Journal of King Saud University – Computer and Information Sciences 35 (2023) 227–237

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages

Pinar Savci [a], Bihter Das [b]

[a] Arçelik A.Ş. Karaağaç Caddesi 2-6, Sütlüce Beyoğlu, 34445 Istanbul, Turkey
[b] Department of Software Engineering, Technology Faculty, Firat University, 23119 Elazig, Turkey

ARTICLE INFO

Article history:
Received 22 November 2022
Revised 25 January 2023
Accepted 19 February 2023
Available online 24 February 2023

Keywords:
Sentiment analysis
Natural language processing
Deep learning
Pre-trained language models
Machine learning
E-commerce

ABSTRACT

In the business world, large companies that can achieve continuity in innovation gain a significant competitive advantage. The sensitivity of these companies to follow and monitor news sources in e-commerce, social media, and forums provides important information to businesses in the decision-making process. With the large amount of data shared in these resources, sentiment analysis can be made from people's comments about services and products, users' emotions can be extracted and important feedback can be obtained. All of this is of course possible with accurate sentiment analysis. In this study, new data sets were created for Turkish, English, and Arabic, and for the first time, comparative sentiment analysis was performed from texts in three different languages. In addition, a very comprehensive study was presented to the researchers by comparing the performances of both the pre-trained language models for Turkish, Arabic, and English, as well as the deep learning and machine learning models. Our paper will guide researchers working on sentiment analysis about which methods will be more successful in texts written in different languages, which contain different types and spelling mistakes, which factors will affect the success, and how much these factors will affect the performance.

1. Introduction

Sentiment analysis (opining mining) is to reveal the general attitude of a community on a subject by evaluating their feelings and thoughts about a subject (Petz, 2014). Sentiment analysis is frequently used in social media analysis, traditional financial markets, and cryptocurrency markets to examine the behavior of users (Obaidi et al., Nov. 2022). In the financial market, users can evaluate their perspectives on the market and their own investments with sentiment analysis. With social media platforms on the Internet, many people can express their thoughts or feelings about a product, brand, or service through text (Chen et al., 2019; Rambocas and Pacheco, 2018). By automatically analyzing customer feedback from survey responses and social media conversations, brand-owning companies can carefully follow their customers' insights and develop various sales strategies. They can also customize their products or services to meet their customers' needs (Chakraborty et al., 2020). With Sentiment analysis, brands are easier to reach customers who need extra care, which will enable marketers to reach all types of customers, thus increasing customer satisfaction and sales. Different methods such as manual operation, keyword processing, and natural language processing are used to determine what type of emotion a comment or thought contains. Manually, the human element is necessary when interpreting the complexity of languages in thought, such as context, ambiguity, and irony. However, there is too much data to handle manually (Vyas, 2019; Dhaoui et al., 2017). With keyword processing, a sensitivity score is assigned to certain words or phrases according to their emotion classes. These scores are used to find the total weight of the text in which they are cited. Thus, with the final total score, it is found to which emotion class the text belongs. Natural language processing, on the other hand, makes use of machine learning methods in which semantic techniques are used in sentiment analysis. As in all machine learning methods, a predictive model is trained to make predictions about the task using datasets called corpus. By using the features of the language in which the dataset is written, it is estimated that the text belongs to the negative, neutral and positive emotion class by capturing the contexts correctly (Naresh, 2021; Kemaloğlu et al., 2021).

E-mail address: bihterdas@firat.edu.tr (B. Das)

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.02.017
1319-1578/© 2023 The Author(s). Published by Elsevier B.V. on behalf of King Saud University.

The paper makes the following main contributions:

- For the first time, a comparative sentiment analysis has been carried out for Turkish, Arabic, and English texts.
- Three different new datasets containing e-commerce data for Turkish, English, and Arabic were created.
- The performances of popular deep learning, machine learning approaches, and pre-trained language models on texts written in different languages that included irregular, different word types and misspellings were compared and the contribution of these approaches on the success was examined.
- The connections between the approach criteria examined and the performance results obtained were investigated.

The rest of this paper is organized as follows: In Section 2, we present the related works in the literature. In Section 3, we provide fundamental information about the dataset, the preprocessing, classifiers, and validation method. Section 4 contains the experimental results and discussion, and Section 5 concludes the paper.

1.1. Motivation

Machine learning (ML) and deep learning (DL) approaches have been used in sentiment analysis studies on texts written in different languages from the past to the present. In addition, pre-trained language models, which are used recently, trying to predict masked words in a text, using transformer architecture and trained on huge amounts of data, have also started to be preferred frequently in natural language processing applications. Since the dataset, which is used by pre-trained language models for training, mostly consists of meaningful, regular, and standard sentence structures, they often achieve high performance in text classification and sentiment analysis applications. In this study, it is aimed to examine the effect of the approaches on the success to be achieved by comparing the performances of the most popular deep learning, machine learning methods and pre-trained language models for sentiment analysis on the texts written in different languages. In the study, product reviews and complaint data received from the e-commerce site, which consists of elements that affect negatively the classification, such as irregular shipments, short texts, different spelling mistakes, and many word types, were used. Thus, it was desired to observe how the high performance of pre-trained language models trained on regular and large data would change in irregular and smaller data. Also, the effect of the language in which the texts were written on the performance of other ML and DL methods in sentiment analysis was observed and the connections between the examined criteria and the results obtained were investigated. Our study will guide on sentiment analysis researchers about which methods will be more successful in texts that contain irregular, different word types and spelling mistakes written in different languages, what factors may affect success, and how much these factors will affect the performance.

2. Related works

In this section, information is given about the methods used for sentiment analysis in texts written in Arabic, Turkish and English and the highest performances obtained with these methods.

2.1. Turkish sentiment analysis

Various sentiment analysis studies have been carried out on e-commerce data written in Turkish, data based on social media, financial studies, Turkish stock market comments, Instagram-Twitter comments, and airline passengers' comments. Demircan et al. (Demircan et al., 2021) executed a sentiment analysis on product comments on e-commerce sites with machine learning methods. Comments were examined in three classes positive, neutral, and negative. The machine learning methods used are support vector machine (SVM), random forest (RF), decision tree (DT), logistic regression (LR), and k-nearest neighbors (k-NN). Experimental results showed that SVM and RF models outperformed other models. Iskender et al. (İskender and Batı, 2015) carried out a sentiment analysis study by comparing the rankings of Turkish Universities with the thoughts and social media comments of the students of the relevant university. In the study, 13,007 tweets (with the entrepreneur keyword) and 14,579 tweets (with the innovation keyword) data, Support Vector Machines, Multinomial Naive Bayes (MNB), and data mining algorithms were used. In the experiment, the MNB method achieved 90 % success for positive and negative classes. Sariman et al. (Sariman ve and Mutaf, 2020) made sentiment analysis using Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) from 2,000,000 tweets during the Covid-19 process. 82 % success has been achieved using Logistic Regression (LR) on the subject of curfew. Pervan et al. (Pervan and Yalım Keleş, 2017) obtained 84.23 % accuracy by using the Random Forest method, positive and negative, from the comments of the mobile phone they received from the e-commerce site.

