IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 27, No. 3, June, 2022, pp. 503-508
Journal homepage: http://iieta.org/journals/isi

Arabic Sentiment Analysis of Eateries' Reviews Using Deep Learning

Leen Muteb Alharbi[1*], Ali Mustafa Qamar[2]

[1] Department of Cyber Security, Onaizah Colleges, Al-Qassim, Saudi Arabia
[2] Department of Computer Science, College of Computer, Buraydah, Saudi Arabia

Corresponding Author Email: Leenmuteb@hotmail.com

https://doi.org/10.18280/isi.270318

Received: 21 May 2022
Accepted: 10 June 2022

Keywords:
sentiment analysis, machine learning, deep learning

ABSTRACT

After air and water, food is the third most essential thing for humans to provide energy and development. More and more customers of eateries, including restaurants and cafes, express their opinion or sentiment about quality, ambiance, and facilities. This research performs a sentiment analysis of the eateries' reviews obtained in Qassim, Saudi Arabia. The reviews are obtained in Arabic, the local language of the region. We apply various models of Long Short-Term Memory, a deep learning technique. The best approach achieved 83% accuracy. Furthermore, we also compared the proposed methods with state-of-the-art machine learning ones, such as support vector machines, nearest neighbor, Naïve Bayes, random forest, and logistic regression. The achieved results are promising.

1. INTRODUCTION

One of the essential aims of Saudi Vision 2030 is a prosperous economy. Therefore, the critical goals of the vision include strengthening economic and investment activities. Therefore, to have a good and successful project and business, we must consider project management as a discipline that initiates, plans, executes, and manages resources.

The project manager, who wants to start a new project or business, needs to know and improve the customers' feedback. The projects considered by our work are the eateries and could be a restaurant or a coffee shop. The overall aim is to improve the income and know the strengths and weaknesses to address them and increase the return on investment (ROI) and the quality of provided services. This all can be attained by using sentiment analysis of the customers' reviews.

Sentiment analysis is used to find a person's perspective on a specific topic. It uses natural language processing (NLP) and text analysis to extract subjective information from various sources [1]. Sentiment analysis deals with extracting feelings from the given text for a particular subject. Since the Internet has become a significant part of our lives, most people use social networking sites to express their opinions on different topics [2]. These sites are also used to learn more about other peoples' opinions. Therefore, mining this data and sentiment extraction has become an important field.

In this research, sentiment analysis techniques and Deep learning (DL) algorithms will be applied to a data set formed by collecting reviews from different people in the Al-Qassim region of Saudi Arabia to help businesses associated with eateries. Alharbi and Qamar [3] applied various machine learning (ML) algorithms to this data. However, to the best of our knowledge, no one has applied DL techniques to this data set. Moreover, we applied Long Short-Term Memory (LSTM), a widely used DL method. Anyone who wants to start a new company could rely upon and benefit from the result of this research founded on sentiment analysis.

The study aims to design an effective system capable of accurately classifying peoples' opinions about eateries in the Qassim region. A data set containing people's Arabic opinions are collected and annotated manually. We compared the obtained results with different machine learning algorithms as well. The rest of the paper is organized as follows: Section 2 presents a detailed literature review, whereas the research methodology is provided in Section 3. The experiments and results are discussed in Section 4, and Section 5 concludes the paper and provides some future research directions.

2. LITERATURE REVIEW

The related works include sentiment analysis using Machine Learning or Deep Learning Algorithms.

Lin et al. [4] performed Linear Discriminant Analysis (LDA) on topic models before sentiment analysis to find the popular comment topics of consumers during a specific period. They further used the characteristic keywords of LDA to classify the corpus using dictionary comparison. Moreover, they proved that the experimental system has good accuracy in emotional polarity classification.

Choudhary and Choudhary [5] performed sentiment analysis on phone reviews downloaded from Twitter. They discovered that Samsung, Motorola, and Oppo are the three most famous brands in the market.

