Evaluating sentiment analysis for Arabic Tweets using machine learning and deep learning

Areej ALSHUTAYRI[1], Huda ALAMOUDI[1], Boushra ALSHEHRI[1], Eman ALDHAHRI[1], Iqbal ALSALEH[2], Nahla ALJOJO[3], Abdullah ALGHOSON[3]

[1] College of Computer Science and Engineering, Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, Saudi Arabia

[2] Faculty of Economic and Administration, Management Information Systems Department, King AbdulAziz University, Jeddah, Saudi Arabia

[3] College of Computer Science and Engineering, Information Systems and Technology Department, University of Jeddah, Jeddah, Saudi Arabia

ealsaleh@kau.edu.sa, (aoalshutayri, eaal-dhahery, nmaljojo, alghoson)@uj.edu.sa

Abstract: Sentiment analysis is concerned with determining whether a certain material contains online information which expresses positive or negative sentiments. The tools for performing this analysis should be able to identify and assess thoughts and feelings with a reasonable degree of accuracy on feelings that are made openly available by people. It is expected that sentiment analysis would be performed for social media. That is why this paper investigates online social media, as sentiment analysis has become an important subject, and it is one of the approaches employed in the field of natural language processing. Sentiment analysis was applied for an Arabic Twitter dataset in order to identify the feelings expressed by the textual tweets and determine whether they were positive, negative, or neutral. Bigrams and unigrams were used when employing the multinomial Naïve Bayes, Gaussian Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM) machine learning algorithms. The Logistic Regression algorithm achieved the highest accuracy, that is with 63.40%. The Long Short-Term Memory (LSTM) neural network was used for the deep learning-based classification, and it reached an accuracy rate of 70%, a figure which proved to be higher than the results shown in the related works.

Keywords: sentiment analysis (SA), machine learning (ML), deep learning, Arabic tweets.

1. Introduction

Sentiment Analysis (SA) is one of the Natural Language Processing (NLP) applications that aims to process and analyze data that is written in human languages (Ghallab et al., 2020). In recent years, SA topics have become popular research areas due to a myriad of social media applications such as Facebook and Twitter (Ahmad et al., 2019). Twitter has a large number of users who post countless written tweets in different languages. Therefore, the text of a tweet needs to be analyzed to determine whether the feeling it expresses is positive or negative. SA was applied to different natural languages including, yet not limited to English, Korean and Arabic. Arabic is the 4th most popular language (Zahidi et al., 2021), and it is the most challenging one due to the richness of its morphology, which makes it complex. Similarly, Arabic sentences don't follow a specific order and sometimes depend on diacritics in order not to be ambiguous (Duwairi & Abu Shaqra, 2021). Additionally, the resources and tools for Arabic languages are limited (Zahidi et al., 2021). This paper employs machine learning and deep learning, the dataset is ready for analysis as data is not being collected starting from scratch. Despite this, challenges can be expected because of factors such as missing labels, imbalanced datasets, inexact values in labels depending on emojis used in the tweets and differences between several dialects. In addition, the accuracy of sentiment analysis for Arabic tweets shall be improved based on the use of machine learning and deep learning classifiers and through the comparison of the obtained results.

The paper is structured as follows: Section 2 presents the background and related works for ML, DL and SA. In Section 3, system architecture and its the main components are discussed, along with the performed experiments and the obtained results. Finally, Section 4 sets forth the conclusion of this paper.