Celik et al. (Celik ve and Aslan, 2019) collected 8770 comments from the Facebook site. They made a sentiment analysis study using Random Forest, Logistic Regression, Naive Bayes, and K-nearest Neighbor (k-NN) methods. In the study, they achieved 74.13 % success with LR. Karayigit et al. (Karayiğit et al., 2021) executed a study to detect abusive comments in Turkish on social media. The CNN method has been compared with traditional machine learning methods. The model proposed in the study offers ideas for the creation and development of Turkish comment filters. 0.946 kappa value obtained by the CNN model. Aktas et al. (Aktas et al., 2021) collected 676 thousand data from the comments of a food website and used Artificial Neural Network(ANN), K-NN, and NB methods. They obtained the highest accuracy with ANN at 86 %. Kumas et al. (Kumas, 2021) collected Twitter data and conducted sentiment analysis with positive and negative outputs. In the study, they compared the performances of NB, SVM, k-NN, and LR methods. It has the highest F-score with RF of 84.23 %. Parlar et al. (Parlar et al., 2017) collected 2043 tweets related to Telecommunications from Twitter. They achieved 78 % success performance with ME.

2.2. English sentiment analysis

The language used in the studies in which sentiment analysis is performed the most is English. Many sentiment analyses studies have been carried out in the fields of social media comments, news, Covid-19 vaccine, health, and financial fields. Devika et al. (Devika et al., 2016) presented a comparative study in terms of performance for sentiment analysis of different approaches such as SVM, NB, and rule-based approaches. They examined opinions about the vaccine in three groups positive, negative and neutral using Machine learning, Rule-based and Lexicon- based approaches. The highest performance accuracy was obtained with the Rule-based approach, 91 % at the review and 86 % at the sentence level. Sunitha et al. (Sunitha et al., 2019.) performed sentiment analysis on 89,743 tweets from Twitter and 14,641 tweets about the US airline. In the study, they compared the performances of NB, RF, SVM, and k-NN methods. They achieved the highest performance with SVM at 83.67 % on Twitter and 69 % with SVM in the airline dataset. Al-Hadhrami et al. (Al-Hadhrami et al., 2019) compared the SVM, RF, and K-Mean Clustering methods in a two-class

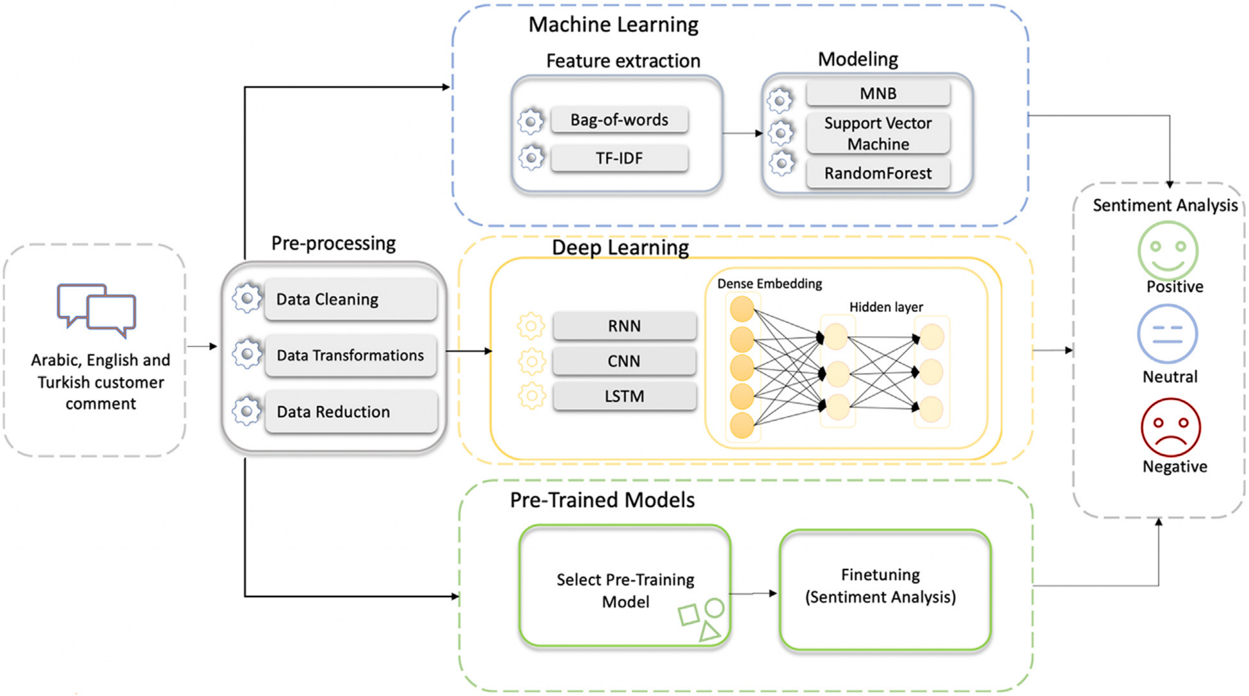


Fig. 1. System architecture.

sentiment analysis study, positive and negative, with data from Twitter. The highest accuracy was obtained with K-means Clustering at 74 %. Yan et al. (Yan et al., 2014) carried out a sentiment analysis using Chinese and English social media comments. Stop words in the texts were removed and *n*-gram and SVM methods were used to process the text. They achieved 98 % accuracy with SVM and 82 % accuracy with *N*-Gram. Kumar et al. (Kumar et al., 2021) executed a study to classify as positive or negative using the US airway dataset for training 4,392 tweets for testing from 10,248 tweets. They compared ANN, Decision Tree (DT), and SVM. The highest accuracy was obtained with ANN at 75.99 %. Studies have stated that CNN, Long Short-Term Network (LSTM), and BiLSTM models show high success in classifying both short and long texts (Trueman and Cambria, 2021; Baziotis et al., 2017; Basiri et al., 2021). Minaee et al. (Minaee et al., 2019) studied a sentiment analysis for IMDB movie reviews using CNN and BiLSTM deep learning methods. They achieved 89 % success with LSTM, 89.3 % with the CNN Model, and 90 % success with the proposed ensemble of LSTM and CNN model in the IMDB dataset. In (Basiri et al., 2021), the authors made a sentiment analysis using social media data. In the study, they used a structure combining LSTM, Gated Recurrent Units (GRU), and CNN models, as well as an attention mechanism to focus on essential words. The highest accuracy performance was obtained with the attention-based bidirectional CNN-RNN deep model at 93.40 in the Kindle dataset.

