Detecting Spam Reviews in Arabic by Deep Learning

Eman Aljadani
Department of Computer Science and Artificial Intelligence
University of Jeddah, Saudi Arabia
ealjadani.stu@uj.edu.sa

Fatmah Assiri
Software Engineering Department
University of Jeddah
Saudi Arabia
fyassiri@uj.edu.sa

Areej Alshutayri
Department of Computer Science and Artificial Intelligence
University of Jeddah, Saudi Arabia
aoalshutayri@uj.edu.sa

Abstract: *Online reviews are frequently used by consumers to make decisions about online purchases, hotel bookings, car rentals, and other choices because online shopping has grown in popularity over the past few years. Reviews are now crucial to both the customer and the business. As writing fake reviews comes with financial gain, opinion spam activities have increased. Some unethical companies may hire workers to write reviews to influence consumers' purchasing decisions; therefore, detecting spam reviews is a very important task. We compiled a large dataset of Arabic reviews consisting of spam and non-spam that are categorized by crowd-sourcing approach. Then, we applied deep learning algorithms to detect spam reviews. To the best of our knowledge, there are no prior studies utilized deep learning to classify reviews that are written in Arabic. Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) models were used, and an accuracy of 97% was achieved by both algorithms. To further improve the results, unbalanced issues were solved by oversampling and undersampling techniques. The results of them are improvements in the precision, recall, and F1-score for spam reviews. For example, in CNN F1-score for spam class increased from 79% to 90% with undersampling and became 82% with oversampling.*

Keywords: *Spam reviews, spam reviews detection, arabic language, deep learning, convolutional neural network, bidirectional long short-term memory.*

Received December 7, 2023; accepted May 4, 2024
https://doi.org/10.34028/iajit/21/3/12

1. Introduction

Following the emergence of the COVID-19 pandemic in December 2019, the practice of online shopping in Saudi Arabia, like in many other nations across the globe, has experienced an unprecedented surge in popularity. It has become the best option for customers; e-commerce websites such as Namshi, Amazon, Shein, Noon, and others have been widely used to fulfill customers' needs due to the movement restrictions and home quarantines implemented in some countries, such as the movement restrictions in Saudi Arabia in March 2020. Due to this, online payments significantly increased by 15% in March 2020 compared to February of the same year and by 239% compared to March 2019 [2].

Accordingly, the impact of online reviews increases daily; reviews can be an important factor since they influence purchase decisions for other customers. People often read product reviews to decide whether to buy the product or not [38]. Therefore, reviews bring significant financial gains or losses for businesses, organizations, and individuals. In other words, positive reviews attract more customers for a particular product or brand. This makes some businesses pay imposters to promote their company to attract new customers or demote competent companies in the same field.

The term "opinion spam" was first used by Jindal and Liu, who also identified three categories of reviews: untruthful reviews, which are those that are not based on consumers' actual usage of the products or services; rather, they are written with hidden motives. Reviews on brands only, which are related to the brand, producer, or seller of the products but do not include comments for the product being reviewed, are the second type. Because they do not specifically mention the product being reviewed, these reviews are regarded as spam. The final category is called a non-review, which includes advertisements and other useless reviews free of opinion, like questions, answers, and random text [21]. Our study will focus on the second and third types of spam reviews.

As previously mentioned, spam reviews can have a negative impact on the online marketplace; therefore, developing methods to help businesses and consumers distinguish truthful reviews from spam reviews is crucial. Machine learning is commonly used in this area, particularly supervised learning techniques [36]. Some studies detect spam reviews that are written in Arabic using machine learning models including Naive Bayes (NB) and Support Vector Machine (SVM) [3, 37, 46]. However, the majority of studies were attentive to detecting spam reviews for English text using machine learning, such as decision tree, NB, SVM, random forest, K-Nearest Neighbor (KNN), and logistic regression [27, 31, 39] and others used deep learning algorithms to detect spam reviews in English, such as Multilayer Perceptron (MLP), Convolutional Neural Network

(CNN) and Long Short-Term Memory (LSTM) [5, 20, 27, 40], to improve spam detection.

Deep learning has recently attracted more attention; deep learning was distinguished from classical machine learning by its ability to extract features from datasets without human intervention. Text review classification benefited from these architectures due to their ability to achieve high accuracy with less engineered features [25]. Shahariar *et al*. [40] compared the performance of both traditional machine learning and deep learning models to detect spam reviews in English. They found that deep learning classifiers performed better and achieved higher accuracy than traditional classifiers such as SVM, KNN, and NB. In addition, pretrained transformer-based language models, such as Bidirectional Encoder Representations From Transformers (BERT) for English language representation [9], obtained better vector representations for words and improved detection accuracy. BERT was developed by Google and has recently been pretrained for the Arabic language, such as MARBERT and ARABERT [1]. It proved to be efficient in different natural language processing domains and significantly better than AraBERT [4], which is the current best performing Arabic pretrained Language Model (LM). However, to the best of our knowledge, there is no research that utilizes deep learning to detect spam reviews written in Arabic.

In this work, we use Recurrent Neural Networks (RNNs) and CNNs to detect spam reviews in Arabic. According to previous research that used deep learning with the English language in the same research area, these models are most efficient and obtain an acceptable accuracy [5, 40, 44]. The main contributions of this paper are:

- Compiling a large Arabic reviews dataset.
- Utilizing deep learning algorithms to detect spam reviews.
- Applying undersampling and oversampling techniques to improve the prediction.

The remainder of the paper is structured as follows: Section 2 offers an overview of prior research on detecting spam reviews. In section 3, we outline the proposed approach, while section 4 details the experimental outcomes and ensuing discussions. Ultimately, in section 5, the paper is concluded, and we outline potential avenues for future research.

2. Literature Review

Due to the widespread use of online stores, there was a need to detect reviews since they affect buyers' decisions. Arabic review spam detection has received very little attention from research studies. As a result, this section provides a summary of earlier research that identifies spam reviews in both Arabic and English. Some Arabic spam reviews detection methods have been applied.