https://doi.org/10.33436/v32i4y202201

2. Related works

2.1. Sentiment analysis

Nowadays, the sentiment analysis approach has grown in popularity with several platforms on the web such as e-commerce, blogs and forums as well as social networks like Twitter, Facebook, and other. Sentiment analysis is a field of natural language processing (NLP), and the basic goal of this science is the extraction of emotion from a text that is written by people (Fouadi et al., 2020). In sentiment analysis in the English language, an opinion is explained by making mention of (o; a; so; h; t) that contains, first of all, object 'o', which is the opinion goal (Fouadi et al., 2020). It can be an output like a service, a topic, an issue, a person, an organization, or an event. It also includes the aspect ‘a’, which is the attribute of the object ‘o’. Sentiment orientation ‘so’ illustrates whether an opinion is positive, negative or neutral, and opinion holder ‘h’, is the person or organization that conveys an attitude or opinion. Time ‘t’ represents the point in which this opinion is expressed (Fouadi et al., 2020; Nassr et al., 2019; Kharde & Sonawane, 2016). The stages of sentiment analysis are the document stage, sentence stage, and aspect stage. Its utilization is a deeply challenging research area which includes different complicated tasks. There are also some topics of interest to most researchers in this field including subjectivity classification, lexicon creation, opinion spam detection, and aspect-level sentiment classification (Nassr et al., 2019).

As it was previously stated, the Arabic language has one of the most challenging collections for sentiment analysis researchers. There are three types of Arabic: classical Arabic, which is the language of the Qur’an (Islam’s Holy Book), modern standard Arabic (MSA) and dialectical Arabic (Boudad et al., 2018). Dialectical Arabic points to all differences in everyday spoken language. These differences exist among Arab countries and they can also exist between some cities within the same country. Arabic is written from right to left and is void of the concept upper or lower cases letters. This language comprises 28 letters, with 25 consonants and just three vowels. However, the Arabic script also uses diacritical marks as short vowels to provide the correct pronunciation and clarify the meaning of a word. The absence of diacritical marks on words is a problem because one cannot always read and understand texts clearly. To present an example, the word (بر) may mean (بُر - Brown Flour), (بَر - Mainland) or (بِر- Alms). The Arabic language has a very rich and complex morphology, a word in Arabic features many morphological aspects such as agglutination, inflection, and derivation. All the aforementioned as well as other unique aspects make sentiment analysis in the Arabic language more complex (Boudad et al., 2018).

2.2. Sentiment analysis by AI strategies (deep learning and machine learning)

Many experiments have been carried out on sentiment analysis through machine learning to predict a certain sentiment. Machine learning algorithms are mainly categorized into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Hemalatha & Ramathmika, 2019). The most common algorithms dealing with text are Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Decision Trees and K-Nearest Neighbor (KNN) (Hemalatha & Ramathmika, 2019). The past few years have displayed a tendency of implementing deep learning models in the field of natural language processing (NLP). Deep Neural Networks (DNNs) are built up of artificial neural networks encompassing multiple invisible layers between the input layer and the output layer. Deep learning has a lot of models, and the most famous ones are Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). Long Short-Term Memory Networks (LSTMs) are also included, which are a type of RNNs (Yadav & Vishwakarma, 2019). There are many pieces of research on sentiment analysis with dissimilar methods and different languages. Table 1 lists some of the related works in a comparative manner.

Table 1. Related works summarization for Sentiment Analysis

RELATED WORK	DATASET	SIZE OF DATASET	LANGUAGE	LABEL	ALGORITHM	ACCURACY
(AL-BAYATI ET AL., 2020)	LABR	16,448 tweets	Arabic	Positive, negative	LSTM	82%

(GOULARAS & KAMIS, 2019)	SemEval	32,000 tweets	English	-	LSTM, CNN	59%
(RAMADHANI & GOO, 2017)	Twitter API	4,000 tweets	English, Korean	Positive, Negative	Deep Feedforward Networks	75%
(MOHAMMED & KORA, 2019)	Movie reviews	6,000 tweets	English	Positive, Negative	LSTM	80.83%
(AL-HASSAN & AL-DOSSARI, 2021)	Twitter API	11,000 tweets	Arabic	None, Religious, Racism, Sexism or general hate	CNN + LTSM LTSM	72% 71%
(CHENG & TSAI, 2019)	Twitter API	40,000 tweets	Arabic	Positive Negative	LSTM	88.05%
(ALSALMAN, 2020)	Twitter API	2,000 tweets	Arabic	Positive Negative	Discriminant Polynomial Naive Bayes	87.5%
(ABUUZNIEN ET AL., 2020)	Twitter API	2,116 tweets	Arabic (Sudanese)	Positive, Negative, Neutral	NB SVM Logistic Regression KNN	SVM (95%)
(HEIKAL, M. ET AL. 2018)	ASTD	10,000 tweets	Arabic	positive, negative, neutral, and objective	CNN + LTSM CNN LSTM	CNN+ LSTM 65.05% CNN 64.30% LSTM 64.75%