2.3. Arabic sentiment analysis

In the last 10 years, studies on sentiment analysis in Arabic datasets have started to increase (Yue et al., 2019; Zhang et al., 2018; Hussein, 2018). People's comments and thoughts on the Internet against the events taking place in the Arab regions have started to attract attention, so the Arabic data content continues to grow. Sentiment analysis studies, sentiment estimation, and various inferences were made by collecting comments on news data, social media, forums, blogs, and websites. Abdelgwad et al. (Abdelgwad et al., 2021) performed an aspect-based sentiment analysis on hotel reviews written in Arabic using GRU and CNN methods. With the Interactive attention network-based bidirectional Gated Recurrent Units (IAN-BGRU) model, they achieved 83.98 % accuracy for aspect polarity detection. Alwehaibi et al. (Alwehaibi et al., 2022) conducted a sentiment classification study on Twitter short texts written in Arabic using LSTM, CNN, and community models. They achieved a classification performance of 96.7 % with the ensemble model. Alassaf et al. (Alassaf and Qamar, 2022) used a one-way analysis of variance (ANOVA) for feature classification to perform sentiment classification on tweets written using Arabic. They compared the performance of SVM, LR, k-NN, NB, and Multilayer Perceptron (MLP) methods. The highest F1-score was obtained with the SVM method 88 % for ANOVA p-value experiment. In (Heikal et al., 2018), the authors performed sentiment analysis on three different datasets written in Arabic and compared the performances of deep learning models such as CNN and LSTM with ensemble method. They have shown that the ensemble method obtained 65.05 % accuracy performance. Brahimi et al. (Brahimi et al., 2021) made a sentiment analysis of the Arabic movie review comments. With F-measure, a classification result of 96 % was obtained. In summary, in Arabic sentiment analysis studies, mostly bigram and trigram feature vectors were used, and supervised machine learning methods were used for classification (Rushdi-Saleh et al., 2011; Duwairi and El-Orfali, 2014; Aly and Atiya, 2013).

3. Materials and methods

In this study, a sentiment analysis study with 3 classes (positive, negative, and neutral) was actualized from the texts written in Arabic, Turkish and English. 3 new datasets were created in 3 different languages in 100 K size, then the data were freed from noise by pre-processing. The preprocessed data was digitized and the data was divided into 70 % training, 20 % validation, and 10 % testing. Data validation was performed with k = 10 for data accuracy. The datasets are made ready for performance comparison in machine learning, deep learning, and pre-trained models. In

Table 1
The detailed information on the data.

Product Category	Language	Number of Samples	Number of Websites
Movie	English, Arabic	80 K	1
Game	English, Arabic	50 K	2
Small appliances	Turkish	40 K	3
Technological products	Turkish, English, Arabic	80 K	4
Large home appliances	Turkish	50 K	4

Table 2
Example of the Turkish/English/Arabic dataset.

Languages	Negative	Neutral	Positive
Example in Turkish	ürün hasarlı ve paketlemesi cok özensiz geldi geldiği gibi iade ettim	biraz yavaş çalışıyor ama fena değil	müthiş performans, sessiz çalışıyor, herkese tavsiye ederim
Example in English	keeps not responding after one solved puzzle very frustrating for my child	i love it but i need to purchase that's really not good	it came just as I expected i really like it you can buy it with confidence the cargo delivery was also fast
Example in Arabic	استلمت المنتج مكسور ومتأخر جداً	المنتج كان جيد ولكن التغليف كان جدا سيئ	اشتريتها لأبنتي وهي تحبه كثيرا، بالنسبة للسعر والحجم جدا ممتاز

Machine Learning(ML) part, the tf-idf and Bag-of-words are applied to the data for features extraction. MNB, SVM, and RF methods, which are among the most popular ML methods, were applied to the dataset in 3 different languages and sentiment analysis process was performed, which created 3-class outputs. In the Deep Learning part, the pre-processing steps were applied again, the sequence length was set to 120 and the short sequences were filled with 0. An embedding layer has been created on the text, which converts the convolution network data into fixed-size vectors. Afterward, CNN, RNN, and LSTM layers were created separately. To prepare the model for training, the loss function was applied with the optimizer 'Adam' and 'categorical_crossentropy'. After the training, a 3-class sentiment analysis output was obtained for the dataset written in three different languages. In the pre-trained model part, 3 different pre-trained models were studied for the dataset in 3 different languages. The bert-base-multilingual-cased model has been applied jointly for datasets in three languages. Also, for the Arabic dataset, xml-roberta which is a pre-trained transformer model on a self-monitoring large body, and distilbert-base-multilingual-cased models, which are twice as fast as mBert-base, were studied. In the Turkish dataset, convbert-base-turkish-cased and distilbert-base-turkish-cased models published as Turkish packages were studied. Finally, in the English dataset, the common bert-base-multilingual-cased model, as well as the distilroberta-base model and *albert-base-v2* pre-trained language models, were used. Fig. 1 shows the flowchart of the experiment of sentiment analysis.

3.1. Datasets

In this study, three new datasets written in Arabic, Turkish and English languages were created. The data was collected from e-commerce sites containing product reviews and complaints. Also, the data size is limited to 100 K for all three languages. An equal number of samples were collected for the positive, negative, and neutral categories. Table 1 provides detailed information about the dataset. Selenium Web driver was used to collect data from websites.

Duplicate data, missing data, numeric data, punctuation marks, and emojis have been cleaned and upper/lower letters have been converted on the created datasets. In addition, the lemmatization process was performed to group different inflected word forms into root forms with the same meaning. Data from multiple channels are aggregated to standardize the data. Data is divided into columns. It was later converted to.csv format. Since the data needs to be scaled to represent it in a smaller range, the data is digitized with a one-hot encoder and the data is separated by data splitting. Afterward, data validation was performed. Table 2 shows an exam-

Table 3
The parameters of the pre-trained models.

Models	Parameters	Hidden Layer Activation	Hidden Size	Initializer range	Intermediate size	Max position embeddings	Attention heads	hidden layers	Batch size	Learning Rate	Number of epochs
bert-base-multilingual-cased	110 M	gelu	768	0.02	3072	512	12	12	16	2e-5	6
xlm-roberta-base	250 M	gelu	768	0.02	3072	514	12	12	16	2e-5	6
distilbert-base-multilingual-cased	134 M	gelu	768	0.02	3072	512	12	6	16	2e-5	6
convbert-base-turkish-cased	106 M	gelu	768	0.02	3072	512	12	12	16	2e-5	6
distilbert-base-turkish-cased	134 M	gelu	768	0.02	3072	512	12	6	16	2e-5	6
albert-base-v2	11 M	gelu_new	768	0.02	3072	512	12	12	16	2e-5	6
distilroberta-base	82 M	gelu	768	0.02	3072	512	12	6	16	2e-5	6

ple of the Turkish/English/Arabic datasets. In addition, the created datasets have been uploaded to GitHub (Pages, 2022).

3.2. Preprocessing

Preprocessing steps consist of data cleaning, data transformation, and data reduction stages. In the data cleaning section, punctuation marks, emojis, numerical data, and missing data in sentences have been cleaned and letters have been reduced. In the data transformations section, the user comments collected from the websites were filtered and made usable. In the data reduction section, the data is analyzed and the inappropriate data is deleted and prepared in.csv format. It is used as such in pre-trained language models. In the text vectorization step of the cleaned data in machine learning algorithms, the document representation model BOW, which is widely used in text processing, and TF-IDF, which is a weighting method commonly used in text processing, are used. In this method, the frequency of each word (TF) is represented by multiplying the inverse document frequency (IDF). Thus, it reduces the importance of words that are repeated a lot and increases the importance of words that contain fewer words. In the preprocessing step of deep learning algorithms, data tokenization preprocessing was performed instead of text vectorization. In this part, the Keras library is used. The fit_on_texts arguments are used to create a word index based on word frequency, and the texts_to_sequences arguments are used to convert each text to an integer sequence. Also, pad_sequences is used to fill the data strings in the dataset with the same length.