Hammad and El-Halees [16] conducted a study aimed at identifying an effective technique for identifying spam reviews sourced from TripAdvisor, Booking, and Agoda. Their approach combined data mining and text mining, employing classifiers such as NB, SVM, and K-NN. The NB classifier yielded the highest accuracy, achieving an impressive 99.2% accuracy rate.

Four different approaches to identifying spam in Arabic text were presented by Saeed *et al*. [37] They employed content-based features that rely on handling negations and n-grams. Rule-based classifiers, machine learning classifiers, majority voting ensembles, and stacking ensembles-which combine K-means and rule-based classifiers were all used to categorize reviews. The experiments revealed that the stacking ensemble outperformed other classifiers in a promising manner. It achieved 95.25% and 99.98% accuracy values for two datasets.

Using a combination of content-and user-based features, Mataoui *et al*. [28] proposed a model to identify spam reviews on social media. These features included comment size, number of hashtags, number of lines, number of emoticons, existence of specific sequences, repetition frequency of a comment, similarity between post, and comment topics. Several classifiers were used under WEKA software to determine that the best models were NB, J48, Sequential Minimal Optimization (SMO), logistic regression classifier, decision table, and locally weighted learning. The best accuracy achieved with J48 was 91.73%.

Deep learning has been used with the English language by some researchers to enhance the spam detection process. Shahariar *et al*. [40] compared the accuracy of deep learning and machine learning. They applied some traditional machine learning classifiers such as NB, KNN, and SVM. They also applied some methods in deep learning, including MLP, CNN, and LSTM. Two datasets were used: one is a labeled dataset from Ott *et al*. [33], and the other is an unlabeled dataset from Yelp, which was labeled using active learning. To develop machine learning classifiers, two features were used: Term Frequency-Inverse Document Frequency (TF-IDF) and n-grams. On the other hand, to generate word embeddings for deep learning, the TF-IDF algorithm with MLP and word2Vec with CNN and LSTM were applied. By the end, LSTM gave the best accuracy: 94.565% and 96.75% for the Ott and Yelp Datasets, respectively.

Archchitha and Charles [5] designed a CNN model that employed pretrained GloVe for Word Representation to identify spam reviews. They conducted a comparative analysis with traditional learning models. For their experimentation, they relied on a labeled dataset obtained from Ott *et al*. [33], which contained 1600 reviews. In their setup, 75% of the dataset was allocated for training, while the remaining 25% was used for testing. The sentences were tokenized by splitting them into lists of words, which were then

associated with 300-dimensional pretrained GloVe word embeddings. CNN models were built with a combination of word embedding and text-based features, including TF-IDF and count vectors, to analyze the effect of various types of features. The results found that the CNN outperformed traditional approaches with an accuracy of 86.25%, and by adding text-based features, the accuracy improved further to reach more than 88%.

The literature shows that there have been a great efforts to detect spam reviews using machine learning and deep learning. However, to the best of our knowledge, there is no prior work utilizing the strength of deep learning to detect spam reviews written in Arabic. Thus, this study investigates the use of deep learning algorithms to build accurate prediction models to detect spam reviews written in Arabic.

3. Spam Reviews Detection

Figure 1 describes our approach to build a detection model for Arabic text. Five steps will be further described in the following sections: data acquisition, data labeling, preprocessing, feature representation, and spam detection model.

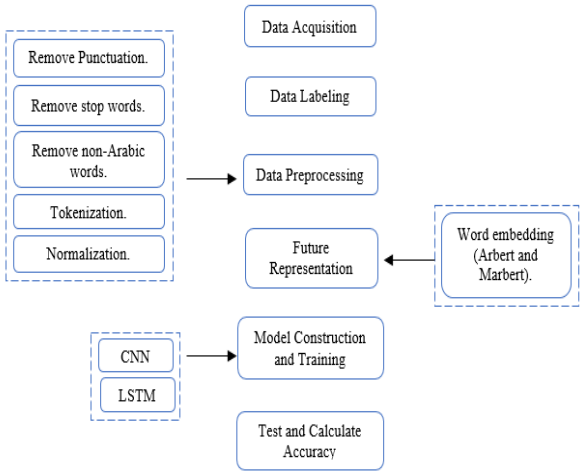

Figure 1. Overall process for the spam reviews detection model.

3.1. Data Acquisition

There was no labeled Arabic dataset for spam reviews; the only public Arabic dataset labeled positive and negative is the Hotel Arabic Reviews Dataset (HARD), which was created by Elnagar *et al*. [11]. It was collected from Booking and consists of 93700 reviews. To be able to use this dataset, data must be labeled as spam and non-spam. Since only customers who have made a reservation and stayed at the hotels are permitted to leave reviews, we have observed that there are no spam reviews and that all of the reviews are genuine in the labelling stage. Booking websites have the ability to screen all guest reviews for offensive language and authenticity before adding them to the website, remove any reviews that are irrelevant or against their policies, and continuously verify that reviews are true and trustworthy.

Therefore, due to the need for a specific dataset that consists of both spam and non-spam reviews, we created an Arabic review dataset from the Shein online store that sells clothes, bags, accessories, shoes, and others. Shein gives points to a customer who writes reviews on any product they purchased. Some people write unrelated text only to receive points even before receiving the shipment. We used web scraping, Selenium and Parsel packages in Python to collect reviews. The selenium package is used for performing HTTP requests, and Parsel is used for handling all HTML processing. Finally, a filter was implemented to remove redundancy from the data. Our dataset contains approximately 107K reviews in total that are written in Arabic.

3.2. Data Labeling

The reviews were categorized (spam or non-spam) manually by using crowdsourcing. To avoid any bias, six people were selected, and 5000 reviews were labeled and reviewed by us. The set of rules was defined and explained to the labeling team to be followed.

Rules were defined depending on two types of spam reviews that were used to detect Arabic spam reviews [16]: general reviews and nonreviews. However, untruthful review was not used since it is difficult to detect by humans and thus will not be able to label them. Additionally, a new rule, which is the existence of repeating words, was defined based on our analysis for reviews. Some examples of spam reviews are shown in Table 1. The following are the identified rules:

- *General review*: these are reviews that pertain to the store or the brand rather than the specific product. While they may be authentic, they are categorized as spam because they lack relevance to the individual products and often exhibit bias.
- *Nonreviews*: this category encompasses advertisements and other irrelevant reviews that do not offer any opinions. This includes questions, answers, requests for likes, or any other unrelated text.
- *Repeating words*: the review consists of repeated words.