In (Al-Bayati et al., 2020) the authors used DL to analyze an Arabic book reviews dataset called LABR which refers to Large-Scale Arabic Book Reviews. The dataset included 16,448 reviews and the predicted output was either positive or negative. This paper employed LSTM neural network and tried different LSTM output sizes with different batch sizes. The best results for accuracy reached approximately 82% when the LSTM output was 50 with a batch size of 256.

In (Goularas & Kamis, 2019) and (Ramadhani & Goo, 2017) the authors used deep learning models to test sentiment analysis for English tweets. In (Goularas & Kamis, 2019), the authors compared LSTM with CNN models for approximately 32,000 tweets from three datasets, which were used at SemEval competitions in 2014, 2016 and 2017. The results for LSTM and CNN are similar and when used together, they returned even better results. As for Ramadhani & Goo (2017), they applied deep feedforward networks on a dataset containing about 4,000 English and Korean tweets labelled as positive or negative, and the accuracy of the result was about 75%.

In (Mohammed & Kora, 2019), authors introduced a framework based on a deep learning model for sentiment analysis of movie reviews and three types of deep learning models were applied on the dataset using LSTM, BILSTM and Gated Recurrent Unit (GRU). The highest accuracy achieved was 80.83% when using LSTM.

The study of Al-Hassan & Al-Dossari (2021) aimed to detect hate in Arabic tweets by classifying them into 5 classes: "none, religious, racism, sexism or general hate". The authors made a comparison of the results of the four deep learning models which were LTSM, CNN + LTSM, GRU and CNN + GRU. These were applied on a dataset of 11,000 tweets, and the best result was produced by CNN + LTSM with a 72% success rate.

In (Cheng & Tsai, 2019), the authors applied three models of deep learning on 40,000 Arabic tweets containing different topics. The employed techniques were RNN and LSTM. The highest improvement was by LSTM with an accuracy rate of 88.05%.

In (AlSalman, 2020) applied machine learning based on the discriminant polynomial Naive Bayes (DMNB) method with 4-gram tokenizer, stemming and word frequency, and inverse document frequency (TF-IDF), which is a technology that improves the corpus-based Arabic sentiment analysis method. The dataset included 2,000 Arabic tweets tagged in two different categories (negative and positive). The accuracy of the DMNB classifier of the proposed method was 87.5%.

In (Abuuznien et al., 2020), the authors focused on extracting and analyzing Sudan's social media feeds about ride-sharing services. They applied four classifiers of machine learning on a dataset of 2,116 tweets, namely, NB, SVM, Logistic Regression, and KNN, with the purpose of measuring their performance. The best accuracy was given by SVM, that is 95%.

In (Heikal et al., 2018), the authors proposed an ensemble model based on DL that combined two classifiers, CNN and LSTM model. The proposed model was used to predict the sentiments in Arabic tweets. They used the ASTD dataset, which consists of 10,000 tweets distributed in 4 categories (positive, negative, neutral, and objective). The best accuracy for the LSTM model was 64.75% with a loss rate of 0.2, while other parameters remained unchanged. In the default configuration, the CNN model used a fully connected layer with a batch size of 100, and achieved an accuracy of 64.30% as the best result, while other parameters remained in the default configuration, and the accuracy of the ensemble model was 65.05%.

The contribution of this paper lies in using ML and DL for Arabic sentiment analysis for Twitter data with different Arabic dialects with the dataset classified into three labels, namely positive, negative and neutral. Similarly, in this paper DL is compared with different ML models.

