Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777803

ANALYSIS AND CLASSIFICATION OF CUSTOMER REVIEWS IN ARABIC USING MACHINE LEARNING AND DEEP LEARNING

Ali Mohammad Salloum[1]*, Muhammad Mazen Almustafa[2]
[1]*Department Of Web Sciences, Syrian Virtual University, Syria, Email: ali_134861@svuonline.org
[2]Department Of Web Sciences, Syrian Virtual University, Syria, Email: t_mmustafa@svuonline.org
***Corresponding Author: Ali Mohammad Salloum**
*Department Of Web Sciences, Syrian Virtual University, Syria, Email: ali_134861@svuonline.org

Abstract
Analyzing and classifying customer reviews in Arabic presents a significant challenge due to the diverse nature of the Arabic language, encompassing various dialects and nuances. In this paper, we address this challenge by employing machine learning and deep learning techniques to analyze and classify Arabic customer reviews. Our study is based on a comprehensive dataset compiled from various sources, containing 33,333 positive and 33,333 negative reviews after filtering out mixed sentiments. To tackle this task, we explore a range of machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Trees, and Logistic Regression, along with various feature extraction methods such as TF-IDF and Word2Vec. Additionally, we delve into the realm of deep learning by employing Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Bidirectional LSTM (BiLSTM) to further enhance sentiment analysis performance. Our experiments in ML reveal that SVM with an RBF kernel and W2V and Ngrams feature extraction achieves the highest F1-score of 82.5% . Among the DL models, BiLSTM with 128 units and dropout = 0.2 emerges as the top performer with an F1-score of 87%. These findings underscore the effectiveness of deep learning techniques in handling the complexities of Arabic text analysis. our research provides valuable insights into sentiment analysis of Arabic customer reviews and presents a comprehensive evaluation of machine learning and deep learning algorithms, paving the way for enhanced customer feedback analysis in the Arabic-speaking world.
Keywords: sentiment analysis, Natural Language Processing, Arabic reviews Analysis, Arabic Language Processing, Deep learning, Machine Learning.

Introduction

The use of modern information technology has become an integral element of work in companies due to its effectiveness, speed of achievement and abundance of information. Companies have begun to shift from manual marketing to electronic marketing that relies on electronic media and social media platforms. This movement represented a qualitative leap in the field of marketing, and Customers are provided with a consistent and robust online environment to benefit from the services provided. In view of this new trend for companies, data collection and analysis through social media has become one of the most important and

accurate steps needed to reach successful marketing and achieve a high sales rate. As these steps contribute to dividing the target customers into segments of different interests by analyzing their data through social media, and thus contribute to the correct orientation to each segment of society with the types and types of products that interest them. Customer satisfaction is the key in assessing how a product or service of a company meets customer expectations [1], [2] and is an important tool that can give organizations major insights into every part of their business, thus helping them to increase earnings or minimize marketing expenses [1], [3]. Customer feedback might help in reviewing the factors that were not previously considered, such as shipping, safe packing, politeness and available customer service consultants and a user-friendly website. Nothing can make customers feel that they are important than asking for their views and valuing their comments. When a customer is asked for any opinion on a product or experience, they feel valued and connected to the organization [1], [4].

Sentiment analysis is one type of analysis method used in natural language processing that mostly focusing on the binary text classification such as positive or negative opinion. This analyzing method can be made automatic by adopting the machine learning technique that is the method being able to learn some patterns like a human does. But machine learning can search for patterns faster and analyze more data than a human could. Machine learning is thus a good tool for developing sentiment analysis model and analyzing text [5]. Deep learning is state of the art learning algorithm developing from the success of neural network. The learning efficiency is increased by stacking more layers of the neural network. Hence, deep learning can analyze data more deeply and get useful features to increase the performance of the predictive model [6]. Presently, deep learning is a popular learning algorithm extensively applied in many application domains such as image processing, image segmentation, and object detection [5].

In the realm of linguistic analysis and information extraction from texts, research and applications have traditionally centered around major languages such as English. However, there is a growing interest in sentiment analysis and customer reviews in the Arabic language. Customer reviews represent a rich source of information about diverse products and services, encapsulating the opinions and sentiments of consumers.

It's worth noting that there is a clear deficiency in research and available tools for sentiment analysis and customer reviews in the Arabic language. This scarcity poses a significant challenge to effectively understand and utilize this data. Therefore, this research aims to fill this gap and provide a comprehensive solution for sentiment analysis of Arabic customer reviews.

In this paper, we leverage standardized data from Arabic customer reviews on Amazon as a primary data source. We will develop and compare several scenarios and methods for analyzing these reviews and extracting valuable insights. We will employ machine learning and deep learning techniques to build models that enhance our understanding of the sentiments and opinions expressed in these reviews. This will enable us to gain a better understanding of customer preferences and sentiments regarding various products and services. We will evaluate the models that will be built using the criteria for evaluating the appropriate classification algorithms, the most famous of which is F1_Score, and compare them on this basis. And then choose the model that gives the highest performance in the rating of customer reviews.

This research aspires to make significant contributions to the field of sentiment analysis in the Arabic language and encourages the development of new tools and techniques for comprehending reactions and responses within this vital linguistic context.

Background

Sentiment analysis

Human can easily read texts and recognize reviewer's sentiment by understanding context, but for computers it is not [7]. Sentiment Analysis is a relatively recent study that deals with the processing of natural texts published in web sites and social networks. However, the processing of texts written in the Arabic language is one of the challenges that specialists face because people do not rely on standard Arabic, writing people in spoken/colloquial languages and use various dialects [7]. sentiment annotation of an utterance, we consider both implicit and explicit affect information. The implicit sentiment of an utterance is determined with the help of context. Whereas, explicit sentiment of an utterance is determined directly from itself, and no external knowledge from the context is required to infer it. We consider three sentiments classes, namely positive, negative and neutral [8].

Sentiment analysis is also called as Opinion analysis or Opinion mining. We have seen a recent growth in the sentiment analysis task. The variuos research works in sentiment analysis [9] published an overview on Opinion mining in the earlier stage. There is a difference between rating opinions and rating sentiments, because the opinion may be descriptive and does not contain any feelings, and it can also contain feelings. The work of Munezero et al [10]. proposes an enlightening discussion, funded on physiological and psychological studies, which tries to eventually establish a fixed frame of definitions for sentiment analysis. We consider these definitions as the basis for our classification of customer reviews.