Table 1. Examples of spam reviews following the defined rules.

Spam Type		Example
General review		كل المنتجات في الموقع جميلة وجودتها عالية.
Nonreview	Advertisements	طلبية زبونه بجيب اي حاجه من شي ان لمصر باقل تكلفه وبدون اي جمارك حبايبي كلموني.
	Random text	استنننننننننتتتتتت
	Request like	عطوني لايك محتاجة نقاط
	Questions	متى توصل الشحنة؟
Repeating one word		حلو حلو حلو حلو حلو

After finishing labeling the first set of reviews, labels were evaluated, and the best three annotators were selected to complete labeling of the rest of the data. This means that we will have three labels for each record from

three different annotators. The majority voting has been applied to decide the final label for each review. At the end of this phase, the dataset contains 100K non-spam reviews and 7K spam reviews.

Data class imbalances occur when one class (typically the non-spam class) is significantly overrepresented with numerous examples, while the other class (usually the spam class) has only a few instances. In such situations, the classifier tends to predict the majority class and often disregards the minority class. To mitigate this problem, sampling techniques like oversampling and undersampling can be employed to equalize class proportions within the training dataset. These methods adjust the distributions of both majority and minority classes, ensuring a more balanced representation of instances in each category [35]. All data are made available in GitHub upon request.

3.3. Data Preprocessing

One of the important steps for Natural Language Processing (NLP) tasks to obtain good accuracy is data preprocessing. Data preprocessing is an approach for cleaning and preparing text data to keep only the words that will impact the prediction goal. Some methods were applied as follows:

- Remove punctuations: removing all punctuation marks such as: [,./;'!#$%^&*":?<>_-()].
- Remove repeated characters: In this step, any repeating characters will be removed, such as (جمييييييل), which will be (جميل).
- Remove Arabic diacritics such as (Fatha ً, Kasrah ٍ, Dhama ٌ, tanween ّ ْ, and others).
- Remove emojis.
- Remove stop words such as (من ، في ، حتى ، إلى ، على ، عن ، لكن ، الا).
- Remove English text: is to remove any English token in the review.
- Tokenization: splits the review into a sequence of a single word for each word called a token.
- Normalization converts the different forms of the word into a common form by transforms each letter to its specified standard form, as shown in Table 2 below.

Table 2. Arabic text normalization.

Letters to replace	Replaced with
ى، ئ، ئ	ي
آ، أ، إ، ٱ، أ	ا
ة	ه
ؤ، ۇ، ۈ	و

3.4. Feature Representation

The process of manually extracting features in NLP applications for Arabic text is quite demanding due to the language's intricate structure and rich morphology, as noted by [29]. In the realm of NLP, representing words as continuous vectors in a multidimensional space has evolved into a crucial step for feature extraction in text data [45]. In the early stages, pretrained text representation models aimed to depict words by capturing their distributed syntactic and semantic attributes. This was achieved through techniques such as Word2vec [30] and GloVe [34]. However, these models did not incorporate the context into its embedding; there is just one vector representation for each word. Different meanings of the word (if any) are combined into one single vector.

The utilization of pretrained contextualized text representation models has driven substantial progress in enhancing the comprehension of natural language understanding, leading to state-of-the-art performance across various NLP tasks [8, 19]. One of them is BERT [9]. BERT stands for Bidirectional Encoder Representations from Transformers. It is NLP model that has undergone extensive training on a substantial volume of data. Rather than employing conventional word tokenization methods, it focuses on interpreting the meanings of words. BERT learns information from text from the left and right sides; BERT is for English language, mBERT multi language and AraBERT for Arabic language [4].

In this work, two more powerful transformer-based language models for Arabic are utilized, i.e., ARBERT and MARBERT [1]. These models are significantly better than AraBERT [4]. ARBERT stands as a substantial pretrained masked language model with a specific emphasis on Modern Standard Arabic. Its training data includes an extensive collection of Arabic datasets, encompassing 61 gigabytes of text, equivalent to 6.2 billion tokens. ARBERT adopts the architecture of BERT-base, which encompasses 12 attention layers, each equipped with 12 attention heads, and a hidden dimensionality of 768. MARBERT focuses on both dialectal Arabic and Modern Standard Arabic; it was trained on Arabic tweets, and the dataset makes up 128 GB of text (15.6 B tokens) [1].

3.5. Spam Detection Model

To detect spam reviews, we used two deep learning models: convolutional neural network and bidirectional long short-term recurrent neural network. According to previous research that used deep learning with Arabic in sentiment analysis [10, 17, 32] and other research in the English language in the same research area [5, 14, 40, 44], these models are most efficient and obtain an acceptable accuracy.

3.5.1. Convolutional Neural Network (CNN)

CNNs represent deep learning models primarily applied in computer vision, often for tasks like image classification. In certain instances, they find utility in natural language processing, including text classification. A CNN typically comprises a

convolutional layer, a pooling layer, and a fully connected layer.

The pivotal elements in CNNs are the convolutional layers, serving as the network's fundamental components. These layers employ filters to generate feature maps and distill the detected features from the input. This process entails applying a small numerical matrix (referred to as a kernel or filter) across the input matrix, transforming it based on the filter values. Subsequently, the results undergo normalization via a nonlinear activation function. This normalization process enables the model to learn complex patterns while mitigating the risk of gradient vanishing and maintaining computational efficiency. The pooling layer further contributes by reducing the feature map's dimension. It achieves this by partitioning the map into smaller segments and selecting the maximum value (max pooling) from each segment. This layer simplifies the network's complexity, enhancing CNN efficiency. The fully connected layer, on the other hand, functions as a multilayer perceptron linked to all activations from previous layers. Neuron activations are computed by matrix multiplication with their respective weights, augmented by an offset value, as described by [13]. In the end, the classification layer performs classification based on the attributes extracted by the previous layers. We adopted the CNN hyperparameter configuration that is utilized for sentiment prediction in Arabic tweets [17], as shown in Table 3. They systematically adjusted each hyperparameter value individually and computed accuracy and F1-score until achieving the best possible result.

