243	104	230	0	0	0	0	0	0	0
5	18	35	4	597	535	0	0	0	0	0
6	9	202	24	2	66	0	0	0	0	0
7	1929	132	43	184	990	0	0	0	0	0
8	184	17	6	25	14	0	0	0	0	0
9	19	3274	819	690	138	0	0	0	0	0
10	9	5	6	25	2220	0	0	0	0	0

4. Model Implementation

The implemented model is summarized in Table 6, detailing the parameter counts for each layer. The

input layer comprises 693,700 parameters, hidden layer-1 has 38,528 parameters, hidden layer-2 contains 49,408 parameters, and hidden layer-3 has 4,160 parameters. The output layer is composed of 65 parameters.

Table 6

Summary of Implemented Model

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 104, 100)	693700
conv1d (Conv1D)	(None, 102, 128)	38528
lstm (LSTM)	(None, 102, 64)	49408
global_max_pooling1d	(Global (None, 64)	0
dense (Dense)	(None, 64)	4160
dense_1 (Dense)	(None, 1)	65
Total params: 785,861		
Trainable params: 785,861		
Non-trainable params: 0		

4.1 Model Training

The model was trained on the provided training data sample from Table 4 using the Adam optimizer for 10 epochs. Details of the training mechanism are illustrated in Fig.3.

Table 7

Training Mechanism

Epochs	Time	Train Loss	Train Accuracy
Epoch-1	100ms/step	0.4682	0.7908
Epoch-2	84ms/step	0.0987	0.9633
Epoch-3	88ms/step	0. 0213	0.9942
Epoch-4	84ms/step	0. 0080	0.9971
Epoch-5	84ms/step	0. 0021	0.9996
Epoch-6	86ms/step	0.0031	0.9996
Epoch-7	86ms/step	4.5846e-04	1.0000
Epoch-8	85ms/step	2.0261e-04	1.0000
Epoch-9	89ms/step	1.1760e-04	1.0000
Epoch-10	92ms/step	8.8824e-05	1.0000

4.2 Testing of Proposed Model

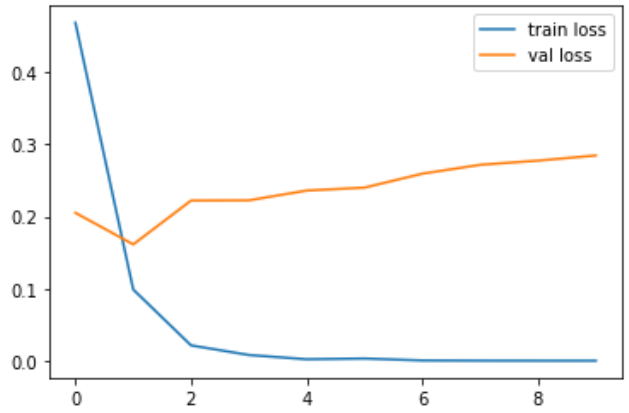
The test data from Table 5 was utilized on the trained model, and the outcomes are presented in Table 7. Instances where the "Actual Class" and "Predicted Decisions" columns share the same value indicate accurate predictions. The model produces predicted values, categorized as 1 (Positive) if they surpass 0.5, otherwise, as 0 (Negative). Table 8 exclusively displays results for a sample from the Test Data in Table 2, a comprehensive description of all 1600 review instances will be elucidated in the confusion matrix below.