Chakraborty et al. [6] introduced sentiment analysis in their studies. Although people's tweets about COVID-19 are primarily positive, they are entirely focused on reposting negative tweets and cannot find helpful words in WordCloud in tweets. The hypothesis was verified by employing a deep learning classifier, and they obtained 81% accuracy. In addition to these, the authors also proposed a fuzzy rule library based on the Gaussian membership function to identify emotions correctly. The proposed model got 79% accuracy.

Ezhilarasan et al. [7] use sentiment analysis to analyze customer needs through online reviews to help organizations improve product development.

Noor and Islam [8] explained how sentiment analysis has been applied to women's e-commerce reviews obtained from Amazon.com. Their dataset was preprocessed, and various ML algorithms were applied. Their best results represented an accuracy of 80.88% while using Sequential Minimal Optimization (SMO).

The authors [9] used the IMDB benchmark dataset for movie reviews to test various machine learning, neural networks, and deep learning models. Furthermore, different word-embedding techniques were tested. The study's findings showed that the LSTM model with Bidirectional Encoder Representations from Transformer (BERT) embeddings produced the best results and was 93% accurate.

The author [10] gave an overview of sentiment analysis for the Twitter dataset, named Sentiment140. They proposed a new approach for analyzing tweets as negative or positive classes from annotated Twitter records using deep learning neural networks combined with the CNN-LSTM method. This model used an efficient deep learning architecture with hyperparameters fine-tuned at the CNN layer, followed by a long-distance-dependent bidirectional LSTM neural network. It performed better than any baseline method in the benchmark dataset. The accuracy of the proposed model is 81.20%.

Another group of researchers [11] proposed a pattern-based approach to classify texts collected from Twitter (i.e., tweets). They classified the tweets into seven different classes. The experiments show that the approach reaches 56.9% accuracy and a precision level of sentimental tweets (other than neutral and sarcastic) equal to 72.58%.

Narr et al. [12] developed a scheme for sentimentally analyzing tweets in any language. For the Arabic language, Elhawary and Elfeky [13] used Arabic lexicon words to extract features and identify reviews.

Khine and Aung [14] have proposed a model that combined Multi-Aspect Attention (MAA) and Bi-LSTM. The model was developed for aspect-based sentiment classification. Two datasets were utilized to evaluate the model. The experiments showed that the use of MAA-BiLSTM gave an accuracy of 89%, which is the highest accuracy compared with baseline LSTM.

Ghallab et al. [15] performed a systematic literature review (SLR) of Arabic SA. A total of 108 relevant papers were analyzed in detail. It was observed that researchers are more focused on building resources and performing sentiment classification. Similarly, most articles processed Modern Standard Arabic (MSA) and dialectal Arabic (DA). As expected, the most common source is Twitter because of the public nature of the data. The most common preprocessing strategies include stemming, normalization, and tokenization, followed by stopword removal. N-grams are the most frequent type of feature. Moreover, SVM and NB are the most common methods employed for Arabic SA. A growing trend of using DL in Arabic sentiment analysis was also observed.

Alosaimi et al. [16] performed Arabic sentiment analysis for Saudi hotels while employing unsupervised ML, K-Means, and Hierarchical clustering techniques. Four thousand six hundred four reviews were obtained from the TripAdvisor website. Euclidean distance and cosine similarity was used as distance metrics. The best results were obtained with K-Means and cosine similarity. The results are not surprising since many previous kinds of research have shown that cosine works better with text.

Obiedat et al. [17] developed a hybrid classification algorithm based on SVM and Particle Swarm Optimization (PSO) for Arabic SA. The dataset contained more than 3000 reviews in the Arabic language about different restaurants in Jordan. The results showed that the hybrid approach performed better than the standard SVM, LR, RF, DT, and NN.

Saleh et al. [18] developed an ensemble approach for Arabic sentiment analysis using three DL models, LSTM, RNN, and Gated Recurrent Units (GRU), along with three traditional ML algorithms (RF, LR, and SVM). Among various DL models, LSTM got the best accuracy of 0.919 on the Arabic Sentiment Twitter Corpus (ASTC), consisting of 56,795 Arabic tweets collected in April 2019.