Table 3. CNN model hyperparameter configuration.

Hyperparameters	Value
Filter sizes	[3, 4, 5]
Number of filters	100
Dropout rate	0.5
Learning rate	0.0001
Number of epochs	10
Batch size	50

3.5.2. Bidirectional Long Short-Term Memory Networks (BiLSTM)

LSTM represents a distinct category of RNNs designed to grasp extended patterns of information. Hochreiter and Schmidhuber introduced LSTMs in 1997 [18]. Typically, an LSTM unit is composed of a memory cell, an input, output, and forget gates, as illustrated in Figure 2. The memory cell's role is to retain information across time intervals, while the other gates regulate the input and output of data from the cell.

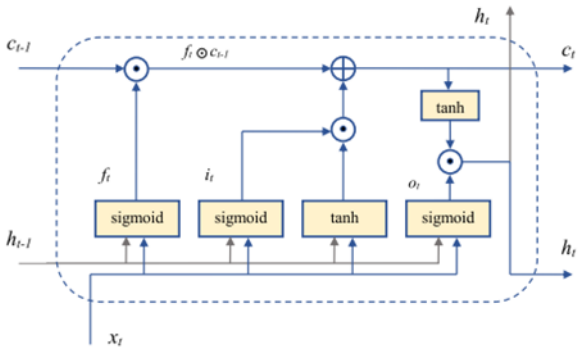

Figure 2. An illustrative block diagram of the LSTM network [15].

The forget gate (f_t) is responsible for determining whether to retain or discard the information from the previous state (c_{t-1}), and this decision is based on the values of the input (x_t) and the hidden state (h_{t-1}). The output of the forget gate can assume a value of 0 or 1. Similarly, the input gate (i_t) plays a role in deciding the extent to which the information from the input text (x_t) and the hidden state (h_{t-1}) should be allowed to update the cell state. The value of c_t represents the resulting cell state, which is computed through mathematical operations involving c_{t-1}, f_t, and i_t. The output gate (o_t) controls the transfer of information from the current cell state to the hidden state. The mathematical expressions for these gates are given in Equations (1) to (5). In the context of LSTM, the inputs at any given time (t) consist of the input vector (x_t), the previously hidden state (h_{t-1}), and the previous cell state (c_{t-1}). Conversely, the outputs include the current hidden state (h_t) and the current cell state (c_t). The symbol $\odot$ denotes elementwise vector multiplication.

$$f_t = sigmoid\left(W_{f_x} x_t + W_{f_h} h_{t-1} + b_f\right) \quad (1)$$

$$i_t = sigmoid\left(W_{i_x} x_t + W_{i_h} h_{t-1} + b_i\right) \quad (2)$$

$$c_t = c_{t-1} \odot f_t + i_t \odot tanh\left(W_{c_x} x_t + W_{c_h} h_{t-1} + b_c\right) \quad (3)$$

$$o_t = sigmoid\left(W_{o_x} x_t + W_{o_h} h_{t-1} + b_o\right) \quad (4)$$

$$h_t = o_t \odot \tanh(c_t) \quad (5)$$

In the LSTM model, information flows strictly in a unidirectional manner, meaning that the state at time 't' relies solely on preceding information. However, to capture the full semantic context of an input review, the following information is just as important as the previous information. To achieve a more comprehensive representation of contextual information, the BiLSTMmodel was introduced [17]. The BiLSTM model consists of two LSTM networks that can process input reviews in both forward and backward directions. We adopted the LSTM hyperparameter configuration utilized for sentiment prediction in Arabic tweets, which is detailed in Table 4 from the work by Heikal *et al*. [17]. They systematically adjusted each hyperparameter value individually and computed accuracy and F1-score until achieving the best possible result.

Table 4. LSTM model hyperparameter configuration.

Hyperparameters	Value
LSTM hidden state dimension	200
Dropout rate	0.2
Learning rate	0.001
Number of epochs	10
Batch size	50

4. Experiment Results and Evaluation

This study investigates the following research questions:

- RQ1: Which word embeddings are better (ArBERT) or (MarBERT) with our dataset?
- RQ2: Which deep learning model performance is best in detecting Arabic spam reviews?
- RQ3: Does solving the issue of data imbalance improve model performance?

All of the experiments were implemented using Google Colaboratory tools (Colab) [7]. Colab is a project by the Google Research Lab. It is a Linux machine, and its interface is based on the Jupyter notebook service. It provides free access to recent computing resources such as Graphical Processing Units (GPUs).

4.1. Evaluation Measures

This study employed a confusion matrix to assess the performance of classifiers on a test dataset [41]. This matrix is essentially a two-dimensional arrangement, illustrated in Table 5, and yields four distinct values:

- True Positive (TP): This category represents cases where both the predicted and actual labels are identified as spam.
- True Negative (TN): This category correspond to the case where both the predicted and actual labels are recognized as non-spam.
- False Positive (FP): It arises when the predicted label is classified as spam, despite the actual label being non-spam.
- False Negative (FN): It occurs when the predicted label designates a non-spam category while the actual label is spam.

Table 5. Confusion matrix.

Actual values \ Predicted values	Non-spam	Spam
Non-spam	TN	FP
Spam	FN	TP

Using the information from the confusion matrix, we computed precision, recall, accuracy, and F1 score using Equations (6) to (9).

- Precision (P): it is calculated by dividing the number of positive predictions by the total number of positive class values predicted.
- Recall (R): it is determined by dividing the number of positive predictions by the number of positive class values in the test data.
- Accuracy (A): it is defined as the degree of correctness of a quantity or expression.
- F1 score: it represents a balance between precision and recall.

$$P = TP/(TP+FP) \quad (6)$$

$$R = TP/(TP+FN) \quad (7)$$

$$A = (TP+TN)/(TP+FP+FN+TN) \quad (8)$$

$$F1\ score = (2*P*R)/(P+R) \quad (9)$$

4.2. Experimental Results

In this section, three experiments have been conducted to answer the research questions. Experiment I evaluated ArBERT and MarBERT as two word embeddings that are well known for Arabic language. Experiment II compares the performance of BiLSTM and CNN. Experiment III provides the results of oversampling and undersampling techniques as techniques used to solve the data imbalance issue.