3. System architecture

This work aims to analyze tweets collected from the twitter to predict the orientation of sentiment expressed by those tweets (positive, negative and natural). Several techniques based on deep learning and machine learning classifiers were employed to classify the tweets in order to determine the most efficient one(s). The proposed system architecture is shown in Figure 1. The architecture illustrates the three main phases before the predicted output. In the beginning, samples are fetched from a Twitter dataset, then these samples are cleared from any noise in each row in the pre-processing phase. After that, the clean dataset enters the classification phase which comprises different models of machine learning or deep learning, and these are applied to get a predicted output.

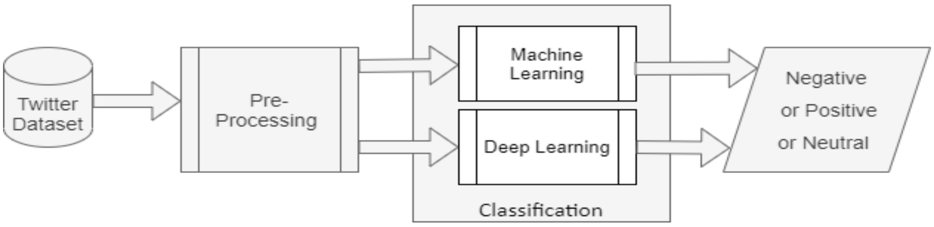


Figure 1. Block diagram for sentiment analysis of Arabic tweets

3.1. Data collection stage

This work uses 32,186 rows that are fetched from Arabic Sentiment Analysis Dataset (ASAD) (Alharbi et al., 2020). Each of them is annotated manually by at least three annotators in either one of three types of sentiments, namely positive, negative, and neutral as it was mentioned previously. The dataset includes more than 4,900 positive tweets, more than 4,800 negative tweets and more than 22,275 neutral tweets, which make up most of the data. The tweets in the dataset were written in several dialects that include 'Khaleeji, Hijazi and Egyptian' dialects (Alharbi et al., 2020). Almost all of the tweets were written in 2020. The biggest challenge is that this analysis is made on an imbalanced set, given the figures mentioned above.

In this paper the ASAD was divided into 85% for training, which equals 27,200 samples, and 15% for testing, which means approximately 4,800 samples.

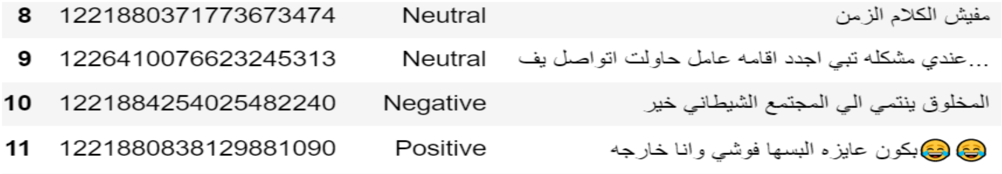
3.2. Tweet pre-processing phase

Any dataset usually has noise characters which reduce the efficiency of the results or can affect them negatively. Wherefore, all characters that affect the training process must be cleared from the dataset. There are some examples of tweets shown before and after pre-processing in Figure 2 and Figure 3.

1-The following list includes the data which was removed from the dataset:

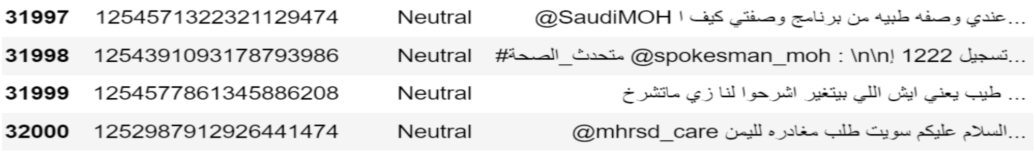
- English letters, repetition (chars/tweets) and special characters
- Diacritics, punctuation marks, URL and HTML tags
- Mentions and hashtags
- Stop words like (’أما‘ , ’هذا‘ ,’الذي‘ ,’التي‘ , ’إذا‘).

2- Normalizing Arabic Text: Some letters can be represented in numerous ways in the Arabic language. An example of this is the letter (Alef), which has different forms (ا-إ-آ-أ), and thus, normalization is applied to use one form which is (أ).