3.3. Classifiers

In this section, information about deep learning methods, machine learning approaches, pre-trained language models, and their parameters in the experiment is presented.

A. Pre-trained language models

In this section, information is given about the pre-trained language models used for Turkish, English and Arabic texts.

Language models for Turkish

Bert-base-multilingual-cased: This model is a pre-trained transformer model with Wikipedia in 104 languages, using masked language modeling (MLM), and is case sensitive. In the model, 15 % of the words are randomly masked as input with MLP and the masked words are predicted by running the entire masked sentence over the model. In addition, next sentence prediction (NSP) in this model predicts whether two sentences follow each other.

Convbert-base-turkish-cased: This language model also trained a ConvBERT model on the Turkish part of the mC4 corpus. In this model, a sequence length of 512 is used during the full training time (Oflazer and Saraclar, 2018).

Distilbert-base-turkish-cased: This language model is a cased distilled BERT model designed for Turkish, over 7 GB of the original training data used in BERTurk's training.

Language models for English

Bert-base-multilingual-cased: In the experiment, this pre-trained language model was used for sentiment analysis from English texts as well as for Turkish texts.

Distilroberta-base model: Distilroberta-base, a distilled version of the RoBERTa-based model, has the same training procedure as Distilbert. It can show different performances for different languages. This pre-trained language model, which has 6 layers, 768 dimensions, 12 heads, is case sensitive and has a total of 82 M parameters. Also, this language model is much faster than the Roberta-base language model.

Albert-base-v2: ALBERT is a transformer model pre-trained on large English data and has self-monitoring capability. With BERT, the number of hidden layers is the same and the computational cost is the same. This caseless model has 128 embedding dimensions, 12 attention heads, and 11 M parameters.

Language models for Arabic

Bert-base-multilingual-cased: In the experiment, this pre-trained language model was used for sentiment analysis from Arabic texts as well as for Turkish and English texts.

Xml-roberta: This language model used for Arabic texts is different from Bert and has more vocabulary. It contains about 250,000 tokens.

Distilbert-base-multilingual-cased: It is a leaner version of the Bert-based multilingual model, a cased pre-trained language model. The case-sensitive model, which has 6 layers, 12 heads, and 134 M parameters, is trained on a Wikipedia dataset written in 104 different language. Table 3 shows the parameters of all pre-trained language models.

B. Deep learning methods

Traditional machine learning methods keep related words in short-term memory in classifying texts for natural language processing applications. Deep learning methods, on the other hand, keep words in memory for a longer time. The most widely used deep learning methods in natural language processing are Recurrent Neural Network, Long Short-term Memory, Bidirectional LSTM, Gated Recurrent Unit, and Convolutional Neural Network.

Recurrent Neural Network

The Recurrent Neural Network method, which is used in natural language processing, keeps in memory that the words in the sentence are related to the words before and after it. Therefore, it achieves very good results in applications such as sentiment analysis, text classification, and question answering. The biggest difference between RNN from deep learning models is that these models remember. In addition, while in other neural networks each input is independent of the other, in RNNs the inputs are related to each other. RNNs make associations between inputs to follow the next step and remember all their associations while they are being trained. The disadvantages of RNNs are slow computation and difficulty in accessing long-ago information (Basiri et al., 2021). Table 4 shows parameters of the RNN model.

Long-Short Term Memory

Long-short Term Memory method is one of the models used to overcome the difficulties of RNN architecture. LSTM models outperform RNNs in speech recognition in learning context-sensitive languages. Unlike standard neural networks, the LSTM architecture has feedback connections. It remembers values at random times with feedback networks. It is frequently used in Recursive Neural Networks. It eliminates problems such as long forgetting problems caused by RNN. This architecture is used in many fields. It is frequently used in areas such as handwriting recognition, speech recognition, and anomaly detection. LSTM basically consists of a cell, an input gate, an output gate, and a forget gate. The cell remembers values at random intervals. All three doors regulate the entry and exit of information entering the cell. LSTM is frequently used in time series analysis because LSTM has the ability

Table 4
Parameters of the RNN model.

Hyper-parameter	Value
Dropout rate	0.50
Learning rate	0.001
Batch size	10
Activation function	Sigmoid
Loss	Categorical_crossentropy
Optimizer	Adam

Table 5
Parameters of the LSTM model.

Hyper-parameter	Value
LSTM hidden state dimension	196
Dropout rate	0.5
Learning rate	0.001
Batch size	10
Activation function	Softmax
Loss	Categorical_crossentropy
Optimizer	Adam

to learn long-term dependencies. It addresses the long-term dependency issues that arise and the vanishing gradient issues in time series analysis (Arbane et al., 2023). Table 5 shows parameters of the LSTM model.

Convolutional Neural Networks

Convolutional neural networks are a model developed for images and multidimensional data. CNN, which consists of cascading trainable sections, consists of the input layer, convolution, pooling, and fully connected layers. In these models, the training process is carried out by performing layer-by-layer operations after receiving the input data. Finally, it gives a final output for comparison with the correct result. An error occurs as much as the difference between the produced result and the desired result. The backpropagation algorithm is used to transfer this error to all weights. The weights are updated with each iteration to reduce the error. In CNN, the input data is an image, audio, video, and more recently text. CNN models have been applied to various problems in the field of natural language processing and successful results have been obtained. CNN's are widely used in text classification, named entity recognition, text sorting, and question-answering applications (Ni et al., 2021). Table 6 shows parameters of the CNN model.

C. Machine learning methods

Machine learning, which is used in natural language processing, has a structure that trains the system with some training data for information extraction. Machine learning methods, which are mostly used in text classification and sentiment analysis, are examined in two parts supervised and unsupervised learning algorithms. While classification is made with labeled training data in supervised learning, unlabeled data is used in unsupervised learning.

Table 6
Parameters of the CNN model.

Hyper-parameter	Value
LSTM hidden state dimension	196
Dropout rate	0.5
Learning rate	0.001
Batch size	10
Activation function	Softmax
Loss	Categorical_crossentropy
Optimizer	Adam

Random Forest

Random Forest, one of the supervised classification algorithms, is formed by the formation of more than one decision tree. It is mostly used in regression and classification problems. In this algorithm, the input vector is assigned to each tree and a result is produced at the output of these trees (Chen et al., 2022).

Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is the application of Bayes' theorem to predict which category a sample is in, with the strong assumption that one feature is independent of another feature. This algorithm is a probabilistic classifier and calculates the probability of each category using Bayes' theorem and determines the category with the highest probability of a sample as the output. Naive Bayes classifiers have been successfully implemented in NLP applications (Singh et al., 2020). This algorithm, predicts a text's label. Multinomial Naive Bayes classification algorithm is also a widely used algorithm in sentiment analysis.

Support Vector Machine

Support Vector Machine is one of the supervised algorithms, a simple but effective algorithm mostly used in text classification. The training data labeled with the created vector model are classified. A boundary line called the hyperplane separates the groups. The goal is to find the hyperplane that creates the greatest separation between the classes. Although it was developed for two-class problems, it is also preferred for solving multi-class problems (Jaya Hidayat et al., 2022).