In the cited work, opinions are considered as personal interpretations of information about a topic, while sentiments are prompted by emotions. As an example, an opinion could be:

(1) "the battery of the smartphone has a good capacity", while a sentence containing a sentiment could be: (2) "I always loved the longevity of the battery of all the smartphones of this manufacturer, since the first model". Considering this difference, it appears clear that an opinion can also be expressed without any emotion, for instance in a descriptive way, as in the sentence: (3) "In my experience, the average duration of the battery is 8 hours with a normal use". Sentiment can be also considered a social construct of emotions that develop over time and are enduring, meaning that the temporal aspects are critical for the sentiment, as is evident from the previous example. Opinions, on the other hand, are just personal interpretations of facts that may or may not be emotionally charged (in the above example, the first sentence contains an opinion with a positive orientation, while the third sentence contains an opinion without orientation or with a neutral orientation). And even when an opinion expresses a certain kind of subjectivity and judgment, it does not necessarily imply that there is a sentiment [11]. the authors conclude that it is important to identify which is the desired output of a SA tool, whether it is to classify an opinion or to identify emotion in the text.

Machine Learning

There is no doubt that analyzing customer reviews in the Arabic language constitutes a great challenge for anyone who wants to do this task, machine learning with its various algorithms and its wide use recently with these tasks is a good technique for analyzing and classifying customer reviews, so we will review the most important machine learning algorithms used in these tasks.

SVM

Support vector machines are statistical- and machine-learning techniques with the primary goal of prediction. They can be applied to continuous, binary, and categorical outcomes analogous to Gaussian, logistic, and multinomial regression [12]. in this paper, we evaluate two distinct kernel models for Support Vector Machine: RBF and Linear. In this research [13], all SVM equations for classifications tasks are explained. The best scenario we achieved using SVM obtained an F1_Score of 82.5%.

KNN

The classic KNN algorithm is a supervised machine learning algorithm that is predominantly used for classification purposes. The algorithm consists of a variable parameter, known as k, which translates to the number of 'nearest neighbors. the equations of which are explained in this research [14]. Using this algorithm, we obtained an F1 Score of 54.8%.

Random Forest

Each review will be classified into positive or negative category. In this paper, we employ random forest for the classification task. Random forest algorithm is a supervised classification algorithm. It is an ensemble learning technique based on decision tree algorithm [15], [16]. Using this algorithm, we obtained an F1_Score of 80.9%.

Decision Tree

Decision Tree is an algorithm that use trees to predict the outcome of an instance. Essentially, a test node computes an outcome based on the attribute values of an instance, where each possible outcome is associated with one of the subtrees. The process of classify an instance starts on the root node of the tree. If the root node is a test, the outcome for the instance Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis it is predicted to one of the subtrees and the process continues until a leaf node it is encountered, in this situation the label of the leaf node gives the predicted class of the instance [17]. Using this algorithm, we obtained an F1_Score of 73.1%.

Logistic Regression

As we are going to classify reviews in a positive or negative class, LR adopted because of high efficacy in binary classification tasks. LR uses a threshold boundary to isolate the positive reviews from the negative ones. LR uses a Logistic function to estimate probabilities between positive or negative label y and data features w given by input x. Thus, LR uses sigmoid function to get the likelihood directly by minimizing infinitive $+\infty$ and $-\infty$ into a scale between 0 to 1 [18]. Using this algorithm, we obtained an F1_Score of 82.1%.

Deep Learning

Sentiment analysis of a large number of user reviews on e-commerce platforms can effectively improve user satisfaction [19]. To create a high-performance model for analyzing and classifying sentiments, it is necessary to use deep neural networks and advanced techniques in the field of deep learning, as deep learning can recognize complex patterns and create better connections than machine learning, especially when it comes to analyzing and understanding texts. In recent years, deep learning has made great achievements in many fields. Compared with traditional machine learning methods, deep learning does not need human intervention features, but deep learning needs massive data as support. Deep learning-based methods automatically extract features from different neural network models and learn from their own errors [19].

CNN is widely popular because it can be used in image datasets by extracting the significant features of the image while the 'convolutional' filter (i.e., kernel) moves through the image . Many researchers in the field of natural language processing rely on CNNs to analyze and classify sentiments, since CNN-based models have proven their ability to deeply understand patterns in texts. This research [20] is a systematic review that contains a group of related research in the same field and explains the techniques and algorithms used in each research.

Researchers have recently been using artificial intelligence (AI) techniques in various aspects of life, and natural language processing (NLP) has had a large share in this. One of the most important techniques that have been used with it is machine learning (ML), but recently (in the past decade) there has been greater reliance on learning. Deep (DL), which is considered one of the sections of (ML) and appeared in 2010, and Figure 1 shows the relationship between artificial intelligence, machine learning, and deep learning.

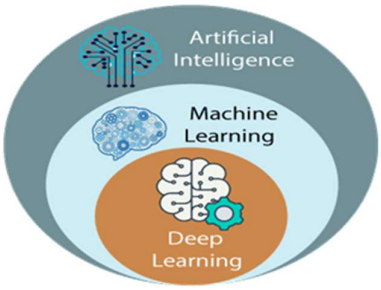

Figure 1 relationship between AI, ML and DL

The architecture of CNNs consists of convolutional and subsampling layers (figure 2). The convolutional layer performs feature extraction from the input data and generates feature maps. The feature map is computed through an element-wise multiplication of the small matrix of weights (kernel) and the matrix representation of the input data, and the result is summed. This weighted sum then passed through the non-linear activation function. One of the most common is the function ReLu, which is given in equation (1) as:

$$ReLu(x) = max(0, x)$$

The output in the case of binary classification is of two labels (either 0 or 1), so the Sigmoid function is used in the output layer, which is given by equation (2).