8	1221880371773673474	Neutral	مفيش الكلام الزمن
9	1226410076623245313	Neutral	...عندي مشكله تبي اجدد اقامه عامل حاولت اتواصل يف
10	1221884254025482240	Negative	المخلوق ينتمي الي المجتمع الشيطاني خير
11	1221880838129881090	Positive	😂😂بكون عايزه البسها فوشي وانا خارجه

Figure 2. Examples of tweets in the dataset after pre-processing



31997	1254571322321129474	Neutral	...عندي وصفه طبيه من برنامج وصفتي كيف ا @SaudiMOH
31998	1254391093178793986	Neutral	...تسجيل 1222 !\n\n : @spokesman_moh متحدث_الصحة#
31999	1254577861345886208	Neutral	... طيب يعني ايش اللي بيتغير اشرحوا لنا زي ماتشرخ
32000	1252987912926441474	Neutral	...السلام عليكم سويت طلب مغادره لليمن @mhrsd_care

Figure 3. Examples of tweets in the dataset before pre-processing

3.3. Experiments and results

3.3.1. Machine learning

The first step in this phase is the construction of a word vector that can be used by the employed classifier(s). This is followed by the use of different popular techniques to classify Arabic tweets based on how many papers use them in sentiment analysis. The classifiers are multinomial Naive Bayes, Logistic Regression, Gaussian Naive Bayes, and SVM.

Classification results for different ML algorithms:

1- A grid search is used for finding the optimal *hyperparameters* of a model which results in the most accurate predictions. Table 3 shows the experiment for grid search with different classifiers and each classifier has many parameters. Notably, changing values helps to reach the best result as it is shown in Table 2.

Table 2. Description of the employed parameters

Description of Parameter	Parameter Name	Classifiers
The regularization parameter informs the SVM how badly you don't want each training example to be misclassified.	C	SVM
The classification accuracy is improved by using kernel parameters as a tuning function. The type of kernel must be determined among the following: ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’. If it is not determined, the default kernel is "rbf".	Kernel	SVM

The alpha parameter in multinomial Naive Bayes and this is what is known as a hyperparameter. As a parameter, it governs the model's shape.	Alpha	Multinomial NB
Whether or not to learn previous probabilities for each class. A uniform prior would be used if false.	Fit_prior	Multinomial NB
Prior class probabilities would not be changed depending on the dataset when priors are given (in an array).	Priors	Gaussian NB
Used to achieve variance smoothing	Var_smoothing	Gaussian NB
Also called primal formulation that is only carried out for 12 penalties with liblinear solver.	Dual	Logistic Regression
Used to determine the standard utilized in the penalization.	Penalty	Logistic Regression

Table 3. The experiment for grid search

Classifiers	Parameters	Rank test score
SVM	{'C': 10, 'kernel': 'rbf'}	0.6340 %
Multinomial NB	{'alpha': 1.0, 'fit_prior': True}	0.6305 %
Gaussian NB	{'priors':None, 'var_smoothing': 1e-08}	0.534 %
Logistic Regression	{'dual': False, 'penalty': 'l2'}	0.635 %

2- Apply Unigrams and Bigrams

Unigram and bigram features, in which the unigram consists of a single word description and the bigram consists of a two-word feature description.

A) Unigram with Multinomial NB

Table 4 shows the experiment results for unigrams with Multinomial Naïve Bayes:

- First experiment with min_df from [1…10].

Min_df: is used for deleting terms that appear too infrequently.

- Pick min_df results with the best accuracy.
- Experiment with max_df from [0.5, 0.6, …1.0].

Max_df: is used for deleting terms that appear too frequently.

- Highest Accuracy is 61.1 with min_df =1.