Table 8

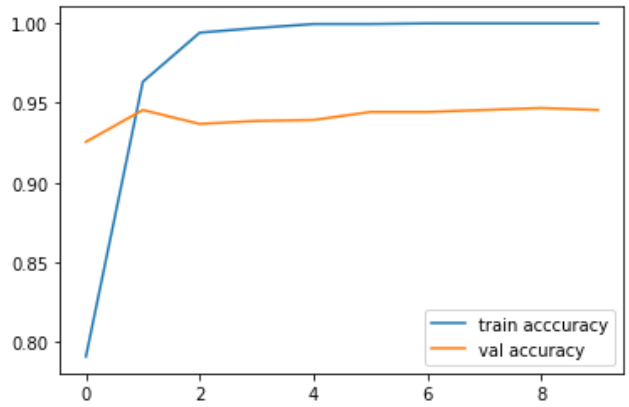
Predicted Values from Test Data

SNo	Review	Actual Class	Predicted Values	Predicted Decisions
1	دبستوني بطاقة اتمانية ماابغاها وكل قرش اخذتوه مني ظلم اخذه منكم يوم القيامة	0	[8.787419e-05]	0
2	ربنا يبارك فيكو	1	[0.9999251]	1
3	فالاول كان البرنامج جيد ويفتح بسرعه وأما الان فلا يفتح	0	[0.94823056]	1
4	كيف يتم إنشاء حساب	1	[0.01949596]	0
5	جميل رائع يوسهل في توصيل الطعام انصح بي تجربته وانصح من الشركه بان تطوره للاجمل	1	[0.999963]	1
6	بعد اخر تحديث التطبيق لايعمل مطلقا انا كل شغلي واقف بسبب عدم عمل التطبيق	0	[0.0038117]	0
7	تحديثكم الاخير هذا زفت ماعاد صار يدخلني للتطبيق..بسرعه حدثوه وصلحوا الخطأ .. بعد تحديثكم زفت الاخير حدثوه وصلحوا الخطأ ..مايدخلني للتطبيق اليوم ..	0	[0.002383]	0
8	زفت برنامج لا يعمل ولا يتحدث ايش نقول نشكركم غصب	0	[2.0066847e-05]	0
9	بنك لايقدر قيمة عملائه،والدليل انك تروح وماتحصل زحمة عند اي فرع لان الناس غسلو يدهم منه،وعطيتهم نجمة بس عشان التعليق	0	[0.01578149]	0
10	بعد التحديث لا يعمل بالمره نرجو حل المشكله وشكرا	0	[1.6286886e-05]	0

Throughout the training and testing phases of the proposed model, there is a decline in loss and an increase in accuracy, as depicted in Fig. 3.



(a)



(b)

Fig. 3. Loss (a) and Accuracy (b) During Training and Testing

Fig.3, loss and accuracy diagrams provide a dynamic view of the model's learning process. They assist in diagnosing issues, such as underfitting, overfitting, or convergence problems, and help researchers and practitioners make informed decisions about model adjustments and improvements during the training and testing phases. Lines for loss and accuracy on epoch during and testing are showing that model is neither underfit nor overfit.

5. Results and Discussions

The confusion matrix will be employed for this type of classification. Among the 1600 Arabic reviews considered as test data, 821 were classified as negative, and 779 were classified as positive. Out of the reviews classified as positive, 740 were indeed positive, while 39 were mistakenly labelled as negative. For those classified as negative, 773 were accurately identified as such, while 48 were erroneously labelled as positive. The representation of these results is presented in Fig.4 and Fig.5, demonstrating that the model performs well on the test data.

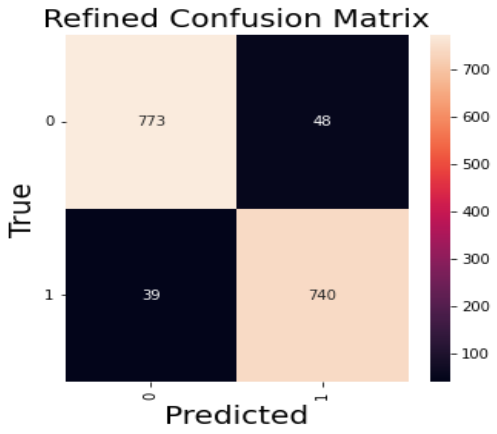


Fig. 4. Visual Representation of Confusion Matrix on Test Data

Fig. 4. contains Graphical representation of confusion matrix a colourful heatmaps, helps us easily understand how well a model is doing. It's like a visual report card for the model's predictions. Light pink colours usually mean the model is doing well, and dark colours show where it's making mistakes. Top left to bottom right diagonal contains light pink means model performed well.

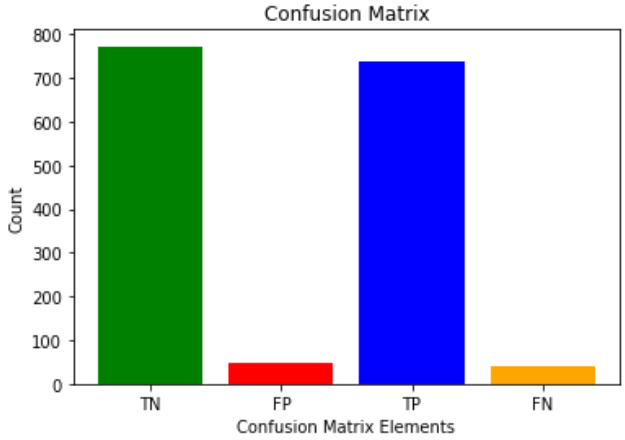


Fig. 5. Visual Representation of TN, FP, TP, FN

Fig. 5. indicates that model predicted true positive and true negative at very high rate while false positive and false negative at very low rate, indicates that efficiency of proposed model is very high.