$$\delta(z) = \frac{1}{1 + e^{-z}}$$

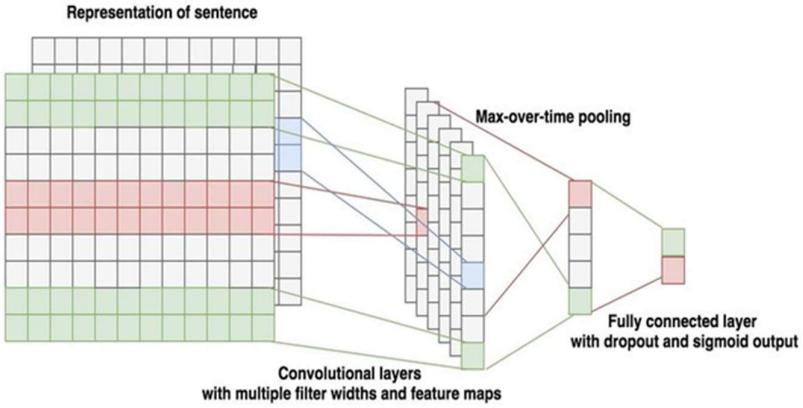

Figure 2 the CNN architecture for binary text classification

Binary_crossentropy is a loss function used in binary classification problems scenarios where each input sample belongs to exactly one of two classes. For instance, an email can be either 'spam' or 'not spam', a patient can be 'diseased' or 'healthy' and a sentiment can be 'positive' or 'negative'. The equation (3) for this function.

$$\log \text{loss} = \frac{1}{N} \sum_{i=1}^{N} - (y_i * \log(P_i) + (1 - y_i) * \log (1 - P_i))$$

LSTM model proposed by Hochreiter and Schmidhuber [21] introduces the concept of a state for each of the layers of a RNN which plays the role of memory. The input signal affects the state of the memory, and this, in turn, affects the output layer, just like in an RNN. But this state of memory persists throughout the time steps of a sequence (for example, time series, sentence, or text document). Therefore, each input signal affects the state of the memory as well as the output signal of the hidden layer [22]. The LSTM model is able to store the previous information thus capturing the prominent long-range dependencies in the given input. LSTM has performed significantly in sequence modeling tasks, such as text classification, sentiment analysis, time series prediction, etc. There are three important elements in the LSTM model, including forget gate, input gate, and output gate. The forget gate will decide to forget or discard irrelevant information from the previous cell state and new input data. The input gate plays the role of a filter to decide which information is worth remembering, thus to be updated into the next state. The value close to zero means that it is less important to be updated. The output gate determines the information that should be the output in the next cell state [23]. The calculations of the single LSTM unit at a single time step t in the forget gate ft , input gate it , output gate ot and cell state ct , are defined as follows (equations 4, 5, 6 , 7, 8, 9):

$$f_t = \boldsymbol{\delta}(\boldsymbol{W_f X_t} + \boldsymbol{U_f h_{t-1}} + \boldsymbol{b_f})$$
$$i_t = \boldsymbol{\delta}(\boldsymbol{W_i X_t} + \boldsymbol{U_i h_{t-1}} + \boldsymbol{b_i})$$

$$o_t = \boldsymbol{\delta}(\boldsymbol{W_o X_t} + \boldsymbol{U_o h_{t-1}} + \boldsymbol{b_o})$$
$$\tilde{c_t} = \boldsymbol{tanh}(\boldsymbol{W_c X_t} + \boldsymbol{U_c h_{t-1}} + \boldsymbol{b_c})$$
$$c_t = \boldsymbol{f_t} * \boldsymbol{c_{t-1}} + \boldsymbol{i_t} * \tilde{c_t}$$
$$h_t = \boldsymbol{o_t} * \boldsymbol{tanh}(\boldsymbol{c_t})$$

where σ is the sigmoid function, Xt denotes the input, $(\boldsymbol{W_f}, \boldsymbol{W_i}, \boldsymbol{W_o}, \boldsymbol{W_c}, \boldsymbol{U_f}, \boldsymbol{U_i}, \boldsymbol{U_o}, \boldsymbol{U_c})$ and $(\boldsymbol{b_f}, \boldsymbol{b_i}, \boldsymbol{b_o}, \boldsymbol{b_c})$ denote the weight matrices and biases in the forget gate, input gate, output gate, and cell state, correspondingly. . $\boldsymbol{h_{t-1}}$ and c_(t-1) are the output of the LSTM at time step t-1 . The operation * is the element wise multiplication. Figure 3 explain the architecture for LSTM cell.

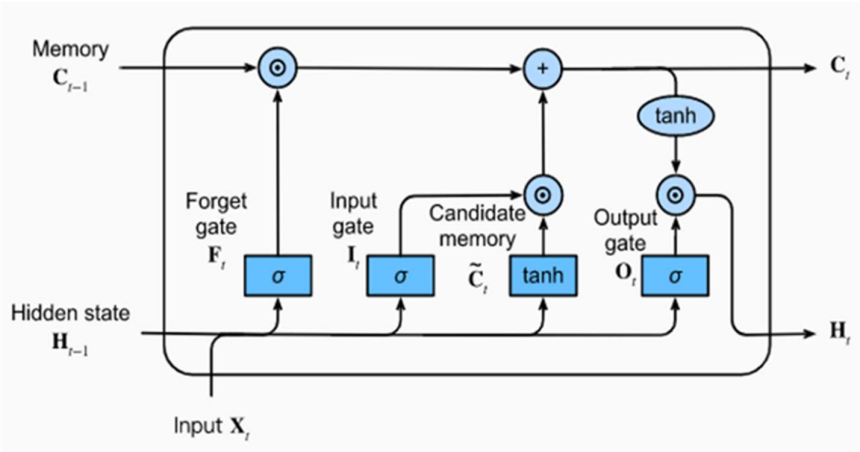

Figure 3 LSTM cell architecture

Unidirectional (standard) LSTM only preserves information of the past because the only inputs it has seen are from the past. Unlike standard LSTM, in BiLSTM (Bidirectional LSTM) model the input flows in both directions and it's capable of utilizing information from both sides. So BiLSTM is a sequence processing model that consists of two LSTMs layers: one taking the input in a forward direction (from "past to future"), and the other in a backwards direction (from "future to past") [22].