Table 4. The experiment results for unigrams with Multinomial Naïve Bayes

min_df	Training Acc.	Test Acc. (%)	Running Time (s)
1	92.937	61.1	14.30
2	85.561	60.6	1.83
3	81.048	61.0	1.03
4	78.172	60.0	0.79
5	75.409	60.0	0.61
6	73.572	59.7	0.59
7	72.247	59.2	0.50
8	71.046	59.2	0.46
9	69.771	58.6	0.43
10	68.784	58.0	0.39

B) Unigrams and bigrams with Multinomial NB

Table 5 shows the results of the experiment for unigrams and bigrams with Multinomial Naïve Bayes:

- First experiment with min_df from [2…10].
- Pick min_df results with the best accuracy.
- Experiment with max_df from [0.5, 0.6, …,1.0].
- Highest accuracy is 61.7 with min_df = 3.

Table 5. The experiment results for unigrams and bigrams with Multinomial Naïve Bayes

min_df	Training Acc.	Test Acc. (%)	Running Time (s)
2	87.086	61.4	2.20
3	81.548	61.7	1.47
4	78.522	61.1	1.07
5	75.672	60.5	0.89
6	73.559	59.8	0.81
7	72.022	59.8	0.77
8	70.921	59.6	0.71
9	69.721	58.8	0.64

C) Unigrams with Logistic Regression

Table 6 shows the results of the experiment for unigrams with logistic regression:

- First experiment with min_df from [1…10].
- Pick min_df results with the best accuracy.
- Experiment with max_df from [0.5, 0.6, … 1.0]
- Highest accuracy is 59.7 with min_df=3

Table 6. The experiment results for unigrams with Logistic Regression

min_df	Training Acc.	Test Acc. (%)	Running Time (s)
1	93.999	59.4	61.23
2	88.799	59.2	33.18
3	85.336	59.7	20.94
4	82.573	59.5	13.47
5	79.835	59.5	10.80
6	77.672	58.3	8.01
7	76.010	58.0	6.50
8	74.534	57.8	5.15
9	73.097	57.1	5.75
10	71.796	57.0	4.64

D) Unigrams and bigrams with Logistic Regression

Table 7 shows the results of the experiment for unigrams and bigrams with Logistic Regression:

- First experiment with min_df from [1…10].
- Pick min_df results with the best accuracy.

- Experiment with max_df from [0.5, 0.6, … 1.0]
- Highest accuracy is 60.6 with min_df=2

Table 7. The experiment results for unigrams and bigrams with Logistic Regression

min_df	Training Acc.	Test Acc. (%)	Running Time (s)
2	90.274	60.6	31.80
3	86.311	59.9	17.89
4	83.760	60.1	14.97
5	80.760	59.2	10.84
6	78.360	58.2	10.25
7	76.560	58.1	8.63
8	75.059	58.2	6.62
9	73.697	57.3	6.34
10	72.284	57.6	5.34

3.3.2. Deep learning

In deep learning, there are three layers, the first one being the embedding layer. The LSTM layer and the dense layer complete the chain as it is shown in Figure 4 below.

Embedding Later
LSTM Layer
Dense Layer

Figure 4. Deep learning flow

The first layer of the proposed model, the embedding layer, is the input of the next layer, it applied embedded word vector which converts the data into vector representations or numeric forms to get efficient results. In this paper an embedding layer created by Keras library was used, which is one of the various libraries that have methods for word embedding.

The second layer is the LSTM layer and in this layer, LSTM neural network is used which is a special kind of RNN. In this paper the LSTM layer was used with a dropout of 20% and a recurrent dropout of 20%. The last layer is the dense layer which is also called the output layer. Softmax probability distribution is used as an activation function. Using Softmax to obtain the output predicts the input class as positive, negative, or neutral.

In this model, the LSTM neural network was used on a Twitter dataset which included 32,186 samples. There were divided into a training and a testing set, which represented 85% and 15% of the given dataset, respectively. The other properties were as follows:

- Input_dim = 2500 words and lstm_out = 196 vectors
- epochs = 10 and batch_size = 32

The best validation accuracy reached 70% with a loss of 2%. Table 8 shows the results for the LSTM model implemented in this work.