In another related research, Alharbi et al. [19] used a combination of character-level and word-level models to effectively represent the Arabic words in tweets. The improvement comes in two aspects: semantics and morphology.

3. RESEARCH METHODOLOGY

Customer reviews play an essential role in modern manufacturing informed decisions. This research applied an LSTM deep neural network architecture and used automatic feature extraction.

3.1 Dataset

The dataset consists of 1371 users' reviews of various eateries in the Qassim region. The reviews are in the native language of the area, i.e., Arabic. This makes sure that the true sentiment of the customers is captured. Each review is classified as positive or negative by three volunteers who are native Arabic speakers. This arrangement avoided deadlocks. A label for a review was finalized only if at least two volunteers agreed with the same label. The neutral reviews were not considered since they mainly provide information only. The final dataset contained 672 positive reviews (57.39%) and 499 negative ones (42.61%). Since the classes are not equally presented, we can consider the dataset as imbalanced, as shown in Figure 1.

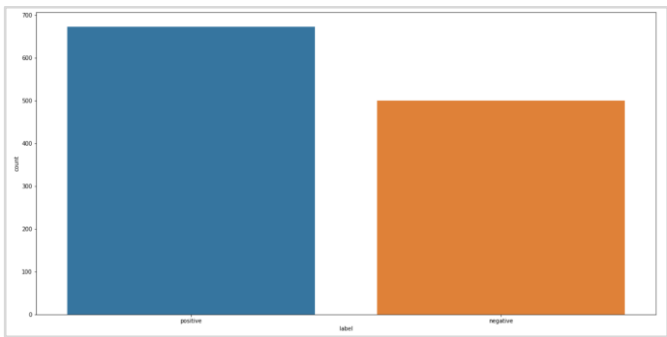

Figure 1. Distribution of positive and negative reviews

3.2 Framework

The primary tool that is used in this research is Python 3.7.4. Python is one of the most widely used general-purpose programming languages. Guido van Rossum designed it in 1991 [20]. It supports Object-Oriented Programming objects, which can be used in many areas of software development. Moreover, it is possible to integrate this language with other languages and tools. It has various ML libraries such as Pandas, Numpy, Scipy, Keras, and Theano. One of the libraries that we

used is Scikit-learn (sklearn).

Moreover, Python provides an excellent library for NLP tasks and text processing, namely Natural Language Toolkit (NLTK) [21]. It can be used on different platforms like Windows, Mac, and Linux. Python code is generally much easier to read and much quicker to write than code written in other languages [22, 23].

3.3 Data preprocessing

We applied standard preprocessing steps such as stemming, text cleaning, stop-words removal, tokenization, and normalization.

In Figure 2, we want to capture the most essential and highly mentioned positive customer review attributes. Therefore, we used a text analyzer to analyze the word counts such as: “ ”من ”، و ”، على ”، في ”،” and “الى.” Besides that, one can observe that the most commonly occurring positive words and the translated ones in English are as follows:

“المقاهي”،”المطاعم”،”ممتازة”،”الخدمة”،”افضل”،”جميلة”،”احب”.

“Love”, “beautiful”, “best”, “service”, “excellent”, “restaurants”, “cafes”.

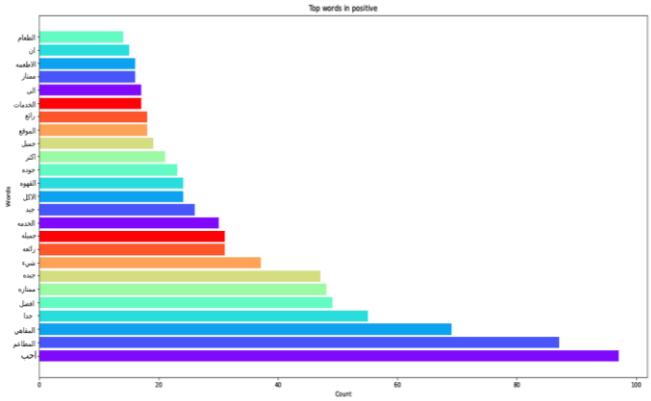

Figure 2. Most frequent Positive customer words

We also analyzed the negative reviews to see which aspects of the eateries the customer dislikes. In Figure 3, one can observe that the most frequent words in the negative text along with their translation in English are as follows:

“الموقع”،”الجوده”،”الانتظار”،”المواقف”،”التنوع”،”الخدمة”.