Although LSTM solves the long-term dependency problem, it is hard to utilize the contextual information of the text. The model design concept of BiLSTM is to make the feature data obtained at time have information between the past and the future at the same time. Experiments have shown that this neural network structure model has better text feature extraction efficiency and performance than a single LSTM structure model. In text sentiment classification, BiLSTM also considers the context of the text, and uses the output of the CNN pooling layer as the input of two LSTM networks with opposite time series. The forward LSTM can obtain the above information of the input sequence, and the backward LSTM can obtain the above information of the input sequence. The context information of the input sequence is then calculated by vector splicing to obtain the final hidden layer representation. It is worth mentioning that the LSTM neural network parameters in BiLSTM are independent of each other, and they only share the word-embedding word vector list [24].

BiLSTM increase the amount of information available to the network, improving the context. It's also more powerful tool for modeling the sequential dependencies between words and phrases in both directions of the sequence than standard LSTM. BiLSTM is usually used when we have the sequence-to-sequence tasks but it should be noted that BiLSTM (compared to LSTM) is a much "slower" model and requires more time for training [22]. Figure 4 explain the architecture for BILSTM.

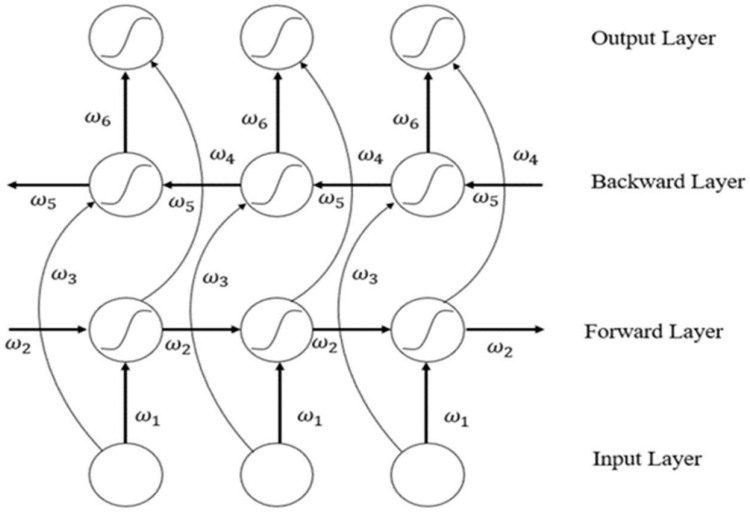

Figure 4 BILSTM architecture

In this research paper, deep learning achieved better results compared to machine learning, and the results will be explained in the Results and Discussion section.

Related works

In social networks people can express and share their opinions and experiences using various types of social data such as text data (comments, tweets, etc.), as well as multimedia data (e.g., videos, sounds). A huge volume of data is generated from social networks on a daily basis, and this data reflects the emotional attitudes of the audience towards various aspects of life such as political, business and social topics. Social data is described as informal, unstructured and rapidly evolving contents, thus processing and analyzing this data using traditional analysis methods is a time-consuming and resource-intensive task [25].

Natural language processing algorithms enable various language-related tasks such as Part-of-Speech (POS) tagging, parsing, machine translation, and dialogue systems. Sentiment analysis has become an important research topic in the field of natural language processing due to its crucial role in analyzing public opinion and making data-driven decisions. Arabic is one of the languages widely used on social networks. However, the richness and diversity of its dialects make it a challenging language for sentiment analysis [26].

In work done by Pandey et .al [27],they built a system for collecting and analyzing opinions about products from various online shopping websites. The main focus of the study was opinion mining, where opinions were categorized into two types:

Direct Opinions: These are textual documents that directly express a positive or negative opinion about a product. For example: "The battery backup for this phone is extremely poor."

Comparison Opinions: These opinions involve comparing the subject with something else. Opinion mining is a subset of web mining.

Initially, they gathered the data for analysis, which consisted of product reviews from the website amazon.com. These reviews covered over 500 reviews of electronic products, including computers, mobile phones, tablets, and electronics. The data was collected from approximately 3.2 million individuals' opinions on 10,001 different products. Each review included the following information:

- Reviewer_ID
- Product Model
- Date and Time of the review
- Review text

They then performed data filtering based on Parts of Speech (POS). They categorized words into eight parts of speech (noun, verb, pronoun, etc.), and certain words that didn't convey sentiment (pronouns, conjunctions, etc.) were filtered out. They adopted the idea that if a sentence contained a single positive word, it would be considered positive, and if it contained a single negative word, it would be considered negative. They emphasized the importance of POS in sentiment classification because certain words like nouns and pronouns don't carry sentiment and can be filtered out, while verbs like "improve" convey a significant amount of sentiment. They also mentioned using a POS tagger that provides 46 tags instead of just 8 [27].

Sentiment analysis is a significant research field in natural language processing (NLP) that focuses on categorizing opinions or emotions toward a product or service into predefined sentiment labels. Text data used for sentiment analysis leverages text mining, linguistics, and statistical knowledge-based techniques to automatically assign sentiment labels (e.g., positive, negative, or neutral) to user-generated text found online [28]. However, the labels can vary depending on the context of sentiment analysis.

Sentiment analysis encompasses various subtasks such as polarity classification, aspect-based sentiment analysis, sarcasm detection, and more. These tasks can be performed at both the sentence and document levels [29], [30].

For decades, numerous machine learning algorithms like Support Vector Machines (SVM) and Logistic Regression have been proposed to address various NLP challenges [31], [32]. These algorithms are known for their effectiveness and their ability to learn automatically [33].

In work done by Shafin et .al [34],They collected more than 1000 customer comments on products for this study, where they analyzed sentiments in the comments using Natural Language Processing (NLP) techniques and several machine learning algorithms. Analyzing customer comments on specific products aids in the future development of these products.

Initially, they gathered the data for the study, which consisted of approximately 1020 customer comments in the Bengali language (Bangla), collected from e-commerce platforms. They then performed data preprocessing, which included removing stop words, punctuation, and emojis from the text [34]. They also added a label column, classifying the data into two categories:

positive and negative comments. They utilized natural language processing techniques, such as tokenization, to segment and understand the comments. After data preprocessing, they entered the phase of using machine learning algorithms to train models capable of classifying the comments. They employed several algorithms, including SVM, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree. During the training and testing of the models, they experimented with multiple scenarios that varied in the proportion of test data, ranging from 30% to 70%. They observed that as the proportion of test data increased, the accuracy of the models tended to decrease. The best result they achieved was an accuracy of 88.81% with the SVM algorithm using a 30% test data proportion [34].