In this study, for feature extraction in machine learning methods, the word frequency was measured with the Bag-of_words used, and the word counts were changed with TF-IDF scores in the whole dataset by giving weight to the words with TF-IDF. "n_estimators = 20" hyperparameter is used for the RandomForestClassifier model created with Random Forest Algorithm. The SVM model was created with the classifier object created from the scikit-learn library, SVC class, and it was specified as "kernel = linear". Other parameters are used by default. Parameters are used by default in the MultinomialNB() class by using the scikit-learn library in the Multinomial Naive Bayes algorithm.

3.4. Validation method

In predictive modeling problems, the cross-validation method, which helps to compare a suitable model, is also used to evaluate and test the performance of models. The cross-validation method tends to have a lower bias than other methods in counting the model's efficiency scores. In this study, k-fold cross-validation with k = 10 was applied for validation and the models were evaluated in terms of estimation accuracy.

4. Experimental results and discussion

The performances of deep learning models such as RNN, CNN, LSTM, machine learning such as MNB, SVM, RF and pre-trained language models were compared for sentiment analysis on customer comments in Turkish, English and Arabic, collected from e-commerce sites. The comparison of machine learning and deep learning methods in the experimental study is given in Fig. 2.

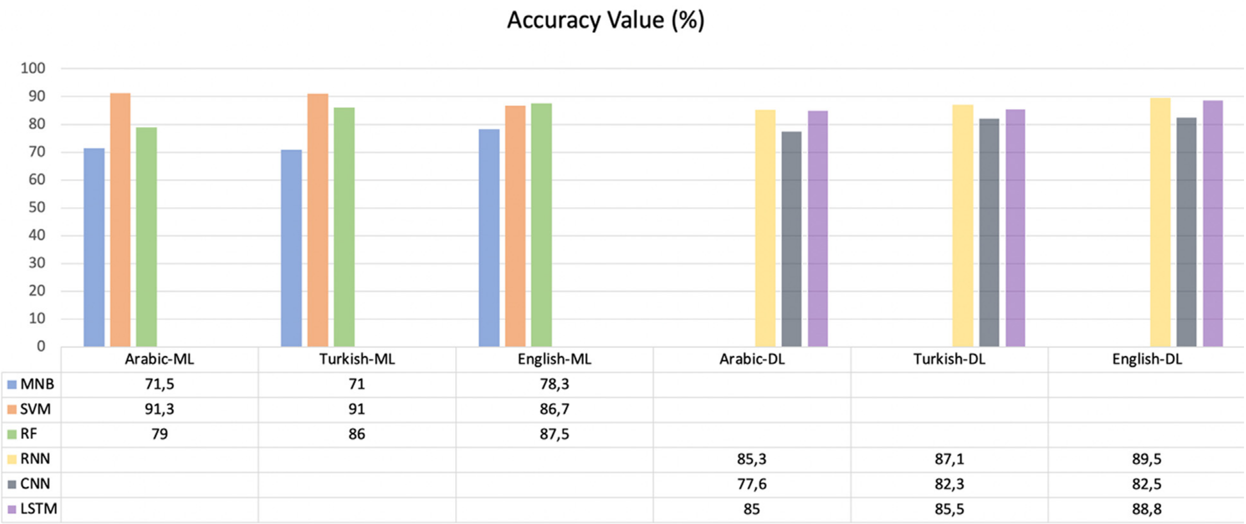


Fig. 2. Performance comparison of DL and ML methods for three languages.

Table 7
Performance comparison of machine learning models for three languages.

		MNB				SVM				RF			
		ACC (%)	Pre	Rec	F1-s	ACC (%)	Pre	Rec	F1-s	ACC (%)	Pre	Rec	F1-s
Turkish	Negative	71	0.56	0.78	0.65	**91**	0.81	0.81	0.81	87	0.66	0.70	0.68
	Neutral		0.64	0.68	0.66		0.86	0.94	0.90		0.72	0.79	0.75
	Positive		0.84	0.47	0.60		0.84	0. 0.77	0.81		0.74	0.62	0.68
English	Negative	78.3	0.71	0.85	0.77	86.7	0.81	0. 0.82	0.82	**87.5**	0.79	0.79	0.79
	Neutral		0.69	0.59	0.63		0.74	0 0.77	0.76		0.70	0.79	0.74
	Positive		0.81	0.77	0.79		0.91	0 0.86	0.88		0.88	0.78	0.83
Arabic	Negative	71.5	0.74	0.72	0.73	**91.3**	0.80	0.85	0.83	79	0.66	0.78	0.71
	Neutral		0.65	0.70	0.67		0.78	0.76	0.77		0.69	0.59	0.63
	Positive		0.76	0.72	0.74		0.84	0.81	0.83		0.73	0.70	0.71

ACC: Accuracy, Pre: Precision, Rec: Recall, F1-s: F1-score.

As seen in Fig. 2, in the RF method, which is one of the machine learning models, the language that shows the best accuracy value over 100 K data is English. In Arabic and Turkish datasets, the SVM method achieved the best performance. Table 7 shows the performance results of deep learning methods such as MNB, SVM, and RF for three languages.

Although RF shows high performance only for the English language, the SVM method has reached high accuracy values for all languages. The reason why the SVM method has better performance than other methods is that it uses overfitting protection that does not need to be dependent on the number of features and can handle high dimensional input space. The SVM method tries to identify a few irrelevant features in a text with feature selection. One way to avoid these high-dimensional input fields is to assume that most of the features are irrelevant. In the SVM method, the corresponding document vector for each document is sparse with only a few non-zero entries. In text classification and sentiment analysis, algorithms with similar inductive bias such as SVM are more suitable for problems with sparse samples. Also, most text classification problems are linearly separable. In the work of SVMs, there is the idea of finding separators such as linear or polynomial, RBF.

In addition, deep learning classification results performance in datasets in three languages have achieved good results in datasets in RNN and LSTM classification tasks. It is seen that deep learning applications have better success on the English dataset compared to other datasets. It has been seen that the LSTM algorithm achieves almost the same performance on three datasets, and the CNN method has very close performance values in Turkish and English. Table 8 shows the performance results of deep learning methods such as RNN, CNN, and LSTM for three languages.

Although the RNN and LSTM models showed very close performance values, the RNN model had the highest accuracy. Since LSTM can process information in memory for a long time compared to RNN, it has a feature that it can memorize the order of data. It works on eliminating unused information that can aid in text classification. The reason why RNN outperforms other methods is that it is designed to allow RNN to exhibit temporal behavior and to capture sequential data when the current step has some type of relationship to previous steps. It makes it a more natural approach when dealing with textual data as text is naturally sequential. RNNs perform very well for applications such as text classification and sentiment analysis where sequential information is clearly important because meaning can be misinterpreted or grammatically incorrect if sequential information is not used. RNN can also model the string of text and capture long-term dependencies. Therefore, RNN also performs better in tasks where the length of the text is important. CNN, on the other hand, works better in applications such as Named Entity Recognition where feature detection is more important in the text, but cannot achieve as well

Table 8
Performance comparison of deep learning models for three languages.