Table 8. The experiment results for LSTM

Epoch	Accuracy (%)	Loss (%)	Running Time (s)
1	69.5	23	85
2	69.7	10	77
3	69.6	6	77

4	69.2	4	78
5	69.6	3	83
6	69.5	2	84
7	69.4	2	81
8	69.5	2	84
9	69.6	2	82
10	70.0	2	80

3.3.3. Results

This subsection illustrates the obtained experimental results. In this study sentiment analysis techniques based on deep learning and machine learning classifiers were applied, and the results for the implemented models were compared. Four popular machine learning classifiers were used for sentiment analysis, namely multinomial NB, Gaussian NB, logistic regression, and SVM, and grid search was applied to find the best hyperparameters of the model with the intent to generate the most accurate predictions. The highest accuracy was that of SVM, namely 63.40%, and the parameters involved {'C': 10,'kernel':'rbf'}. Furthermore, when applying unigrams and bigrams, the highest accuracy was that of multinomial NB, namely 61.7%, and min_df = 3. In deep learning, LSTM was applied. In this model, the LSTM neural network was used for the analysed Twitter dataset. The dataset included 32,186 samples, which were divided into a training and a testing set which represented 85% and 15% of the given dataset, respectively. The best validation accuracy rate was 70%, and the loss was 2%. As for the comparative results for ML and DL technology, the LSTM used in DL had the highest accuracy rate.

4. Conclusion

This paper allows for comparing different machine learning and deep learning algorithms for more than 32,000 Arabic Tweets in order to classify the sentiments they express as positive, negative and neutral. Different machine learning algorithms were applied, namely multinomial and Gaussian Naive Bayes, Logistic Regression and SVM. The best accuracy was achieved when using SVM, namely 63%. The deep learning algorithm increased the achieved accuracy up to 70% using LSTM. The reason for the low accuracy lied in the fact that the analysed dataset was imbalanced. It included 22,275 neutral tweets, while the positive tweets amounted to 4,920 and the negative tweets to 4,806. Likewise, the labels related to the text of some samples were not accurate. All these things contributed to lowering the achieved accuracy and raised the challenge of increasing the accuracy.

REFERENCES

1. Abuuznien, S., Abdelmohsin, Z., Abdu, E. & Amin, I. (2020). Sentiment Analysis for Sudanese Arabic Dialect using comparative Supervised Learning approach. In *International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, (pp.1-6).

2. Ahmad, S., Asghar, M. Z., Alotaibi, F. & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 9(1), 24.

3. Al-Bayati, A., Al-Araji, A. & Ameen, S. (2020). Arabic Sentiment Analysis (ASA) using Deep Learning Approach. *Journal of Engineering*, 26(6), 85-93.

4. Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. & Zhang, X. (2020). ASAD: *A Twitter-based Benchmark Arabic Sentiment Analysis Data set.* [Online]. Available from http://arxiv.org/abs/2011.00578.

5. Al-Hassan, A. & Al-Dossari, H. (2021). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*. https://doi.org/10.1007/s00530-020-00742-w.

6. AlSalman, H. (2020). An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media. In *3rd International Conference on Computer Applications & Information Security (ICCAIS)*, (pp.1-4).

7. Boudad, N., Faizi, R., Thami, R. O. & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4), 2479-2490.

8. Cheng, L.-C. & Tsai, S.-L. (2019). Deep Learning for Automated Sentiment Analysis of Social Media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, (pp. 1001-1004).

9. Duwairi, R. & Abu Shaqra, F. (2021). Syntactic- and morphology-based text augmentation framework for Arabic sentiment analysis. *PeerJ. Computer Science*, 7(4), e469.

10. Fouadi, H., Satori, K., El Moubtahij, H., Yahyaouy, A. & Lamtougui, H. (2020). Applications Of Deep Learning In Arabic Sentiment Analysis: Research Perspective. In *1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, (pp. 1-6).

11. Ghallab, A., Mohsen, A. & Ali, Y. (2020). Arabic Sentiment Analysis: A Systematic Literature Review. *Applied Computational Intelligence and Soft Computing*, 1-21, (10.1155/2020/7403128).