5.1 ROC Curve and Absolute Mean Error

The acronym ROC stands for Receiver Operating Characteristic, and ROC curves are commonly employed to visually illustrate the trade-off between clinical sensitivity and specificity across various cutoff points for a test or a combination of tests. Moreover, the area under the ROC curve provides insight into the overall efficacy of the test(s) within a model. Since a larger area under the ROC curve indicates a more effective test, these areas are utilized to compare the utility of different tests [45]. Fig. 6. displays the optimal curve for the Hybrid Deep Learning Model on Test Arabic Data. The mean absolute error of the proposed model is 0.054,

indicating that the model is excellent, as reflected by this minimal value.

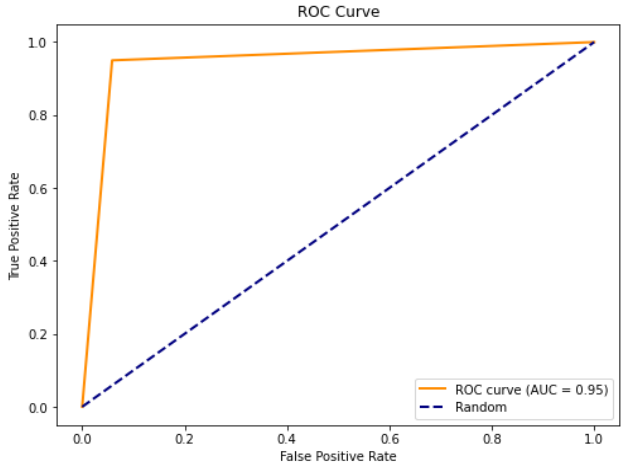


Fig. 6. ROC Curve of Proposed Model

5.2 Confusion Matrix Measures

Confusion matrices serve as a widely employed metric in addressing classification challenges, applicable to both binary and multiclass scenarios. The formula outlined below is employed to assess a model's accuracy through the construction of a confusion matrix. The matrix encapsulates true positive (TP) values, false positive (FP) values when misclassified into the relevant class, false negative (FN) values when misclassified into another class, and true negative (TN) values correctly categorized into another class. Based on these elements, commonly used classification performance metrics include accuracy (ACC) derived from Eq-1, precision (P) from Eq-2, recall (R) from Eq-3, and F-score from Eq-4 [46].

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$P = \frac{TP}{TP+FP} \quad (2)$$

$$R = \frac{TP}{TP+FN} \quad (3)$$

$$F - Score = 2 * \frac{P*R}{P+R} \quad (4)$$

Table 9 illustrates that the proposed model attained a precision of 95%, recall of 94%, and an F1 score of 95% for negative Arabic reviews. Regarding positive sentiment, the precision, recall, and F1 score are 94%, 95%, and 94%, respectively. The average values for positive and negative precision, recall, and F1 score stand at 95%, 95%, and 95%, and the overall accuracy is also 95%. These averaged metrics indicate the commendable performance of the proposed Hybrid Deep Learning Model for Arabic Reviews.

Table 9

Confusion Matrix Measures

Class	Precision	Recall	F1 Score	Support
Negative (0)	0.95	0.94	0.95	821
Positive (1)	0.94	0.95	0.94	779
Average	0.95	0.95	0.95	1600
Accuracy	0.95			