“Service”, “Diversity”, “Parking”, “Waiting”, “Quality”, “Location”.

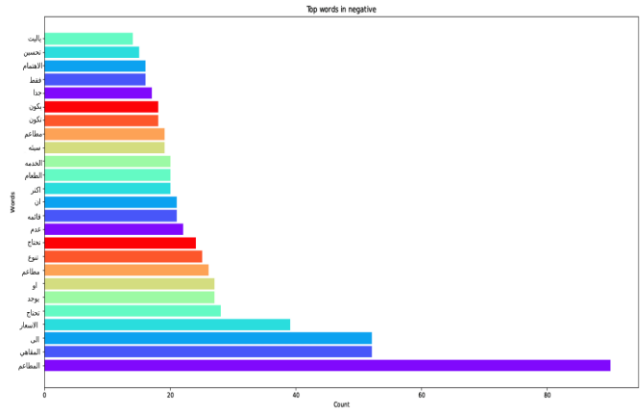

Figure 3. Most-frequent Negative customer words

3.4 Data learning techniques

This study proposes deep learning techniques to develop a Sentiment analysis model. We mainly used a deep learning approach: Long Short-Term Memory (LSTM). Here, we present the method and the proposed architecture.

LSTM classifier:

One can learn from long-term dependencies to understand the context behind a sentence [9]. LSTM is an extension of a recurrent neural network (RNN), and unlike the simple memory in RNN, LSTM has an advanced memory that can remember inputs over a long time. The model comprises the embedding layer as the input layer and one LSTM layer containing 128 neurons. The activation function of the sigmoid computes the output of the FC layer. ADAM is used to adapt the learning rate, whereas binary cross-entropy is used as the objective function. The number of epochs was 10 based on the early stopping. Once the parameter selection was completed, the model was trained and evaluated using 20% of the training set. The architectural design of the proposed LSTM sentiment model is shown in Figure 4. The LSTM layers are sandwiched between the Input layer and the Output layer. The input of the Input layer is composed of the pre-processed Arabic words.

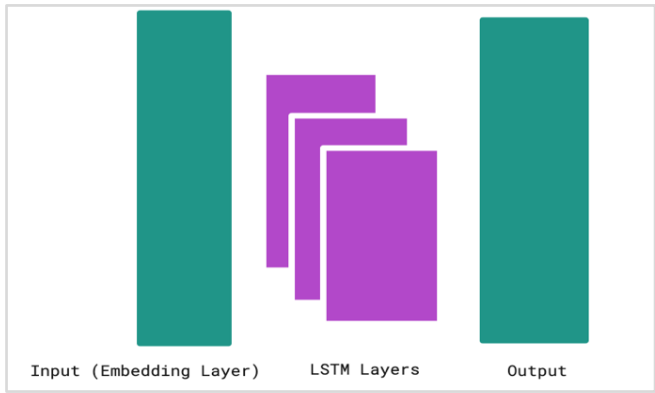

Figure 4. Proposed LSTM sentiment model

4. EXPERIMENTS AND RESULTS

In this research, we used LSTM with different hyperparameter settings. Three different models are evaluated in detail. As with many natural language processing applications, we used an embedding layer to create the word embeddings. The vocabulary size is 1754. The hyperparameters for the 1st model are provided in Table 1.

Table 1. Parameters settings for the first LSTM model

Parameters	Values
Number of LSTM layers	1
Number of nodes	100
Dropout rate	0.2
Recurrent dropout	0.2
Activation function	Sigmoid
Loss function	Binary cross-entropy
Optimizer	ADAM
Epochs	10
Batch size	128

Table 2. The model summary of the first LSTM model

Layer (type)	Output shape	Number of Parameters
Embedding	(None, 100, 100)	175,400
LSTM	(None, 100)	80,400
Dense	(100, 1)	101
Total parameters	255,901	
rainable parameters	255,901	

The model summary is given in Table 2.