			RNN				CNN				LSTM		
		ACC (%)	Pre	Rec	F1-s	ACC (%)	Pre	Rec	F1-s	ACC (%)	Pre	Rec	F1-s
Turkish	Negative	**87.1**	0.89	0.82	0.85	82.3	0.76	0.84	0.80	85.5	0.84	0.88	0.86
	Neutral		0.92	0.93	0.92		0.89	0.89	0.89		0.88	0.88	0.88
	Positive		0.82	0.87	0.84		0.82	0. 0.74	0.78		0.86	0.81	0.83
English	Negative	**89.5**	0.93	0.89	0.91	82.5	0.88	0. 0.81	0.84	88.8	0.90	0.89	0.89
	Neutral		0.85	0.86	0.85		0.71	0 0.86	0.78		0.83	0.86	0.84
	Positive		0.91	0.94	0.92		0.94	0 0.80	0.87		0.94	0.92	0.93
Arabic	Negative	**85.3**	0.86	0.90	0.88	77.6	0.79	0.85	0.82	85.0	0.86	0.89	0.87
	Neutral		0.82	0.79	0.81		0.73	0.72	0.72		0.84	0.78	0.81
	Positive		0.87	0.85	0.86		0.82	0.73	0.77		0.85	0.87	0.86

ACC: Accuracy, Pre: Precision, Rec: Recall, F1-s: F1-score.

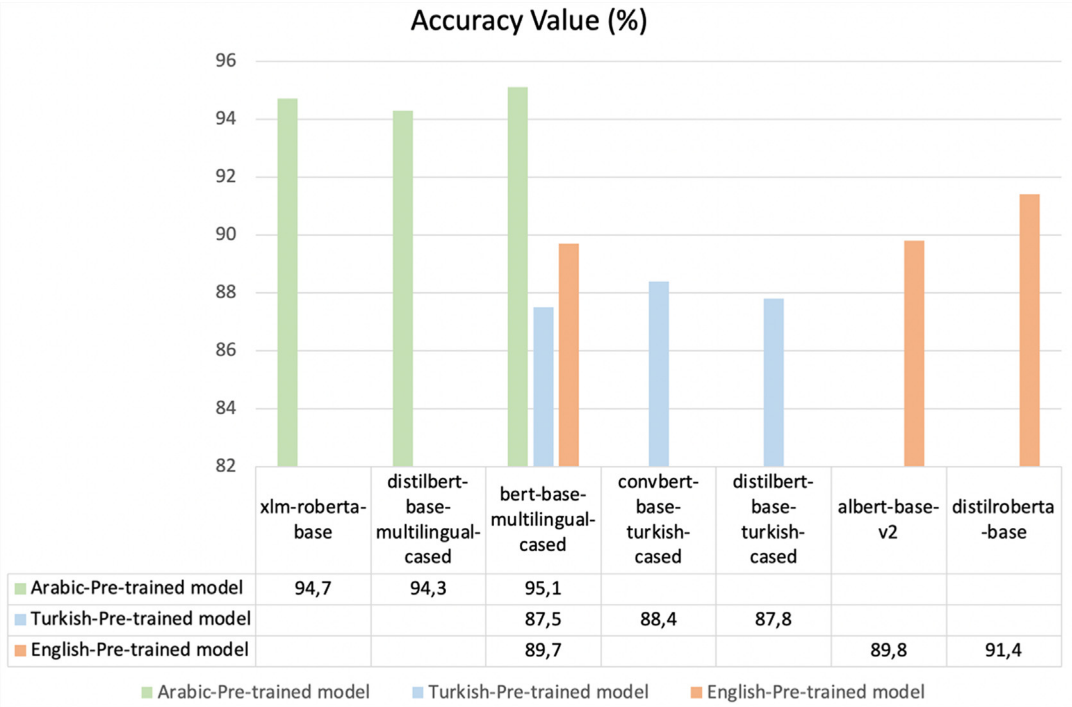


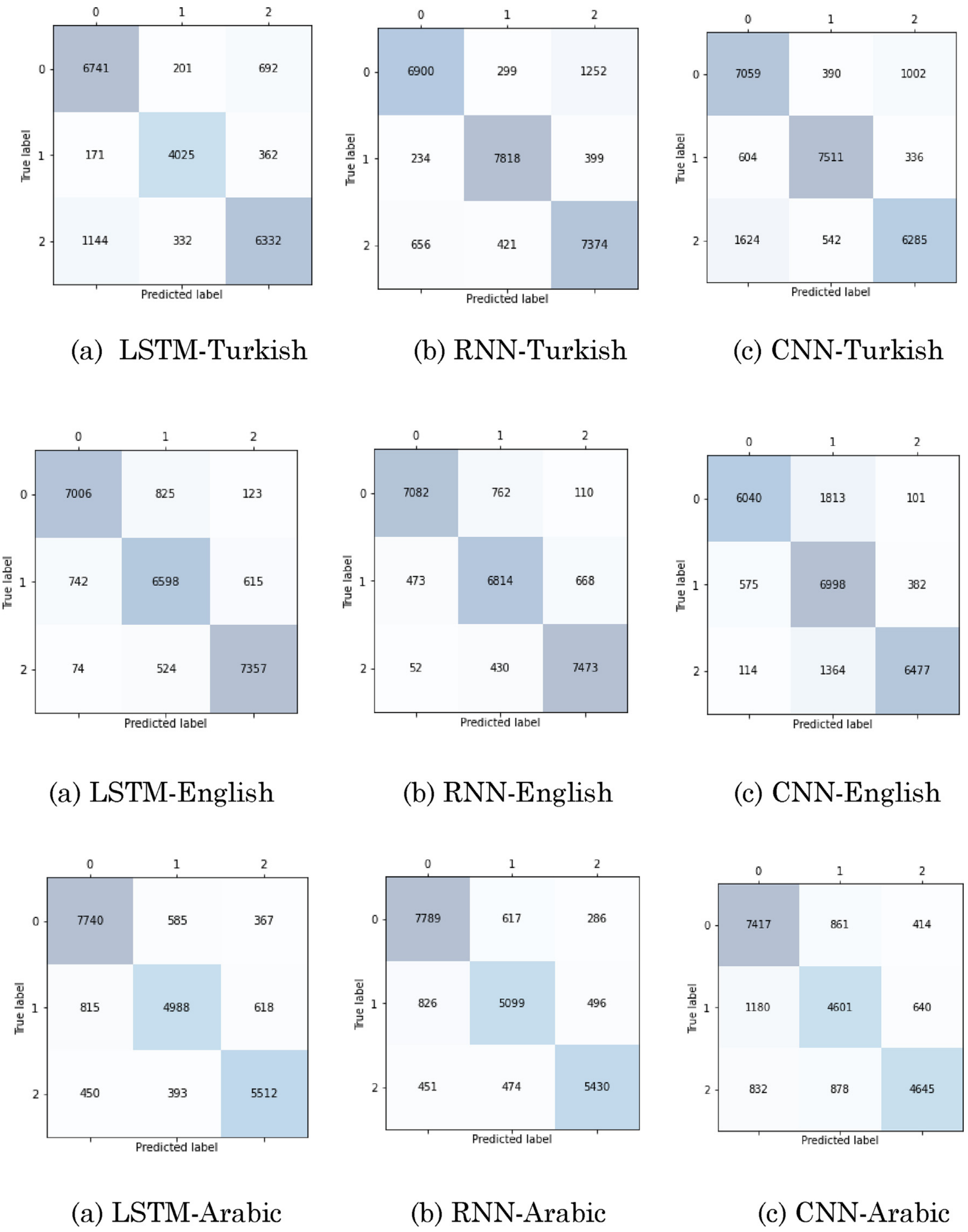
Fig. 3. Performance comparison of pre-trained language models for Turkish, English, Arabic.

as RNN and LSTM in tasks where sequential modeling is more important.