In work done by Shah et .al [35], they categorized sentiment analysis of customer reviews on Amazon products into two methods. The first method relies on machine learning approaches, where they mentioned several algorithms that can be employed in this domain. The second method relies on lexicon-based approaches, and there are three types:

- Dictionary-based approach
- Manual opinion approach
- Corpus-based approach

They utilized customer review data from the Amazon website, which is available on Kaggle. They extracted several features from this data, including customer opinions, which included:

- Product_ID
- Review text
- Review rating

However, the data exhibited a bias towards positive comments (most of the comments were positive). They performed data preprocessing, which involved removing null values, tokenization, converting all words to lowercase, removing stop words, and stemming. Ratings on Amazon range from 1 to 5 for each review. They categorized 1 and 2 as negative comments, 3 as neutral, and 4 and 5 as positive. So, the classification had three categories:

- Score 0: Negative
- Score 1: Neutral
- Score 2: Positive

They also used the tf-idf method for feature extraction. They split the dataset into training and testing sets (80% for training and 20% for testing) and employed several classification algorithms, including:

- Logistic Regression
- Bernoulli Naïve Bayes
- Multinomial Naïve Bayes
- Random Forest

They evaluated the testing data using various evaluation metrics typically used for classification problems, including precision, recall, accuracy, and F1-Score. The results indicated that the Random Forest algorithm outperformed others in terms of performance across various evaluation metrics, achieving a final accuracy of 93.17%, which was the highest among the tested algorithms [35].

In work done by Hakimi et .al [36], The aim of this paper was to utilize artificial intelligence techniques to enhance sustainable product design by extracting positive linguistic words related to products. A novel methodology for text analysis was proposed using techniques to break down the text into individual words, text embeddings, deep learning models, and attention mechanisms.

Extensive experiments were conducted using 323,150 reviews from real customers on Arabic e-commerce websites. Positive evaluations from customers regarding products were extracted. The impact of interrelated parameters, such as the size of the used vocabulary, the size of internal reviews, and the number of training model iterations, was studied. They employed pre-trained models in review processing, using the w-Bert model to transform each word into an embedding. Feature extraction was performed on these word embeddings using a convolutional layer, followed by max-pooling to reduce the feature set obtained from the previous layer. An attention layer was used to calculate the importance of each word, and the features were then input into a fully connected layer with a sigmoid function to classify the review as either positive or negative. They discussed various scenarios to precisely determine the model's parameters and achieved high accuracy. The result, with an F1-Score of 96.1%, indicated the best-performing model [36].

In work done by Elnagar et .al [37], They conducted sentiment analysis on a large and well-known dataset in the Arabic language known as BRAD, which contains 692,586 reviews. This dataset includes reviews in Modern Standard Arabic (MSA) as well as various dialects such as Levantine, Gulf, and Egyptian Arabic. Additionally, the dataset contains some reviews in Persian, which is structurally similar to Arabic. They processed the data using natural language processing techniques like unigrams, bigrams, and trigrams. After preparing and cleaning the data and extracting features, they applied several supervised learning algorithms, including Naïve Bayes, Decision Tree, Random Forest, XGBoost, and SVM. They also applied unsupervised learning algorithms, namely CNN and RNN. They achieved good results in both supervised and unsupervised learning, with F1-Score values ranging from 90% to 91%, indicating strong performance [37].

Dataset

Dataset resources

The quality and nature of the data is one of the most important factors affecting the success and evaluation of the research, The data related to customer reviews in Arabic are somewhat limited, especially considering the broad and extensive nature of this topic. In this paper, we relied on data for customer reviews in Arabic on Amazon products. The dataset in [38].

This dataset is mainly a compilation of several available datasets and a sampling of 100k rows (99999 to be exact). The dataset combines reviews from hotels, books, movies, products and a few airlines. It has three classes (Mixed, Negative and Positive). Most were mapped from reviewers' ratings with 3 being mixed, above 3 positive and below 3 negatives. The dataset has no duplicate reviews. In this paper, We deleted the mixed data because in this research, our aim is to identify only positive and negative reviews, and we do not give importance to neutral customers.

Dataset description and analysis

The dataset that we have consists of two features as follows:

- Text: Contains customer reviews on products in Arabic.
- Label: 'positive' or 'negative' (after deleted mixed).

Table 1 contains examples from dataset:

Table 1 examples from dataset

label	text
Positive	من أجمل ما قرأت.. رواية تستحق القراءة فعلا..
Positive	جيد . . لا يوجد غير حمام واحد في الشقة
Positive	روايه مشوقه مشاعر راقيه اختيار اسماء متناغم و الاهتمام بتفاصيل ممتع سعدت بقراءتها
Negative	ضعيف جدا. قريب من الحرم ..... لكن يوجد طلعة الفندق صعبة لكبار السن. كل شيء
Negative	اقامه سيبه جددددا. تعامل الموظفين في خدمات الغرف والاستقبال مغفنين جدا وقليلي الحياء

The other challenge in dealing with this Arabic dataset is the diversity of dialects. The reviews include both standard Arabic language and various dialects such as Egyptian, Gulf, and Levantine. This variation poses a significant challenge in creating a model capable of analyzing and classifying comments despite their dialectal diversity.

After analyzing and examining this dataset, we observed that it is balanced, containing 33,333 positive reviews and 33,333 negative reviews (after removing mixed reviews). We calculated the percentage of both positive and negative reviews. Figure 5 shows the percentage of positive and negative reviews in the dataset.

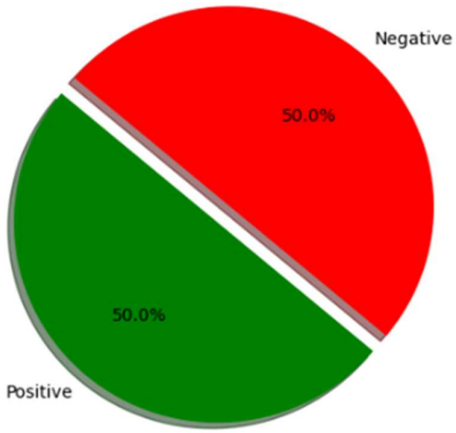

Figure 5 the percentage of positive and negative reviews in the dataset

During the data analysis, we did not calculate the percentage of reviews written in Egyptian, Gulf, Levantine, or Standard Arabic dialects due to the absence of a distinct feature in the dataset that indicates the dialect of each review. If we want to calculate the percentage of each dialect, it would require manual effort and be very time-consuming, which is unnecessary for our purposes.