Figure 5 shows the accuracy and loss curves for both training and validation sets. The maximum accuracy for the validation set is 0.819, whereas the same value for the training set is 0.9425. One can observe that the accuracies increase as the epochs are increased. The difference between training and validation sets' accuracy is almost the same as the beginning and the end. Nevertheless, the difference was reduced at the 6th epoch. On the other hand, the training and the validation loss are almost the same at epoch 1. The difference gradually increased and reached more than 21.7% at epoch 10.

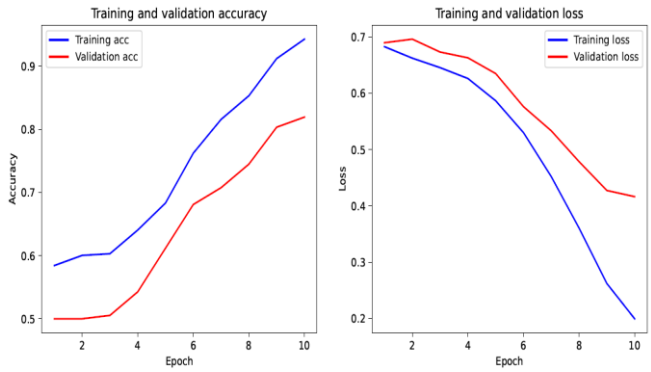

Figure 5. Model Accuracy and loss curves for training and validation sets for the 1st LSTM model with maximum epochs as 10

For the second set of experiments, the number of epochs increases to 20. The rest of the hyperparameters are kept the same as in the first experiment. A noticeable difference is that much more time is taken by this experiment than the first set of experiments.

Figure 6 shows the accuracy and loss curves for training and validation sets once the number of epochs increases to 20.

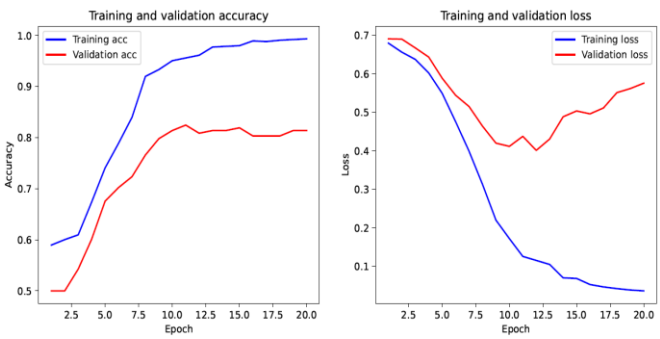

Figure 6. Model Accuracy and loss curves for training and validation sets for the 1st LSTM model once the epochs are increased to 20

The maximum accuracy for the validation set is 0.814, whereas the same value for the training set is 0.9933. One can observe that the accuracies, in general, increase as the epochs are increased. The validation accuracy sometimes decreases after epoch 10. The difference between training and validation sets' accuracy increases to 17.95% at epoch 20. On the other hand, the training and the validation loss are almost the same at epoch 1. The difference gradually increased and reached more than 53.93% at epoch 20. In the case of validation loss, it can be observed that whereas the loss decreased till the 10th epoch, it gradually increased afterward.

A different activation function, Rectified Linear Units (ReLU), is employed in the next set of experiments. The other hyperparameters are kept the same to have a fair comparison. No significant difference in the accuracy or loss was observed, as shown in Figure 7.