4.2.1. Experiment 1

First, we build two models CNN and BiLSTM to compare two word embeddings that have recently been used with Arabic ArBERT and MarBERT. As shown in Tables 6 and 7, the results of ArBERT and MarBERT are almost the same in both models, but MarBERT exceeded ArBERT with a small difference. The accuracy for CNN with MarBERT was 97.47%, which is approximately 0.09% higher than that of ArBERT. Moreover, BiLSTM with MarBERT was 97.35% and BiLSTM with ArBERT was 97.21%, MarBERT is approximately 0.14% higher than that of ArBERT. Additionally, the precision, recall and F1-score of spam and non-spam reviews improved compared to ArBERT. This could be due to the data used to train MarBERT, which was tweets that used both dialectal Arabic and Modern Standard Arabic. On the other hand, ArBERT focused only on Modern Standard Arabic. Figures 3, 4, 7, 8 show how the accuracy improved over epochs.

Table 6. Performance of CNN with ArBERT and MarBERT to detect arabic spam reviews.

	Precision		Recall		F1-score			Accuracy
	Spam	Non-Spam	Spam	Non-Spam	Spam	Non-Spam	Average	
CNN with (ArBERT)	87%	98%	72%	99%	78%	99%	89%	97.38%
CNN with (MarBERT)	87%	98%	73%	99%	79%	99%	89%	97.47%

Table 7. Performance of BiLSTM with ArBERT and MarBERT to detect arabic spam reviews.

	Precision		Recall		F1-score			Accuracy
	Spam	Non-Spam	Spam	Non-Spam	Spam	Non-Spam	Average	
BiLSTM with (ArBERT)	81%	98%	76%	99%	78%	99%	88.5%	97.21%
BiLSTM with (MarBERT)	82%	98%	77%	99%	79%	99%	89.04%	97.35%

4.2.2. Experiment 2

In the second experiment, two classification algorithms are used: CNN and BiLSTM. The MarBERT work embedding model was used with both since it showed better performance from the previous experiment. As seen in Table 8, the results of both algorithms are very close, with a slight superiority of the CNN algorithm, as the accuracy is 97.47%, while the BiLSTM accuracy is 97.35%. Although CNNs are generally used in computer vision [6], they have been applied to various NLP tasks in English such as sentiment analysis and topic categorization [12, 22, 23, 43, 44] and the results were promising. Figures 4 and 8 show improved accuracy over epochs, which supports the findings in the literature.

Table 8. Performance of BiLSTM and CNN.

	Precision		Recall		F1-score			Accuracy
	Spam	Non-Spam	Spam	Non-Spam	Spam	Non-Spam	Average	
CNN	87%	98%	73%	99%	79%	99%	89%	97.47%
BiLSTM	82%	98%	77%	99%	79%	99%	89%	97.35%

4.2.3. Experiment 3

The dataset used in this paper is unbalanced, it contains only 7K spam reviews and 100K non-spam reviews. The unbalance in datasets creates a challenge for learning algorithms as they tend to be biased towards the majority group. Although usually, the minority class is more important, despite its rarity. It may contain valuable insights and knowledge that are essential for accurate analysis and classification [24].

To solve the unbalanced issue, sampling techniques such as oversampling or undersampling are used to create more balanced classes in the dataset, which leads to increased detection of spam reviews. The undersampling approach aims to decrease the number of samples from the majority class to achieve a more balanced class distribution. This technique involves reducing the size of the majority class by selecting the same number of instances as found in the minority class. By doing so, the skewed distribution between the majority and minority classes can be alleviated [42]. However, it has notable limitations. First, it involves removing instances from the majority class, potentially leading to a loss of valuable information and affecting the model's ability to learn essential patterns. Moreover, the random nature of undersampling can introduce bias into the training data, impacting the model's generalization performance. Additionally, the reduction in the size of the training set may limit the model's capacity to learn complex patterns and increase the risk of overfitting [26].

The second technique is oversampling, which involves increasing the samples of the minority class and adding them to the dataset. This method differs from the undersampling approach in that no information is lost, as all instances are employed [42]. Thus, new reviews were collected from the Shein website by web scraping, and spam reviews were selected manually. The total number of spam reviews increased to 14K reviews; it was a difficult task and time-consuming since spam reviews were very scarce, so it was hard to reach 100K spam reviews. The classwise distribution is shown in Table 9.

Table 9. Dataset distribution.

class	Base dataset	Undersampling	Oversampling
No. of Spam	7000	7000	14000
No. of Non spam	100K	7000	100K
Total	107K	14000	114K

The results of oversampling and undersampling for both the CNN and BiLSTM models are shown in Tables 10 and 11. There are improvements in the precision, recall, and F1-score for spam reviews in the two models, which is the aim of this experiment. For example, the precision in BiLSTM was 82%, then after undersampling, it became 93%, and after oversampling, it became 88%. However, in undersampling, the precision, recall, and F1-score for non-spam reviews were reduced as the number of non-spam reviews decreased from 100K to 7K, which led to a decrease in the accuracy of the model. In addition, Oversampling improves the results for detecting spam reviews but does not have a negative impact on detecting non-spam reviews, unlike undersampling. For example, in CNN, the F1-score for spam reviews was 78%, then after oversampling, it became 82%. It is noteworthy that the F1-score for non-spam reviews remains almost the same before and after oversampling. Figures 5, 6, 9, and 10 show how the accuracy improves for BiLSTM and CNN with oversampling and undersampling.

Table 10. Performance of BiLSTM with undersampling and oversampling to detect Arabic spam reviews.