6. Comparison with Benchmarks

The work of [47] utilized SGRUs, SBGRU AraBERT, and the Ensemble deep learning algorithm, achieving an accuracy of 90%. In contrast, the work of [48] attained an F1-score of 81% using a GRU model. A stacking model based on RNN, GRU, and LSTM with a meta-learner SVM [49] achieved 83.12% accuracy, 83.85% precision, 83.12% recall, and 83.06% F1, while the proposed model outperformed with 95% accuracy, 95% precision, 95% recall, and 95% F1. All metrics of the proposed works are superior to those of previous studies. Additionally, the work of [13] employed two different techniques based on BiLSTM and CNN, resulting in accuracies of 92% and 93%, respectively. The proposed model achieves an accuracy of 95%, surpassing all the works. In Aspect-Based Sentiment Analysis on Arabic reviews [12], a Naïve Bayesian model achieved positive accuracies of 42%, 71%, 52%, and 60%, and negative accuracies of 54%, 93%, 36%, and 52% for accuracy, precision, recall, and F1 score, respectively. In comparison, the proposed work achieved significantly higher results with positive and negative accuracies, precisions, recalls, and F1 scores all at 95%, 94%, 95%, and 94%, respectively. The average results for positive and negative sentiment analysis based on accuracy, precision, recall, and F1 score in previous studies were 48%, 82%, 44%, and 56%, respectively. However, the proposed work demonstrated superior average results across all metrics with 95% accuracy, 95% precision, 95% recall, and 95% F1 score. Table 10 illustrates the comparison of the proposed model with various studies conducted on an Arabic dataset for sentiment analysis. The proposed approach achieved the highest accuracy of 95%. Benchmarks used dataset in Arabic form on their own model while proposed model used same dataset on hybrid proposed work.

Table 10

Proposed Model Comparison with Previous Studies

Work	Accuracy
Aspect-based using Support Vector Machine (SVM)[50]	76%
Deep Learning using Recursive Neural Tensor Networks (RNTN) [51]	80%
Syntax-based Aspect Detection [52]	74%

10 M Tweets After preprocessing 416,292 Tweets [47]	90%
The pre-trained RNN, GRU and LSTM with meta-learner SVM [49]	83%
Ontology-based for Domain Features Extraction and Weighting [10]	79%
BiLSTM, CNN [13]	92%, 93%
AraMAMS [12]	82%
Proposed Work	**95%**

7. Conclusion

Sentiment analysis in Arabic involves the extraction of subjective opinions from various users on different subjects. These opinions and sentiments play a crucial role in understanding individual perspectives and judgments within a specific domain. While numerous text classification (TC) studies have been conducted in languages like English, French, German, Spanish, Chinese, Greek, and Japanese, there is limited current research on automatically classifying text documents in Arabic due to challenges such as varying spellings, diverse character combinations, and the presence of short and long vowels (diacritics). In this study, the authors endeavoured to create a matrix from Arabic text data that is compatible with deep learning models. To enhance the accuracy of sentiment analysis for Arabic data, the proposed work combined powerful deep learning features, specifically Long Short-Term Memory Networks (LSTN) and Convolutional Neural Networks (CNN), to form a hybrid model for Arabic text. The experimental results demonstrated that the proposed model achieved impressive precision, recall, and F1 score, all standing at 95%, along with an overall accuracy of 95%.

Despite the commendable results obtained, it is essential to acknowledge certain limitations for future investigations. Firstly, this research is confined to reviews written in Arabic. Additionally, the proposed model utilizes a data split of 40% for testing and 60% for training, optimized over 10 epochs. Future studies may explore altering this ratio and duration based on different hyperparameters to assess their impact on accuracy. Furthermore, the proposed model employs embedded vectors with post-padding, but alternative hyperparameters and vector techniques, such as word2vec, bag-of-words, and TF/IDF, could be explored to further improve accuracy.

8. References

[1] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," Expert Syst. Appl., vol. 72, pp. 221–230, 2017, doi: 10.1016/j.eswa.2016.10.065.

[2] "3 reasons urgent care facilities should care about online reviews," 31 july 2017, 2017. https://resources.reputation.com/reputation-com-blog/3-reasons-urgent-care-facilities-should-care-about-online-reviews.

[3] F. Wu, Y. Huang, and Z. Yuan, "Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources," Inf. Fusion, vol. 35, pp. 26–37, 2017, doi: 10.1016/j.inffus.2016.09.001.

[4] B. Liu, Sentiment Analysis and Opinion Mining, vol. 5, no. 1. 2012.

[5] J. Jin, P. Ji, and R. Gu, "Identifying comparative customer requirements from product online reviews for competitor analysis," Eng. Appl. Artif. Intell., vol. 49, pp. 61–73, 2016, doi: 10.1016/j.engappai.2015.12.005.

[6] L. Shu, H. Xu, and B. Liu, "Lifelong learning CRF for supervised aspect extraction,", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 148–154, doi: 10.18653/v1/P17-2023.

[7] K. Khan, B. B. Baharudin, A. Khan, and Fazal-E-Malik, "Mining opinion from text documents: A survey," 2009 3rd IEEE Int. Conf. Digit. Ecosyst. Technol. DEST '09, vol. 3, no. 7, pp. 217–222, 2009, doi: 10.1109/DEST.2009.5276756.