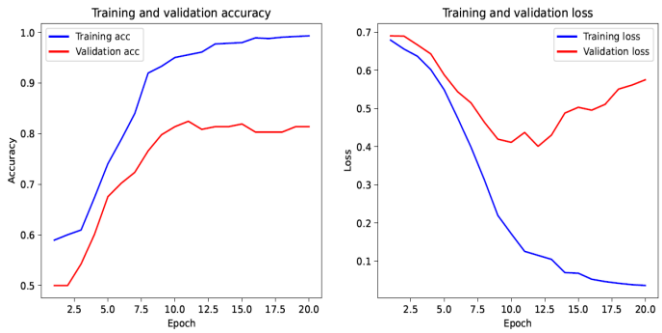

Figure 7. Accuracy and loss curves for training and validation sets for the LSTM model with ReLU as the activation function

Next, we present the accuracy of the test set, as shown in Table 3. The best results are put in bold. We first ran our experiments using the sigmoid function as the activation function, and the optimizer was chosen as ADAM. A gradual increase in accuracy is observed as the number of epochs increases from 10 to 20. Furthermore, the accuracy decreased while increasing the epochs to 30. On the other hand, a reduction in accuracy is seen once the activation function is changed to ReLU and the epochs are increased. Comparing the two activation functions, it can be noticed that the accuracy with sigmoid is much better than using ReLU as an activation function.

Table 3. Accuracy of the test for the various LSTM models

Experiment	Accuracy
Activation as sigmoid, optimizer as ADAM, ten epochs	0.8043
Activation as sigmoid, optimizer as ADAM, 15 epochs	0.8128
Activation as sigmoid, optimizer as ADAM, 20 epochs	**0.8298**
Activation as sigmoid, optimizer as ADAM, 30 epochs	0.7915
Activation as ReLU, optimizer as ADAM, 20 epochs	0.7702
Activation as ReLU, optimizer as ADAM, 15 epochs	0.7915

We now present a detailed comparison between the DL methods and ML-based ones, as shown in Table 4. The comparison is made with Support Vector Machines (SVM), logistic regression (LR), k nearest neighbor (kNN), Naïve

Bayes (NB), and Random Forest (RF). The best accuracy is obtained with SVM. Next in the line is LR, followed by kNN. RF and the proposed LSTM got 83% accuracy, 6% less than the best-performing method.

Table 4. Comparison of DL and ML methods

Algorithms	Accuracy
SVM	**0.89**
Logistic Regression	0.86
K Nearest Neighbor	0.84
Naïve Bayes	0.72
Random Forest	0.83
LSTM (Proposed Method)	0.83

5. CONCLUSIONS

In recent years, the sentiment analysis of product reviews has gained more and more attention. This paper applied long short-term memory (LSTM) for Arabic text mining. The text consists of reviews of different eateries in the Qassim region of Saudi Arabia. LSTM was able to get 83% accuracy, which can be considered a good result in sentiment analysis. Besides that, the results were compared with different machine learning algorithms, such as support vector machines, logistic regression, nearest neighbor, naïve Bayes, random forest, and logistic regression. SVM achieved the highest accuracy of 89%. In the future, we plan to increase the size of the dataset and apply other deep learning approaches, such as convolutional neural networks.

ACKNOWLEDGMENT

This work is supported by the Onaizah Colleges.

REFERENCES

[1] Mouthami, K., Devi, K.N., Bhaskaran, V.M. (2013). Sentiment analysis and classification based on textual reviews. International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India, pp. 271-276. https://doi.org/10.1109/ICICES.2013.6508366

[2] Kaur, H., Mangat, V., Nidhi. (2017). A survey of sentiment analysis techniques. International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 921-925. https://doi.org/10.1109/I-SMAC.2017.8058315

[3] Alharbi, L.M., Qamar, A.M. (2021). Arabic sentiment analysis of eateries' reviews: Qassim region case study. 2021 National Computing Colleges Conference (NCCC), Taif, Saudi Arabia, pp. 1-6. https://doi.org/10.1109/NCCC49330.2021.9428788

[4] Lin, H.C.K., Wang, T.H., Lin, G.C., et al. (2020). Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects. Applied Soft Computing, 97: 106755. https://doi.org/10.1016/j.asoc.2020.106755

[5] Choudhary, M., Choudhary, P.K. (2018). Sentiment analysis of text reviewing algorithm using data mining. International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 532-538. https://doi.org/10.1109/ICSSIT.2018.8748599

[6] Chakraborty, K., Bhatia, S., Bhattacharyya, S., et al. (2020). Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97: 106754. https://doi.org/10.1016/j.asoc.2020.106754