In the study, the performances of pretrained models for sentiment analysis in English, Turkish and Arabic datasets were compared. Fig. 3 shows the comparison of the pre-trained language models. The common language model for the three languages is the 'Bert-base-multilingual-cased' model and this model achieved the best performance in the Arabic dataset. On the Arabic dataset, the other pre-trained language models 'xml-roberta' and 'distilbert-base-multilingual-cased' models achieved almost the same performance. The 'distilroberta-base model', a transformer-based language model in the English dataset, outperformed the other pre-trained models used for the English dataset. In the Turkish dataset, although the 3 pre-trained models show almost similar performances, the model that produces the best results is 'convbert-base-turkish-cased'. As a result, in the comparison made, Arabic pre-trained models performed very well in base packages.

The confusion matrix is given in Fig. 4 to examine the statistical success of the deep learning methods according to Turkish, English, and Arabic, respectively.

Although Multilingual BERT is not specifically designed for Arabic, it supports multiple languages, including Arabic. In the text written in Arabic, Bert-base-multilingual-cased outperformed the others. The ability to pre-train from unlabeled text and predict both left-hand and left-hand conditional words affect this performance. Bert-base-multilingual-cased, works well on cross-lingual transfer tasks, than static non-contextualized word embeddings. The Bert-base-multilingual-cased language model was able to find billions of Arabic token texts to train their models. Wikipedia and news sites and Twitter have also been an important source of data, because Saudi Arabia and Egypt are among the leading countries by the number of Twitter users since January 2022. For the Turkish language, Conbert-base-turkish-cased outperformed other language models. Since the extraction of local features increases con-



(a) LSTM-Turkish (b) RNN-Turkish (c) CNN-Turkish

(a) LSTM-English (b) RNN-English (c) CNN-English

(a) LSTM-Arabic (b) RNN-Arabic (c) CNN-Arabic

Fig. 4. Confusion matrix of the sentiment analysis for deep learning methods.

volution success, convolution layers are used as a complement to personal attention in the pre-training phase in this language model. In addition, the reason why the Convbert-base-turkish-cased model is more successful may be due to the fact that this model uses the convolution kernel to better capture local similarities between words and combines them with self-attention. For the English language, diltilroberta-base has the highest performance among all models. While Bert uses static masking in the text by masking the same part of the sentence in each epoch, distilroberta-base uses dynamic masking and mask different parts of the sentences for different epochs. Thus, it solidify the model and improve the performance.

5. Conclusion

In this study, three corpus containing e-commerce data belonging to different languages such as Turkish, Arabic, and English were created and the performances of deep learning and machine learning methods were examined comparatively by performing sentiment analysis on them. In order to extract positive, negative, and neutral sentiment classes from the data, different pre-trained language models were also used, and the contribution of these models to the accuracy performance in sentiment analysis was investigated. Experimental results showed that pre-trained language models achieved good performance results in three class sentiment analyses. In Arabic, the bert-base-multilingual-cased pre-trained model achieved the best performance with 95.1 %. In English, distilroberta-base model achieved the highest accuracy with 91.4 %. While the RNN model, one of the deep learning models, achieved the best performance on the English dataset with 89.5 %, SVM, one of the machine learning methods, also achieved the highest performance on the Arabic dataset with 91.3 %. In future studies, we plan a sentiment analysis study in the fields of education, politics, finance, health and technology. In addition, we create a very large unlabeled dataset and design a study.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the Republic of Turkey, Ministry of Science, Technology and Industry project named "AI-Based Smart Digital Assistant Customer Dialog Bot project" and project code AR-22-087-0001. It is funded by an R&D project within the scope of law 5746 by the Arçelik Digital Transformation, Big Data, and Artificial Intelligence R&D Center.

References

Abdelgwad, M.M., Soliman, T.H.A., Taloba, A.I., Farghaly, M.F., Sep. 2021. Arabic aspect based sentiment analysis using bidirectional GRU based models. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2021.08.030.

Aktas, O., Coskuner ve, B., Soner, I. 2021. "Turkish Sentiment Analysis Using Machine Learning Methods: Application on Online Food Order Site Reviews", Journal of Artificial Intelligence and Data Science, c.1, sayı.1 ss.1-10, 2021.

Alassaf, M., Qamar, A.M., 2022. Improving Sentiment Analysis of Arabic Tweets by One-way ANOVA. Journal of King Saud University - Computer and Information Sciences 34 (6, Part A), 2849–2859. https://doi.org/10.1016/j.jksuci.2020.10.023.

Al-Hadhrami, S., Al-Fassam, N., Benhidour, H. 2019. "Sentiment Analysis Of English Tweets: A Comparative Study Of Supervised And Unsupervised Approaches", . In 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyad, Suudi Arabaistan, 1-3 Mayıs.

Alwehaibi, A., Bikdash, M., Albogmi, M., Roy, K., 2022. A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches. Journal of King Saud University - Computer and Information Sciences 34 (8, Part B), 6140–6149. https://doi.org/10.1016/j.jksuci.2021.07.011.

Aly, M.A., Atiya, A.F., 2013. Labr: a large scale arabic book reviews dataset. ACL, 494–498.

Arbane, M., Benlamri, R., Brik, Y., Alahmar, A.D., 2023. Social media-based COVID-19 sentiment classification model using Bi-LSTM. Expert Systems with Applications 212,. https://doi.org/10.1016/j.eswa.2022.118710 118710.

Basiri, M.E., Nemati, S., Abdar, M., Cambria, E., Acharya, U.R., 2021. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Gener Comput. Syst. 115, 279–294. https://doi.org/10.1016/j.future.2020.08.005.

Baziotis, C., Pelekis, N., Doulkeridis, C. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis Proceedings of the 11th International Workshop on Semantic Evaluations, SemEval-2017, pp. 747-754, 10.1109/iembs.1997.757075.

Brahimi, B., Touahria, M., Tari, A., 2021. Improving sentiment analysis in Arabic: A combined approach. Journal of King Saud University - Computer and Information Sciences 33 (10), 1242–1250. https://doi.org/10.1016/j.jksuci.2019.07.011.

Celik ve, O., Aslan, A. 2019. "Gender Prediction From Social Media Comments With Artificial Intelligence", Sakarya University Journal of Science, c.23, sayı.6, ss.1256-1264, 2019, DOI: 10.16984/saufenbilder.559452.

K. Chakraborty, S. Bhattacharyya, R. Bag, "A survey of sentiment analysis from social media data", In Proc.of IEEE Transactions on Computational Social Systems, 7 (2) (2020), pp. 450-464, 10.1109/TCSS.2019.2956957.

Chen, Z., Cao, Y., Lu, X., Mei, Q., Liu, X. 2019. SEntiMoji: An emoji-powered learning approach for sentiment analysis in software engineering, Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2019, Association for Computing Machinery, New York, NY, USA (2019), pp. 841-852, 10.1145/3338906.3338977.