[8] 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.

[9] M. Z. Asghar, A. Khan, S. Ahmad, and F. M. Kundi, "A review of feature extraction in sentiment analysis," J. Basic. Appl. Sci. Res, vol. 4, no. 3, pp. 181–186, 2014.

[10] S. M. Khabour, Q. A. Al-Radaideh, and D. Mustafa, "A new ontology-based method for arabic sentiment analysis," Big Data Cogn. Comput., vol. 6, no. 2, pp. 1–19, 2022, doi: 10.3390/bdcc6020048.

[11] A. Louati, H. Louati, E. Kariri, F. Alaskar, and A. Alotaibi, "Sentiment analysis of arabic course reviews of a saudi university using support vector machine," Appl. Sci., vol. 13, 2023, doi: https://doi.org/10.3390/app132312539.

[12] A. AlMasaud and H. H. Al-Baity, "AraMAMS: arabic multi-aspect, multi-sentiment restaurants reviews corpus for aspect-based sentiment analysis," Sustain., vol. 15, no. 16, 2023, doi: 10.3390/su151612268.

[13] A. Alqarni and A. Rahman, "Arabic tweets-based sentiment analysis to investigate the impact of covid-19 in KSA: a deep learning approach," Big Data Cogn. Comput., vol. 7, no. 1, 2023, doi: 10.3390/bdcc7010016.

[14] M. Akhtar and S. U. Rehman, "A machine learning approach for Urdu text sentiment analysis," Mehran Univ. Res. J. Eng. Technol., vol. 42, no. 2, p. 75, 2023, doi: 10.22581/muet1982.2302.09.

[15] J. Bu et al., "ASAP: A chinese review dataset towards aspect category sentiment analysis and rating prediction," 2021, [Online]. Available: http://arxiv.org/abs/2103.06605.

[16] A. al Owisheq, S. al Humoud, N. al Twairesh, and T. al Buhairi, "Arabic sentiment analysis resources: A survey," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9742, pp. 267–278, 2016, doi: 10.1007/978-3-319-39910-2_25.

[17] M. Iram, S. U. Rehman, S. Shahid, and S. A. Mehmood, "Anatomy of sentiment analysis of tweets using machine learning approach," Proc. Pakistan Acad. Sci. Part A, vol. 59, no. 2, pp. 61–73, 2022, doi: 10.53560/PPASA(59-2)771.

[18] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electron., vol. 9, no. 3, 2020, doi: 10.3390/electronics9030483.

[19] P. Cen, K. Zhang, and D. Zheng, "Sentiment analysis using deep learning approach," J. Artif. Intell., vol. 2, no. 1, pp. 17–27, 2020, doi: 10.32604/jai.2020.010132.

[20] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of persian movie reviews using deep learning," Entropy, vol. 23, no. 5, pp. 1–16, 2021, doi: 10.3390/e23050596.

[21] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, 2018, doi: 10.1002/widm.1253.

[22] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 2015, no. 2011, pp. 103–112, doi: 10.3115/v1/n15-1011.

[23] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," Proceedings of the 8th International Joint Conference on Natural Language Processing, 2015, pp. 253–263, [Online]. Available: http://arxiv.org/abs/1510.03820.

[24] S. Zhao, Z. Xu, L. Liu, M. Guo, and J. Yun, "Towards accurate deceptive opinions detection based on word order-preserving CNN," Math. Probl. Eng., vol. 2018, pp. 1–8, 2018, doi: 10.1155/2018/2410206.

[25] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 2015, vol. 1, pp. 1343–1353, doi: 10.3115/v1/p15-1130.

[26] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015, no. September, pp. 1422–1432, doi: 10.18653/v1/d15-1167.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] L. M. Al Qassem, D. Wang, Z. Al Mahmoud, H. Barada, A. Al-Rubaie, and N. I. Almoosa, “Automatic arabic summarization: a survey of methodologies and systems,” Procedia Comput. Sci., vol. 117, pp. 10–18, 2017, doi: 10.1016/j.procs.2017.10.088.

[32] 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.

[33] 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.

[34] L. Fodil, H. Sayoud, and S. Ouamour, “Theme classification of arabic text: a statistical approach,” Terminol. Knowl. Eng., no. Jun, p. 10 p, 2014, [Online]. Available: https://hal.archives-ouvertes.fr/hal-01005873/.