12. Goularas, D. & Kamis, S. (2019). Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data. *International Conference on Deep Learning and Machine Learning in Emerging Applications* (Deep-ML), (pp. 12–17).

13. Heikal, M., Torki, M. & El-Makky, N. (2018). Sentiment Analysis of Arabic Tweets using Deep Learning. *Procedia Computer Science*, 142, 114-122.

14. Hemalatha, S. & Ramathmika, R. (2019). Sentiment Analysis of Yelp Reviews by Machine Learning. In *International Conference on Intelligent Computing and Control Systems (ICCS)*, (pp. 700-704).

15. Kharde, V. A. & Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*, 139(11), 5–15.

16. Mohammed, A. & Kora, R. (2019). Deep learning approaches for Arabic sentiment analysis. *Social Network Analysis and Mining*, 9(52). https://doi.org/10.1007/s13278-019-0596-4

17. Nassr, Z., Sael, N. & Benabbou, F. (2019). A comparative study of sentiment analysis approaches. In *SCA '19: Proceedings of the 4th International Conference on Smart City Applications*, (pp. 1-8).

18. Ramadhani, A. M. & Goo, H. S. (2017). Twitter sentiment analysis using deep learning methods. In *International Annual Engineering Seminar (InAES)*, (pp. 13–16).

19. Yadav, A. & Vishwakarma, D. (2019). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6), 4335-4385.

20. Zahidi, Y., El Younoussi, Y. & Al-Amrani, Y. (2021). Different valuable tools for Arabic sentiment analysis: a comparative evaluation. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(1), 753.

* * *

Areej ALSHUTAYRI is an Assistant Professor at the Department of Computer Science and Artificial Intelligence of the University of Jeddah, Jeddah, Saudi Arabia. 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 include machine learning and natural language processing to understand languages, especially Arabic and its dialects.

* * *

Huda ALAMOUDI is a Master's student at the College of Computer Science and Engineering, Computer Science and Artificial Intelligence Department at the University of Jeddah, Jeddah, Saudi Arabia.

* * *

Boushra ALSHEHRI is a Master's student at the College of Computer Science and Engineering, the Department of Computer Science and Artificial Intelligence at the University of Jeddah, Jeddah, Saudi Arabia.

* * *

Eman ALDHAHRI received her Ph.D. Degree in Computer Science from the University of Memphis, Memphis, United States of America in 2019. She received her M.Sc. Degree in Computer Science from the Southern Illinois University, Carbondale, United States of America in 2014, and her B.Sc. Degree in Computer Science from Taibah University, Al-Madinah, Saudi Arabia in 2006. Dr. Aldhahri is currently an Assistant Professor at the Department of Computer Science and Artificial Intelligence of the College of Computer Science and Engineering, at the University of Jeddah, Jeddah, Saudi Arabia.

* * *

Iqbal ALSALEH obtained her Ph.D. in Management Information Systems at Portsmouth University. She is currently working as an Associate Professor at the Faculty of Economic and Administration, the Management Information Systems Department at King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include Business, Management Information Systems, Leadership, AI, Machine Learning and Deep Learning.

* * *

Nahla ALJOJO obtained her Ph.D. in Computing at Portsmouth University. She is currently working as an Associate Professor at the College of Computer Science and Engineering, the Information systems and Technology Department at the University of Jeddah, Jeddah, Saudi Arabia. Her research interests include adaptivity in web-based educational systems, e-Business, leadership studies, information security and data integrity, e-Learning, education, AI, Machine Learning, Deep Learning, health informatics, environment and ecology, and logistics and supply chain management. Her contributions have been published in prestigious peer-reviewed journals.

* * *

Abdullah ALGHOSON obtained his Ph.D. in Information Systems and Information Technology at Claremont Graduate University. He is currently working as an Assistant Professor at the College of Computer Science and Engineering, Information Systems and Technology Department, at the University of Jeddah, Jeddah, Saudi Arabia. His research interests include Information Retrieval, Natural Language Processing (NLP), Machine Learning, Health Informatics and e-Learning.