	Precision		Recall		F1-score			Accuracy
	Spam	Non Spam	Spam	Non Spam	Spam	Non Spam	Average	
BiLSTM base dataset	82%	98%	77%	99%	79%	99%	89.04%	97.35%
BiLSTM with under-Sample	93%	91%	91%	93%	92%	92%	91.93%	91.93%
BiLSTM with over-Sample	88%	98%	82%	89%	85%	98%	91.4%	96.43%

Table 11. Performance of CNN with undersampling and oversampling to detect Arabic spam reviews.

	Precision		Recall		F1-score			Accuracy
	Spam	Non Spam	Spam	Non Spam	Spam	Non Spam	Average	
CNN base dataset	87%	98%	73%	99%	79%	99%	89%	**97.47%**
CNN with under-Sample	91%	90%	89%	91%	90%	90%	90.39%	90.39%
CNN with over-Sample	86%	97%	78%	98%	82%	98%	89.96%	95.87%

After completing our experiments, we discovered that deep learning techniques, specifically CNN and LSTM models, equipped with pre-trained word embeddings such as MarBERT, achieved good accuracy

levels. Moreover, the utilization of oversampling and undersampling techniques improved the results for detecting spam reviews. This suggests promising implications for the application of deep learning methods in effectively classifying textual data in Arabic, especially within online retail platforms. By efficiently identifying spam reviews, these methods can enhance the quality of the reviews section, ensuring a more authentic and enjoyable shopping experience for customers.

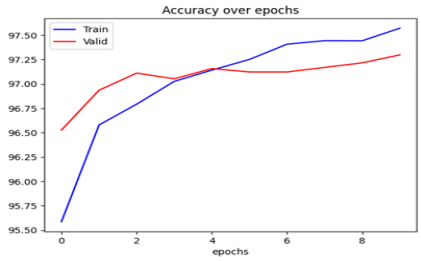

Figure 3. The accuracy for CNN on the training and validation set (Arbert).

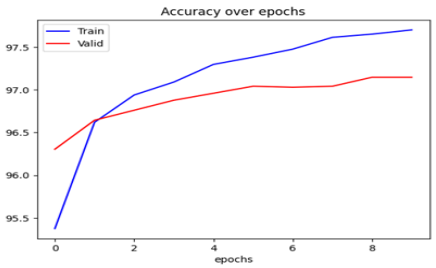

Figure 4. The accuracy for CNN on the training and validation set (Marbert).

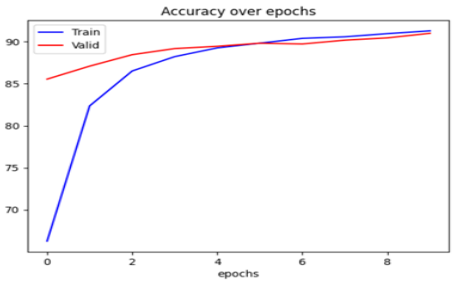

Figure 5. The accuracy for CNN on the training and validation set (undersampling).

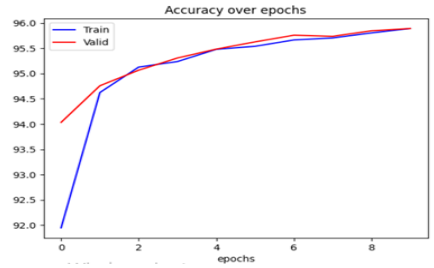

Figure 6. The accuracy for CNN on the training and validation set (oversampling).

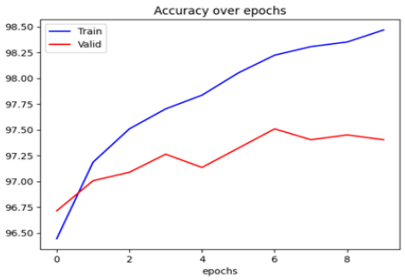

Figure 7. The accuracy for BiLSTM on the training and validation set (Arbert).

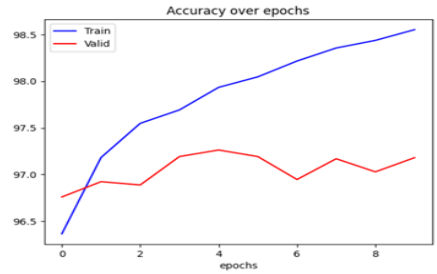

Figure 8. The accuracy for BiLSTM on the training and validation set (Marbert).

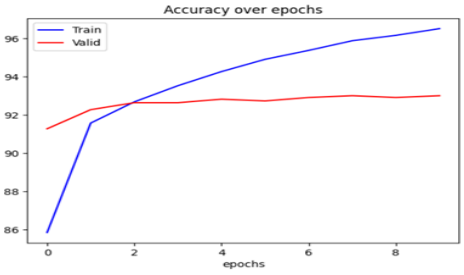

Figure 9. The accuracy for BiLSTM on the training and validation set (undersampling).

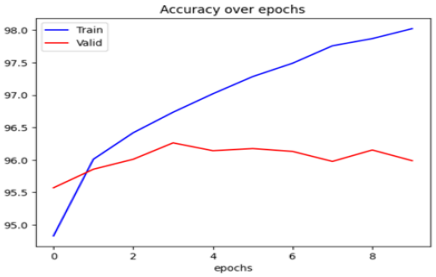

Figure 10. The accuracy for BiLSTM on the training and validation set (oversampling).

5. Conclusions

Deep learning methods typically address classification problems in a comprehensive manner. When it comes to text review classification in English, deep learning architectures have proven advantageous for their ability to attain high accuracy without heavy reliance on engineered features. In our research, we accomplished a notable classification accuracy exceeding 97%. However, for identifying spam reviews using deep learning algorithms, a considerably larger training dataset is needed compared to traditional machine learning techniques. Consequently, we assembled an extensive dataset of Arabic reviews, encompassing approximately 114,000 reviews.