[35] N. Hicham, H. Nassera, and S. Karim, “Enhancing arabic e-commerce review sentiment analysis using a hybrid deep learning model and fasttext word embedding,” EAI Endorsed Trans. Internet Things, vol. 10, 2024, doi: 10.4108/eetiot.4601.

[36] M. R. R. Rana et al., “Aspect-based sentiment analysis for social multimedia: a hybrid computational framework,” Comput. Syst. Sci. Eng., vol. 46, no. 2, pp. 2415–2428, 2023, doi: 10.32604/csse.2023.035149.

[37] A. M. A. MESLEH, “Chi square feature extraction based svms arabic language text categorization system,” J. Comput. Sci., vol. 3, no. 6, pp. 430–435, 2007, doi: 10.3844/jcssp.2007.430.435.

[38] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed, and A. Al-Rajeh, “Automatic arabic text classification,” Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, 2008, pp. 77–84, doi: 10.1109/ICCIS.2008.4670777.

[39] M. N., I. M., A. H., and H. A., “Opinion mining and analysis for arabic language,” Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 5, pp. 181–195, 2014, doi: 10.14569/IJACSA.2014.050528.

[40] Z. A. Abutiheen, A. H. Aliwy, and K. B. S. Aljanabi, “Arabic text classification using master-slaves technique,” J. Phys. Conf. Ser., vol. 1032, no. 1, 2018, doi: 10.1088/1742-6596/1032/1/012052.

[41] M. Ramzy and B. Ibrahim, “User satisfaction with arabic covid-19 apps: sentiment analysis of users’ reviews using machine learning techniques,” Inf. Process. Manag., vol. 61, no. 3, 2024, doi: 10.1016/j.ipm.2024.103644.

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

[43] A. Khattak, A. Habib, M. Z. Asghar, F. Subhan, I. Razzak, and A. Habib, “Applying deep neural networks for user intention identification,” Soft Comput., vol. 25, no. 3, pp. 2191–2220, 2021, doi: 10.1007/s00500-020-05290-z.

[44] H. Ahmad, M. U. Asghar, M. Z. Asghar, A. Khan, and A. H. Mosavi, “A hybrid deep learning technique for personality trait classification from text,” IEEE Access, vol. 9, pp. 146214–146232, 2021, doi: 10.1109/ACCESS.2021.3121791.

[45] S. Ekelund, “ROC curves – what are they and how are they used?,” 2011. [Online]. Available: https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used.

[46] A. Kulkarni, “Confusion matrix confusion matrix is a very popular measure used while solving classification problems.,” Data Democracy, Elviewer, 2020. https://www.sciencedirect.com/topics/engineering/confusion-matrix#:~:text=A confusion matrix is a,malignant tissue is considered cancerous.

[47] S. Alhumoud, “Arabic sentiment analysis using deep learning for covid-19 twitter data,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 20, no. 9, p. 132, 2020.

[48] L. A. Alhuri, H. R. Aljohani, R. M. Almutairi, and F. Haron, “Sentiment analysis of covid-19 on saudi trending hashtags using recurrent neural network,” in Proceedings - International Conference on Developments in eSystems Engineering, DeSE, 2020, vol. 2020-Decem, pp. 299–304, doi: 10.1109/DeSE51703.2020.9450746.

[49] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, and T. Alkhalifah, “Heterogeneous ensemble deep learning model for enhanced arabic sentiment analysis,” Sensors, vol. 22, no. 10, 2022, doi: 10.3390/s22103707.

[50] M. Al-Smadi, O. Qwasmeh, B. Talafha, M. Al-Ayyoub, Y. Jararweh, and E. Benkhelifa, “An enhanced framework for aspect-based sentiment analysis of Hotels’ reviews: Arabic reviews case study,” 2016 11th Int. Conf. Internet Technol. Secur. Trans. ICITST 2016, pp. 98–103, 2017, doi: 10.1109/ICITST.2016.7856675.

[51] R. Baly, H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj, “A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 16, no. 4, 2017, doi: 10.1145/3086576.

[52] M. Mataoui, T. E. Bendali Hacine, I. Tellache, A. Bakhtouchi, and O. Zelmati, “A new syntax-based aspect detection approach for sentiment analysis in Arabic reviews,” 2nd International Conference on Natural Language and Speech Processing, ICNLSP 2018, 2018, pp. 1–6, doi: 10.1109/ICNLSP.2018.8374373.