Methodology

In this paper, We analyzed and classified customer reviews in Arabic on Amazon products using machine learning algorithms and using deep learning. The implementation of the process we followed is shown in Figure 6. The input is text (reviews) and the output is (positive or negative).

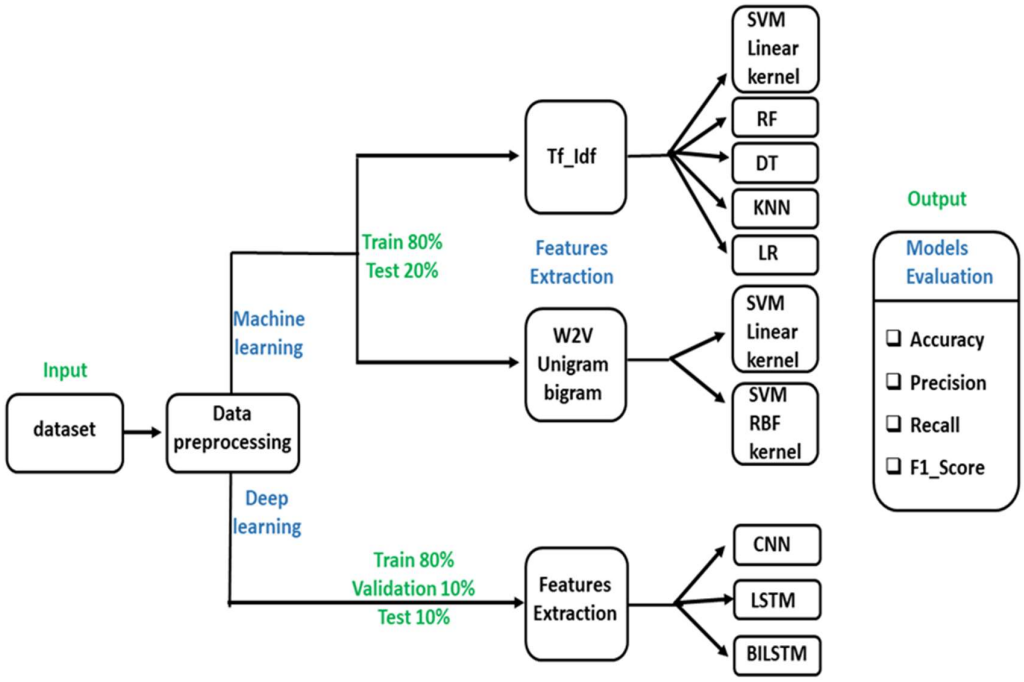

Figure 6 the methodology used Our methodology can be divided and explained into four stages.

Data Preprocessing

The goal of this process is to prepare the data for model training, which varies depending on the type of data. Each type of data requires different processing methods. In this stage, we cleaned the data and applied appropriate processing operations, such as removing numbers, links, non-Arabic characters, punctuation marks, emojis, diacritics from Arabic letters, eliminating stop words, and then stemming the words, along with other processing operations. We used Python for the entire coding process. In this stage, we relied on data processing modules such as pandas, numpy, nltk, re, and string. We converted the output column into numbers, performing a mapping operation where each negative value was assigned -1, and each positive value was assigned 1. We saved the clean data in a new file for use in the subsequent steps. We did not need the data augmentation process because our data is balanced. We conducted a simple data analysis, as explained in section Four (Dataset).

Features Extraction

Feature extraction is an important and fundamental step before creating and training machine learning and deep learning models. This step involves converting text into numbers so that algorithms can understand, analyze, and classify texts. There are many algorithms involved in

this process (as discussed in section Two). For machine learning, we used the tf-idf algorithm to extract features from the texts before training the various algorithms we used. TF-IDF is often used to represent documents as vectors in a high-dimensional space, where each term corresponds to a dimension, and the TF-IDF score for each term is the value along that dimension. These vectors can then be used for various text analysis tasks. This paper [22] explains Feature Extraction methods and how the TF-IDF method works in detail .

We also experimented with the word2Vec method along with unigram and bigram approaches for feature extraction from texts, but only with the SVM algorithm (both linear and rbf) because we did not observe a significant difference between feature extraction methods.

For CNN and LSTM neural networks, we extracted features using the tokenizer and pad_sequence. We also used them with BILSTM neural networks, but with the addition of W2V.

Machine Learning

After preparing the dataset and extracting the features, we divided it into 80% training data and 20% testing data. We applied several different algorithms with various feature extraction methods as follows:

- SVM (kernel 'linear', C=1) with Tf_Idf.
- SVM (kernel 'linear', C=1) with W2V and Ngrams.
- SVM (kernel 'rbf', C=0.5) with W2V and Ngrams.

C in SVM refers to regularization.

- KNN (k=5) with Tf_Idf.
- Random Forest (number of trees are 100) with Tf_Idf.
- Decision Tree with Tf_Idf.
- Logistic Regression with Tf_Idf.

Deep Learning

For deep learning, we split the data into 80% training, 10% validation, and 10% testing. We created several models using neural networks: CNN, LSTM, and BILSTM.

First model: This model consists of six layers in the following order: embedding, conv1d, global max pooling, dense (64), dropout (0.5), and dense (1) as an output layer with a sigmoid function. The model was trained with a learning rate (lr = 0.001) using the Adam optimizer and techniques like early stopping and reduce learning rate. It was trained for ten epochs (epochs = 10) with a batch size of 32.

Second model: This model consists of three layers in the following order: embedding, LSTM (64 units with dropout = 0.2) and dense (1) as an output layer with a sigmoid function. The model was trained with a learning rate (lr = 0.001) using the Adam optimizer. It was trained for ten epochs (epochs = 10) with a batch size of 64.