We evaluated our approach with two models: CNN and BiLSTM. Pretrained word embedding MarBERT and ArBERT are also used to obtain better vector

representations for words and improve the accuracy of trained classifiers. The results for both algorithms are very close, but with a slight superiority of the CNN algorithm, as the accuracy is 97.47%, while the BiLSTM accuracy is 97.35%. The drawbacks of deep learning in our work are that only 7K spam reviews instances and 100K non-spam reviews were collected from Shein. Therefore, we used undersampling and oversampling approaches to solve the imbalanced class distribution problem, and they improved the result for detecting spam reviews, but in undersampling, the total accuracy was decreased by reducing the number of non-spam reviews from 100K to 7K.

In future research, it's possible to augment the quantity of spam reviews sourced from Shein. Moreover, our current focus has been exclusively on text reviews, with no consideration for review spammers. To enhance our study in the future, we can incorporate the detection of both review spam and review spammers. Furthermore, we can explore various CNN and RNN adaptations, as well as potentially introducing a hybrid CNN-RNN model.

References

[1] Abdul-Mageed M., Elmadany A., and Nagoudi E., "ARBERT and MARBERT: Deep Bidirectional Transformers for Arabic," *in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Virtual, pp. 7088-7105, 2021. https://aclanthology.org/2021.acl-long.551

[2] Abid M., Benlaria H., and Gheraia Z., "The Impact of the Emerging Coronavirus (COVID-19) on E-Commerce in the Kingdom of Saudi Arabia," *WSEAS Transaction on Bussiness and Economics*, vol. 19, pp. 825-836, 2022. DOI: 10.37394/23207.2022.19.72

[3] Abu-Hammad A. and El-Halees A., An Approach for Detecting Spam in Arabic Opinion, Master Thesis, Islamic University, 2013. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b350cd14f88e5848392b6417c731832679472eb8

[4] Antoun W., Baly F., and Hajj H., "AraBERT: Transformer-based Model for Arabic Language Understanding," *arXiv Preprint*, vol. arXiv:2003.00104v4, pp. 1-7, 2020. http://arxiv.org/abs/2003.00104

[5] Archchitha K. and Charles E., "Opinion Spam Detection in Online Reviews Using Neural Networks," *in Proceedings of the 19th International Conference on Advances in ICT for Emerging Regions*, Colombo, pp. 1-6, 2019. doi: 10.1109/ICTer48817.2019.9023695

[6] Bhatt D., Patel C., Talsania H., Patel J., and Vaghela R., "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope," *Electronics*, vol. 10, no. 20, pp. 1-28, 2021. https://doi.org/10.3390/electronics10202470

[7] Bisong E., *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, Apress, 2019. https://link.springer.com/book/10.1007/978-1-4842-4470-8

[8] Bourahouat G., Abourezq M., and Daoudi N., "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal on Information Technology*, vol. 21, no. 2, pp. 313-325, 2024. doi: 10.34028/21/2/13

[9] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, pp. 4171-4186, 2019. https://aclanthology.org/N19-1423

[10] Elfaik H. and Nfaoui E., "Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for Arabic Text," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 395-412, 2021. doi: 10.1515/jisys-2020-0021

[11] Elnagar A., Khalifa Y., and Einea A., *Intelligent Natural Language Processing: Trends and Applications*, Springer, 2018. https://www.springerprofessional.de/en/using-deep-neural-networks-for-extracting-sentiment-targets-in-a/15234276

[12] Fang H., Lu C., Hong F., Jiang W., and Wang T., "Convolutional Neural Network for Sentence Classification," *in Proceedings of the 15th IEEE International Conference on Electronic Measurement and Instruments*, Nanjing, pp. 253-258, 2021. doi: 10.1109/ICEMI52946.2021.9679581

[13] Guo T., Dong J., Li H., and Gao Y., "Simple Convolutional Neural Network on Image Classification," *IEEE 2nd International Conference on Big Data Analysis*, Beijing, pp. 721-724, 2017. doi: 10.1109/ICBDA.2017.8078730

[14] Hajek P. and Munk M., "Fake Consumer Review Detection Using Deep Neural Networks Integrating Word Embeddings and Emotion Mining," *Neural Computing and Applications*, vol. 2, pp. 17259-17274, 2020. https://doi.org/10.1007/s00521-020-04757-2

[15] Hameed Z. and Garcia-Zapirain B., "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access*, vol. 8, pp. 73992-74001, 2020. doi: 10.1109/ACCESS.2020.2988550

[16] Hammad A. and El-Halees A., "An Approach for Detecting Spam in Arabic Opinion Reviews," *The International Arab Journal of Information Technology*, vol. 12, no. 1, pp. 10-16, 2015. https://iajit.org/PDF/vol.12,no.1/7006.pdf

[17] Heikal M., Torki M., and El-Makky N., "Sentiment Analysis of Arabic Tweets Using Deep Learning," *Procedia Computer Science*, vol. 142, pp. 114-122, 2018. doi: 10.1016/j.procs.2018.10.466

[18] Hochreiter S. and Schmidhuber J., "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. doi: 10.1162/neco.1997.9.8.1735

[19] Howard J. and Ruder S., "Universal Language Model Fine-tuning for Text Classification," *arXiv Preprint*, vol. arXiv:1801.06146, pp. 1-12, 2018. https://arxiv.org/abs/1801.06146

[20] Jain G., Sharma M., and Agarwal B., "Spam Detection in Social Media Using Convolutional and Long Short Term Memory Neural Network," *Annals of Mathemetics and Artificial Intelligence*, vol. 85, no. 1, pp. 21-44, 2019. doi: 10.1007/s10472-018-9612-z

[21] Jindal N. and Liu B., "Opinion Spam and Analysis," *in Proceedings of the International Conference on Web Search Data Mining*, Palo Alto, pp. 219-230, 2008. https://doi.org/10.1145/1341531.1341560

[22] Johnson R. and Zhang T., "Semi-supervised Convolutional Neural Networks for Text Categorization Via Region Embedding," *Advances in Neural Information Processing* Systems, vol. 28, pp. 919-927, 2015. https://api.semanticscholar.org/CorpusID:168925 0

[23] Kalchbrenner N., Grefenstette E., and Blunsom P., "A Convolutional Neural Network for Modelling Sentences," *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, pp. 655-665, 2014. doi: 10.3115/v1/p14-1062