[7] Ezhilarasan, M., Govindasamy, V., Akila, V., Vadivelan, K. (2019). Sentiment analysis on product review: A survey. International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), Melmaruvathur, India, pp. 180-192. https://doi.org/10.1109/ICCPEIC45300.2019.9082346

[8] Noor, A., Islam, M. (2019). Sentiment analysis for women's e-commerce reviews using machine learning algorithms. 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1-6. https://doi.org/10.1109/ICCCNT45670.2019.8944436

[9] Rizk, Y.E., Asal, W.M. (2021). Sentiment analysis using machine learning and deep learning models on movies reviews. 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, pp. 129-132. https://doi.org/10.1109/NILES53778.2021.9600548

[10] Tyagi, V., Kumar, A., Das, S. (2020). Sentiment analysis on Twitter data using deep learning approach. 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, pp. 187-190. https://doi.org/10.1109/ICACCCN51052.2020.9362853

[11] Bouazizi, M., Ohtsuki, T. (2016). Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter. IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia. https://doi.org/10.1109/ICC.2016.7511392

[12] Narr, S., De Luca, E.W., Albayrak, S. (2011). Extracting semantic annotations from Twitter. Fourth workshop on Exploiting semantic annotations in information retrieval - ESAIR '11, Glasgow, Scotland, UK, pp. 15-16. https://doi.org/10.1145/2064713.2064723

[13] Elhawary, M., Elfeky, M. (2010). Mining Arabic business reviews. IEEE International Conference on Data Mining Workshops, Sydney, NSW, Australia, pp. 1108-1113. https://doi.org/10.1109/ICDMW.2010.24

[14] Khine, W.L.K., Aung, N.T.T. (2020). Multi-aspect attention model for aspect-based sentiment classification using deep learning. International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar, pp. 206-211. https://doi.org/10.1109/ICAIT51105.2020.9261803

[15] Ghallab, A., Mohsen, A., Ali, Y. (2020). Arabic sentiment analysis: A systematic literature review. Appied Computational Intelligence and Soft Computing, 2020: 7403128. https://doi.org/10.1109/ICAIT51105.2020.9261803

[16] Alosaimi, S., Alharthi, M., Alghamdi, K., Alsubait, T., Alqurashi, T. (2020). Sentiment analysis of Arabic reviews for Saudi hotels using unsupervised machine learning. Journal of Computer Science, 16(9): 1258-1267. https://doi.org/10.3844/jcssp.2020.1258.1267

[17] Obiedat, R., Qaddoura, R., Al-Zoubi, A.M., et al. (2022). Sentiment analysis of customers' reviews using a hybrid evolutionary SVM-based approach in an imbalanced data

distribution. IEEE Access, 10: 22260-22273. https://doi.org/10.1109/ACCESS.2022.3149482
[18] Saleh, H., Mostafa, S., Alharbi, A., El-Sappagh, S., Alkhalifah, T. (2022). Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. Sensors, 22(10): 3707. https://doi.org/10.3390/s22103707
[19] Alharbi, A.I., Smith, P., Lee, M. (2022). Integrating character-level and word-level representation for affect in Arabic tweets. Data & Knowledge Engineering, 138: 101973. https://doi.org/10.1016/j.datak.2021.101973
[20] van Rossum, G., Python Development Team. (2018). An introduction to Python, Python Softw. Found., pp. 1-155.
[21] Solangi, Y.A., Solangi, Z.A., Aarain, S., et al. (2018). Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis. IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bangkok, Thailand, pp. 1-4. https://doi.org/10.1109/ICETAS.2018.8629198
[22] Matthes, E. (2019). Python Crash Course: A Hands-On, Project-Based Introduction to Programming, 2nd Edition, San Francisco, CA, United States: No Starch Press.
[23] Yechuri, P.K., Ramadass, S. (2021). Semantic web mining for analyzing retail environment using Word2Vec and CNN-FK. Ingénierie des Systèmes d'Information, 26(3): 311-318. https://doi.org/10.18280/isi.260308