Third model: This model consists of three layers in the following order: embedding, BILSTM (64 units with dropout = 0.2) and dense (1) as an output layer with a sigmoid function. The model was trained with a learning rate (lr = 0.001) using the Adam optimizer. It was trained for ten epochs (epochs = 10) with a batch size of 64. We used learning rate scheduler to reduce learning rate within training.

Fourth model: This model consists of eight layers in the following order: embedding, conv1d (256), max pooling (5), LSTM (128), BILSTM(64), dense (128), dropout (0.5), and dense (1) as an output layer with a sigmoid function. The model was trained with a learning rate (lr = 0.001) using the Adam optimizer and techniques like early stopping and reduce learning rate. It was trained for ten epochs (epochs = 10) with a batch size of 32.

The last model: This model consists of three layers in the following order: embedding, BILSTM (128 units with dropout = 0.2) and dense (1) as an output layer with a sigmoid function. The model was trained with a learning rate (lr = 0.001) using the Adam optimizer. It was trained for ten epochs (epochs = 10) with a batch size of 32. We used learning rate scheduler to reduce learning rate within training.

Results and Discussion

ML results

We will show the results of evaluating the algorithms on the test data. In Table 2 we will show the results of the machine learning algorithms.

Table 2 the results of ML algorithms

model	F1-Score (%)
SVM (kernel 'linear', C=1) with Tf_Idf.	81.9
SVM (kernel 'linear', C=1) with W2V and Ngrams.	81.8
SVM (kernel 'rbf', C=0.5) with W2V and Ngrams	82.5
KNN (k=5) with Tf_Idf	54.8
Random Forest (number of trees are 100) with Tf_Idf	80.9
Decision Tree with Tf_Idf	73.1
Logistic Regression with Tf_Idf	82.1

Figure 7 shows a comparison of machine learning algorithms based on the F1_Score value of the test data.

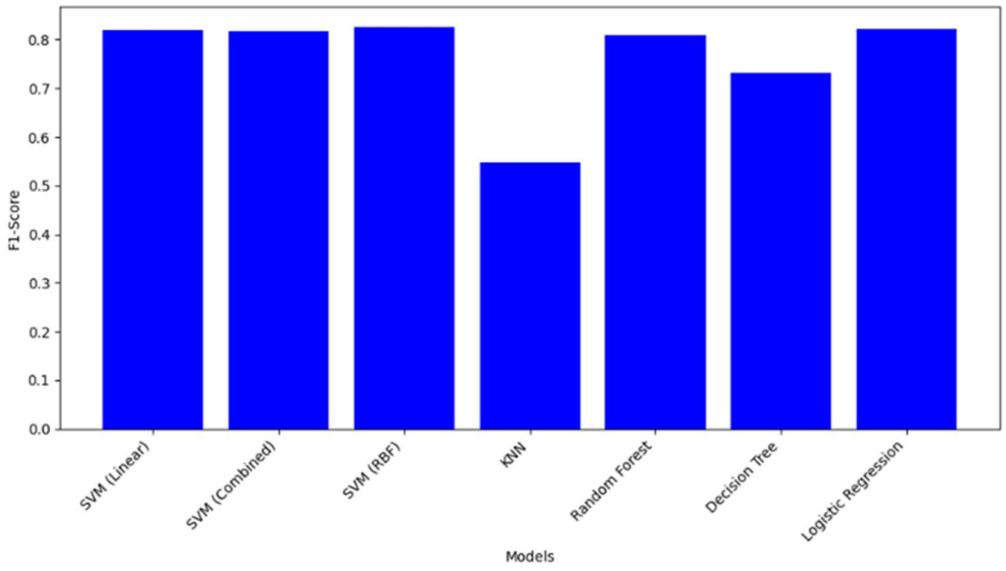

Figure 7 comparison of ML algorithms based on the F1_Score

The best model utilized a Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel and a C value of 0.5. This combination allowed the model to effectively capture

complex patterns in the data. Additionally, it incorporated Word2Vec and Ngrams, which improved its ability to understand the nuances of Arabic text, resulting in a high F1-Score. the worst-performing model used k-Nearest Neighbors (KNN) with k=5 and Tf-Idf features. KNN tends to perform poorly when dealing with high-dimensional data and may struggle to generalize patterns effectively, leading to its lower F1-Score.

DL results

We will compare the performance of deep learning models on the test data. Table 3 shows the results of deep learning models. We will only write the model number (if you want to know more details, you can refer to the second part of section five).

Note: the 1 refers to positive and -1 refers to negative.

Table 3 the results of DL Models

model	label	Precision	Recall	F1-Score (%)
First model	-1	81	86	83
	1	85	79	82
Second model	-1	81	86	83
	1	85	80	82
Third model	-1	88	84	86
	1	85	88	86
Fourth model	-1	85	86	85
	1	85	84	85
Last model	-1	88	84	86
	1	85	89	87

The Last model, which employs bidirectional context and effective sequential processing, outperformed other models in sentiment analysis with an F1-Score of 87%. This demonstrates that capturing contextual information from both directions in text is crucial for accurate sentiment classification. The model's simplicity and generalization ability make it a valuable choice for tasks like customer review sentiment analysis.

Conclusion

In this paper, we have undertaken an in-depth exploration of Classification reviews in Arabic. We analyzed Arabic reviews from Amazon using Machine Learning and Deep Learning, and we compared the results between them after discussing and comparing previous related work in the same field. Through this paper, We noticed AI's high analytics and classification capabilities, the power of deep neural networks compared to other machine learning algorithms. We discussed the huge impact of social media on the global market, and how customer reviews and opinions affect the sale of a particular product. In the end we noticed that using BILSTM networks the performance was better than all other models. In the future, we can improve performance further, analyze and understand customer reviews better by using transformers and other advanced techniques such as the attention mechanism. This paper opens doors to new opportunities for further investigation, improvement, and advancement in this context. It serves as a significant step towards Create deep models that are able to automatically

understand and classify customer reviews written in any language. and lays the groundwork for future research.

References

[1] A. Adak, B. Pradhan, and N. Shukla, "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review," Foods, vol. 11, no. 10, 2022, doi: 10.3390/foods11101500.

[2] Z. K. Chepukaka and F. K. Kirugi, "Service Quality and Customer Satisfaction At Kenya National Archives and Documentation Service, Nairobi County: Servqual Model Revisited," nternational Journal on Customer Relations, vol. 7, no. 1, 2019.