[24] Krawczyk B., "Learning from Imbalanced Data: Open Challenges and Future Directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221-232, 2016. doi: 10.1007/s13748-016-0094-0

[25] LeCun Y., Bengio Y., and Hinton G., "Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015. doi: 10.1038/nature14539

[26] Liu X., Wu J., and Zhou Z., "Exploratory Undersampling For Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 39, no. 2, pp. 539-550, 2009. doi: 10.1109/TSMCB.2008.2007853

[27] Mani S., Kumari S., Jain A., and Kumar P., "Spam Review Detection Using Ensemble Machine Learning," *in Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, New York, pp. 198-209, 2018. https://doi.org/10.1007/978-3-319-96133-0_15

[28] Mataoui M., Zelmati O., Boughaci D., Chaouche M., and Lagoug F., "A Proposed Spam Detection Approach For Arabic Social Networks Content," *in Proceedings of the International Conference on Mathematics and Information Technology*, Adrar, pp. 222-226 2017. doi: 10.1109/MATHIT.2017.8259721

[29] Mesleh A., "Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study," *in Proceedings of the Advances in Computer and Information Sciences and Engineering Conference*, Massachusetts, pp. 11-16, 2008. https://doi.org/10.1007/978-1-4020-8741-7_3

[30] Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J., *Advances in Neural Information Processing Systems*, Springer, 2013. https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[31] Narayan R., Rout J., and Jena S., *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Springer, 2018. DOI: 10.1007/978-981-10-3373-5

[32] Ombabi A., Ouarda W., and Alimi A., "Deep Learning CNN–LSTM Framework for Arabic Sentiment Analysis Using Textual Information Shared in Social Networks," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 53, 2020. doi: 10.1007/s13278-020-00668-1

[33] Ott M., Choi Y., Cardie C., and Hancock J., "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, pp. 309-319, 2011. https://aclanthology.org/P11-1032.pdf

[34] Pennington J., Socher R., and Manning C., "GloVe : Global Vectors for Word Representation," *in Proceedings of the Empirical Methods in Natural Language Processing Conference*, Doha, pp. 1532-1543, 2014. DOI: 10.3115/v1/D14-1162

[35] Phung S., Bouzerdoum A., and Nguyen G., *Pattern Recognition*, InTech, 2009. https://ro.uow.edu.au/infopapers/792

[36] Ren Y. and Ji D., "Learning to Detect Deceptive Opinion Spam: A Survey," *IEEE Access*, vol. 7, pp. 42934-42945, 2019. doi: 10.1109/ACCESS.2019.2908495

[37] Saeed R., Rady S., and Gharib T., "An Ensemble Approach for Spam Detection in Arabic Opinion Texts," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1407-1416, 2022. doi: 10.1016/j.jksuci.2019.10.002

[38] Samha A., Li Y., and Zhang J., "Apect-Based Opinion Extraction from Customer Reviews," *Computation and Language*, *arXiv Preprint*, vol. arXiv:1404.1982, pp. 149-160, 2014. https://doi.org/10.48550/arXiv.1404.1982

[39] Saumya S. and Singh J., "Detection of Spam Reviews: A Sentiment Analysis Approach," *CSI Transaction on ICT*, vol. 6, no. 2, pp. 137-148, 2018. doi: 10.1007/s40012-018-0193-0

[40] Shahariar G., Biswas S., Omar F., Shah F., and Hassan S., "Spam Review Detection Using Deep Learning," *in Proceedings of the 10th Annual Information Technology, Electronics and Mobile Communication Conference*, Vancouver, pp. 0027-0033, 2019. doi: 10.1109/IEMCON.2019.8936148

[41] Sokolova M. and Lapalme G., "A Systematic Analysis of Performance Measures for Classification Tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009. doi: 10.1016/j.ipm.2009.03.002

[42] Syah M., Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets, Master Thesis, The University of Texas at Austin, 2004. http://dx.doi.org/10.26153/tsw/12300

[43] Tammina S. and Annareddy S., "Sentiment Analysis on Customer Reviews Using Convolutional Neural Network," *in Proceedings of the International Conference on Computer Communication and Informatics*, Coimbatore, pp. 1-6, 2020. doi: 10.1109/ICCCI48352.2020.9104086

[44] Wang C., Day M., Chen C., and Liou J., "Detecting Spamming Reviews Using Long Short-Term Memory Recurrent Neural Network Framework," *in Proceedings of the 2nd International Conference on E-Commerce, E-Business and E-Government*, New York, pp. 16-20, 2018.doi: 10.1145/3234781.3234794

[45] Zahran M., Magooda A., Mahgoub A., Raafat H., and Rashwan M., "Word Representations in Vector Space and their Applications for Arabic," *in Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing*, Cairo, pp. 430-443, 2015. https://doi.org/10.1007/978-3-319-18111-0_32

[46] Ziani A., Azizi N., Schwab D., Zenakhra D., and Aldwairi M., "Deceptive Opinions Detection Using New Proposed Arabic Semantic Features," *Procedia Computer Science*, vol. 189, pp. 29-36, 2021. doi: 10.1016/j.procs.2021.05.067

Eman Aljadani, received a B.S. degree in Computer Engineering from King Abdulaziz University. She is currently pursuing a master's degree with the Computer Science and Artificial Intelligence Department at the University of Jeddah. Her research interests include deep learning and natural language processing.

Fatmah Assiri, an associate professor of Software Engineering and the former supervisor of Computer Science and Artificial intelligence department in the College of Computer Science and Engineering, University of Jeddah, KSA. I hold a Ph.D degree of Computer Science from Colorado State University, Colorado, United States. I served as a consultant for the entrepreneurship and innovation center at the University of Jeddah. My research interests are software testing and validation, automation, and I am currently interested in data and machine learning to develop smart solutions.

Areej Alshutayri, an Associate Professor in the department of computer science and Artificial Intelligence at the university of Jeddah. Areej collected and created a social media Arabic dialect text corpus (SMADC) using Twitter, Facebook, and Online newspapers. Areej's research interests in using Artificial Intelligence which includes machine learning and natural language processing to understand languages especially Arabic language and its dialects.