Chen, H., Wu, L., Chen, J., Lu, W., Ding, J., Mar. 2022. A comparative study of automated legal text classification using random forests and deep learning. Information Processing & Management 59, (2). https://doi.org/10.1016/j.ipm.2021.102798 102798.

Demircan, M., Seller, A., Abut, F., Akay, M.F., Jun. 2021. Developing Turkish sentiment analysis models using machine learning and e-commerce data. International Journal of Cognitive Computing in Engineering 2, 202–207. https://doi.org/10.1016/j.ijcce.2021.11.003.

Devika, M., Sunitha, C., Ganesh, A., 2016. Sentiment analysis: A comparative study on different approaches. Procedia Comput. Sci. 87, 44–49. https://doi.org/10.1016/j.procs.2016.05.124.

Dhaoui, C., Webster, C.M., Tan, L.P., 2017. Social media sentiment analysis: Lexicon versus machine learning. Journal of Consumer Marketing 34 (6), 480–488. https://doi.org/10.1108/JCM-03-2017-2141.

Duwairi, R., El-Orfali, M., 2014. A study of the effects of preprocessing strategies on sentiment analysis for arabic text. J. Inf. Sci. 40 (4), 501–513.

GitHub Pages, https://github.com/BihterDass/ArabicTextClassificationDataset/turkish-nlp-dataset/EnglishTextClassificationDataset, 2022 [accessed 27 August 2022].

Heikal, M., Torki, M., El-Makky, N., 2018. Sentiment analysis of Arabic tweets using deep learning. Procedia Comput. Sci. 142, 114–122. https://doi.org/10.1016/j.procs.2018.10.466.

Hussein, D.-M.-E.-D.-M., 2018. **A survey on sentiment analysis challenges**, Journal of King Saud University -. Engineering Sciences 30 (4), 330–338.

İskender, E., Batı, G.B., 2015. Comparing Turkish Universities Entrepreneurship and Innovativeness Index's Rankings with Sentiment Analysis Results on Social Media. Procedia - Social and Behavioral Sciences 195, 1543–1552. https://doi.org/10.1016/j.sbspro.2015.06.457.

Jaya Hidayat, T.H., Ruldeviyani, Y., Aditama, A.R., Madya, G.R., Nugraha, A.W., Adisaputra, M.W., 2022. Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. Procedia Computer Science 197, 660–667. https://doi.org/10.1016/j.procs.2021.12.187.

Karayiğit, H., İnan Acı, Ç., Akdağlı, A., 2021. Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods. Expert Systems with Applications 174,. https://doi.org/10.1016/j.eswa.2021.114802 114802.

Kemaloğlu, N., Küçüksille, E.U., Özgünsür, M.E., 2021. **Turkish sentiment analysis on social media**, Sakarya University Journal of. Science 25 (3), 629–638. https://doi.org/10.16984/-saufenbilder.872227.

Kumar, G.R., Sheshanna, K.V., Anjan Babu, G., 2021. Sentiment Analysis For Airline Tweets Utilizing Machine Learning Techniques. In: international Conference on Mobile Computing and Sustainable informatics EAI/Springer innovations in Communication and Computing, Chapter 75. Springer Nature Switzerland AG, pp. 791–799.

Kumas, E., 2021. Comparison of Classifiers When Making Sentiment Analysis from Turkish Twitter Data. Eskisehir Turkish World Application and Research Center Informatics Journal c.2, no.2, 1–5.

Minaee, S., Azimi, E., Abdolrashidi, A. 2019. Deep-sentiment: Sentiment analysis using ensemble of CNN and Bi-LSTM models**.** http://arxiv.org/abs/1904.04206.

Naresh, A., 2021. Recommender system for sentiment analysis using machine learning models. Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (10), 583–588.

Ni, P., Li, G., Hung, P.C.K., Chang, V., Dec. 2021. StaResGRU-CNN with CMedLMs: A stacked residual GRU-CNN with pre-trained biomedical language models for predictive intelligence. Applied Soft Computing 113,. https://doi.org/10.1016/j.asoc.2021.107975 107975.

Obaidi, M., Nagel, L., Specht, A., Klünder, J., 2022. Sentiment analysis tools in software engineering: A systematic mapping study. Information and Software Technology 151,. https://doi.org/10.1016/j.infsof.2022.107018 107018.

Oflazer, K., Saraclar, M. 2018. Turkish Natural Language Processing, Springer Verlag-Theory and Aplications of Natural Language Processing Book Series.

Parlar, T., Sarac, E., Ozel, S. A. 2017. "Comparison of feature selection methods for Sentiment analysis on Trukish Twitter Data, 2017 25th Signal Processing and Communications Applications Conference (SIU), pp 1-4, 2017.

Pervan, N., Yalım Keleş, H. 2017. "Sentiment Analysis Using A Random Forest Classifier On Turkish Web Comments", Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering, c. 59, sayı. 2, ss. 69-79, 2017, DOI:10.1501/commua1-2_0000000105.

Petz, G. et al., 2014. computational approaches for mining user's opinions on the Web 2.0. Information Processing & Management 50, (6), 899–908. https://doi.org/10.1016/j.ipm.2014.07.005.

Rambocas, M., Pacheco, B.G., 2018. Online sentiment analysis in marketing research: A review. Journal of Research in Interactive Marketing 12 (2), 146–163. https://doi.org/10.1108/JRIM-05-2017-0030.

Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M., 2011. Oca: opinion corpus for arabic. J. Assoc. Inf. Sci. Technol. 62 (10), 2045–2054.

G. Sariman ve E. Mutaf, "Covid-19 Sürecinde Twitter Mesajlarinin Duygu Analizi", Euroasia Journal of Mathematics Engineering Natural and Medical Sciences, c.7, sayı. 10, ss. 137-148, 2020, DOI: 10.38065/euroasiaorg.149.

Singh, M., Wasim Bhatt, M., Bedi, H.S., Mishra, U., Dec. 2020. Performance of bernoulli's naive bayes classifier in the detection of fake news. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.10.896.

Sunitha, P.B., Joseph ve, S., Akhil, P.V. 2019. ''A study on the performance of supervised algorithms for classification in sentiment analysis'', In IEEE Region 10 Conference (TENCO), Kochi, Hindistan, 17-20 Ekim, 2019.

A.K. J., Trueman, T.E., Cambria, E. 2021. A convolutional stacked bidirectional LSTM with a multiplicative attention mechanism for aspect category and sentiment detection Cogn. Comput., 13, pp. 1423-1432, 10.1007/s12559-021-09948-0.

Vyas, V.U., 2019. Approaches to sentiment analysis on product reviews. sentiment analysis and knowledge discovery in contemporary business. IGI Global, 15–30.

Yan, G., He, W., Shen, J. and Tang, C. 2014. ''A bilingual approach for conducting Chinese and English social media sentiment analysis,'' *Computer Networks*, vol. 75, pp. 491–503, Dec. 2014, doi: 10.1016/j.comnet.2014.08.021.

Yue, L., Chen, W., Li, X., Zuo, W., Yin, M., 2019. A survey of sentiment analysis in social media. Knowledge and Information Systems 60 (2), 617–663.

Zhang, L., Wang, S., Liu, B., 2018. **Deep learning for sentiment analysis: A survey** Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery 8 (4), 1–25.