[3] J. D. Barsky and R. Labagh, "A Strategy for Customer Satisfaction," Cornell Hotel and Restaurant Administration Quarterly, vol. 33, no. 5, 1992, doi: 10.1177/001088049203300524.

[4] D. Suhartanto, M. Helmi Ali, K. H. Tan, F. Sjahroeddin, and L. Kusdibyo, "Loyalty toward online food delivery service: the role of e-service quality and food quality," Journal of Foodservice Business Research, vol. 22, no. 1, 2019, doi: 10.1080/15378020.2018.1546076.

[5] P. Poomka, N. Kerdprasop, and K. Kerdprasop, "Machine Learning Versus Deep Learning Performances on the Sentiment Analysis of Product Reviews," Int J Mach Learn Comput, vol. 11, no. 2, pp. 103–109, Mar. 2021, doi: 10.18178/ijmlc.2021.11.2.1021.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[7] G. Alhamad and M. B. Kurdy, "Feature-Based Sentiment Analysis for Arabic Language," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, pp. 455–462, 2020, doi: 10.14569/IJACSA.2020.0111158.

[8] D. S. Chauhan, S. R. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.401.

[9] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," Artif Intell Rev, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, doi: 10.1007/s10462-021-09973-3.

[10] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," IEEE Trans Affect Comput, vol. 5, no. 2, pp. 101–111, 2014, doi: 10.1109/TAFFC.2014.2317187.

[11] S.-M. Kim and E. Hovy, "Determining the Sentiment of Opinions."

[12] N. Guenther and M. Schonlau, "Support vector machines," Stata Journal, vol. 16, no. 4, 2016, doi: 10.1177/1536867x1601600407.

[13] S. Amarappa and S. V Sathyanarayana, "Data classification using Support vector Machine (SVM), a simplified approach." [Online]. Available: www.ijecse.org

[14] V. B. S. Prasath et al., "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review," Aug. 2017, doi: 10.1089/big.2018.0175.

[15] L. Breiman, "Random Forests," 2001.

[16] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, no. 1, pp. 46–50, Oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.
[17] M. Guia, R. R. Silva, and J. Bernardino, "Comparison of Naive Bayes, support vector machine, decision trees and random forest on sentiment analysis," in IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, SciTePress, 2019, pp. 525–531. doi: 10.5220/0008364105250531.
[18] Jouf University and Institute of Electrical and Electronics Engineers, 2019 International Conference on Computer and Information Sciences (ICCIS) : Jouf University - Aljouf - kingdom of Saudi Arabia, 03-04 April 2019.
[19] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," IEEE Access, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
[20] A. Adak, B. Pradhan, and N. Shukla, "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review," Foods, vol. 11, no. 10, May 2022, doi: 10.3390/foods11101500.
[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[22] V. D. Derbentsev, V. S. Bezkorovainyi, A. V Matviychuk, O. M. Pomazun, A. V Hrabariev, and A. M. Hostryk, "A comparative study of deep learning models for sentiment analysis of social media texts ⋆," 2023.
[23] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," IEEE Access, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.
[24] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews," J Sens, vol. 2022, 2022, doi: 10.1155/2022/7212366.
[25] A. Elouardighi, M. Maghfour, H. Hammia, and F.-Z. Aazi, "A Machine Learning Approach for Sentiment Analysis in the Standard or Dialectal Arabic Facebook Comments."
[26] M. Al-Ayyoub, A. Nuseir, G. Kanaan, and R. Al-Shalabi, "Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews." [Online]. Available: www.ijacsa.thesai.org
[27] Institute of Electrical and Electronics Engineers and Manav Rachna International Institute of Research and Studies, Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing : trends, prespectives and prospects : COMITCON-2019 : 14th-16th February, 2019.
[28] S. Alowaidi, M. Saleh, and O. Abulnaja, "Semantic Sentiment Analysis of Arabic Texts," 2017. [Online]. Available: www.ijacsa.thesai.org
[29] A. M. Mostafa, "An Evaluation of Sentiment Analysis and Classification Algorithms for Arabic Textual Data," 2017.
[30] IEEE Computational Intelligence Society, International Neural Network Society, and Institute of Electrical and Electronics Engineers, 2018 International Joint Conference on Neural Networks (IJCNN) : 2018 proceedings.

[31] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proceedings of the IEEE, vol. 105, no. 12. Institute of Electrical and Electronics Engineers Inc., pp. 2295–2329, Dec. 01, 2017. doi: 10.1109/JPROC.2017.2761740.

[32] M. S. Haydar, M. Al Helal, and S. A. Hossain, "Sentiment Extraction From Bangla Text : A Character Level Supervised Recurrent Neural Network Approach."

[33] Q. T. Ain et al., "Sentiment Analysis Using Deep Learning Techniques: A Review," 2017. [Online]. Available: www.ijacsa.thesai.org

[34] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, and M. O. Faruk, "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language," in ICCIT 2020 - 23rd International Conference on Computer and Information Technology, Proceedings, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICCIT51783.2020.9392733.

[35] B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O. N. Kushwaha, and N. K. Shah, "Sentiments Detection for Amazon Product Review," in 2021 International Conference on Computer Communication and Informatics, ICCCI 2021, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ICCCI50826.2021.9402414.

[36] N. Ali Hakami and H. A. Hosni Mahmoud, "Deep Learning Analysis for Reviews in Arabic E-Commerce Sites to Detect Consumer Behavior towards Sustainability," Sustainability (Switzerland), vol. 14, no. 19, Oct. 2022, doi: 10.3390/su141912860.

[37] A. Elnagar, L. Lulu, and O. Einea, "An Annotated Huge Dataset for Standard and Colloquial Arabic Reviews for Subjective Sentiment Analysis," in Procedia Computer Science, Elsevier B.V., 2018, pp. 182–189. doi: 10.1016/j.procs.2018.10.474.

[38] Abed Khooli, "arabic-100k-reviews," Mar. 07, 2020. https://datasetsearch.research.google.com/search?src=0&query=arabic-100kreviews&docid=L2cvMTFxOGJrYjJ6Zw%3D%3D (accessed Jun. 01, 2023).