Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES
and
ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access
Volume 05
Issue 01

April, 2021

Research Article

Analysis of whether news on the Internet is real or fake by using deep learning methods and the TF-IDF algorithm

Tilbe Korkmaz [a], ***Ali Çetinkaya*** [b,c,]*, ***Hakan Aydın*** [a] **and** ***Mehmet Ali Barışkan*** [a]

[a]*Istanbul Gelisim University, Faculty of Engineering and Architecture, Computer Engineering, Istanbul, 34310, Turkey*
[b]*Istanbul Gelisim University, Technology Transfer Office Application and Research Center, Istanbul, 34310, Turkey*
[c]*Kocaeli University, Faculty of Engineering, Department of Electronics and Communications Engineering, Kocaeli, 41001, Turkey*

ARTICLE INFO

Article history:
Received 12 August 2020
Revised 08 October 2020
Accepted 21 October 2020

Keywords:
Natural language processing
Text analysis
Text classification
TF-IDF algorithms

ABSTRACT

Internet use has become increasingly widespread nowadays. In addition, there is a significant increase in the amount of text content produced in digital media. However, the accuracy and inaccuracy of the news we read and the content produced in a large number are also unknown. In this study, classification and analysis of whether the news is real or not were done by using Deep Learning methods. For the English news, the data set created by Katharine Jarmul was used. The data set contained a total of 6336 news items. The distribution of this data set, which consisted of political and political news, was 50% fake and 50% real. The method used in text classification was Term Frequency - Inverse Document Frequency (TF-IDF). The classification was made with the data set used and 93.88% success and 6.12% error were obtained as a result of the analysis.

1. Introduction

Especially with the increase of news on the internet, the importance of obtaining accurate news and clear information has also increased. The increase in the number of news has also led to an increase in false of fake news. In this article, text mining methods and deep learning techniques are used together to classify real and fake news. This study is important in terms of automatically and efficiently determining real and fake news, making news analysis faster and more efficient, and eliminating such problems in news analysis.

Deep learning is an artificial intelligence method used in areas such as object recognition, speech recognition, and natural language processing, and it uses multi-layer artificial neural networks. Today, many researchers in different fields such as big data [1-6], autonomous vehicles [7, 8], handwritten character recognition [9, 10], medical image processing, natural language processing [11, 12], signature verification, voice and video recognition, are using deep learning method in their studies conducted in the most popular and challenging areas of the world [13-16].

In the deep learning method, the learning process is performed through examples. There is no need to use rule sets to solve the problem. It is sufficient to select a model that provides a solution to the problem by evaluating the samples or to create it. Therefore, model selection should be done well [17-19]. Deep learning structures often enable the design, training, and accurate analysis of artificial neural networks using highly programmed interfaces. Some of the most used deep learning libraries are Caffe2, PyTorch, Tensorflow, etc. As a working style, they use multiple graphics processing units (GPU) for high performance; the biggest reason for graphic processors, such as CUDA and cuDNN, to use accelerated libraries is that they can be trained quickly [20].

Today, many studies are carried out on text and sentence classification problems. One of the most important problems of the text classification [21] is that the texts to be classified are not structural [22]. The success rate can be increased by

* Corresponding author. Tel.: +90 212 422 7000 \ 7187; Fax: 0212 422 74 01
E-mail addresses: tilbekorkmaz95@gmail.com *(T. Korkmaz)*, alcetinkaya@gelisim.edu.tr *(A. Çetinkaya)*, haaydin@gelisim.edu.tr *(H. Aydın)*, mabariskan@gelisim.edu.tr *(M. A. Barışkan)*
ORCID: 0000-0002-6882-5004 *(T. Korkmaz)*, 0000-0003-4535-3953 *(A. Çetinkaya)*, 0000-0002-0122-8512 *(H. Aydın)*, *0000-0002-8039-2686 (M. A. Barışkan)*
DOI: 10.35860/iarej.779019

using deep learning method in analysis studies conducted with classification algorithms [23] on text processing [24, 25]. As a success criterion, the quality of word representations comes to the fore.

Term Frequency-Inverse Document Frequency (TF-IDF) is used for word representation and Word2Vec [26] is used for fast text [23]. TF-IDF can be used in the so-called subject grouping together with the detection of the articles, in the inference of the author and the classification of the author and the subject [11, 27]. The text classification algorithm developed for author recognition has been successful as a result of experiments [28]. The classification success of texts has been seen as 81.2% [29]. It has been shown that with a 93.3% accuracy rate, much better performance achieved with classification algorithms based on machine learning [9, 10]. In a study on the problem of Turkish text classification [30], as a result of the analysis, it was observed that the stationary words were removed and there was an increase in the classification performance. However, it was not observed that taking word roots had a positive effect on classification accuracy. In other studies, it has been determined that the most successful is the term weights TF and TF * IDF [31].

The purpose is to perform automatic detections by using text mining methods in the interpretations on the internet. In this study, together with 444 comments related to this project, by using the TF-IDF classification algorithm, the success measurements and comparisons ensured in the model can be made. The most successful algorithm among the compared algorithms has been found to be consecutive minimal optimization (88.73%) [32]. It is observed that the number of academic studies published on the internet increases day by day [33]. The words in the news texts are created using the FastText model, which is used very popularly in the recent literature as word embedding. The model is tested individually first by a single sentence and then by the first two sentences through training the full text in each news [34].

A system of detection from comments was developed with analysis on Turkish texts such as social media posts, customer reviews, novels, etc. [35, 36]. The increase in internet and social media use has also increased the sharing of textual information [33]. Classification of the tweets written on the sector has been carried out [37, 38]. It has been investigated that whether there is any theme of emotion among Turkish and English tweets, and if any, a classification of emotions has been performed [39, 40]. The parts obtained from the Turkish texts and the TREMO data set used in the area of emotion extraction have been compared with the classification results obtained using different machine learning algorithms [10]. Classification and modeling are applied to the data set prepared via Facebook. Knowing that it is effective in social media mining, this method has been studied and shown as a text classification [31, 41]. Test data collected using web scraping methods as summaries of news on Turkish news websites are classified using vector learning methods and depth learning methods together with using a "hot coding". A 90% classification success is achieved [10].

To solve the classification problems of websites by using the artificial neural network [42, 43], two different approaches (dual classification or multiple classifications) are applied. Both approaches were tested on Web sites collected within the scope of a study and a comparison of their performance was observed [44]. ANN is one of the self-learning methods. It leads to the emergence of deep learning [12]. The Regional Based Convolutional Neural Network (RCNN), long and short-term memory (LSTM) [31, 39], and convolutional neural network (CNN) models can be used in sentence classification studies [36].

In the data set used in our deep learning study we conducted classification by using TF-IDF, and 93.88% success and 6.12% errors were detected. Compared to the algorithms used in the references, the success rate was 2.12 percent less than the DSA algorithm and 0.52 percent less than the Random Forest algorithm, but it showed a noticeable success compared to other algorithms.

2. Materials and Methods

2.1 Parts of the System's Control Structure

ANN (artificial neural network) is called logical software working similarly to the human brain and it has been developed to think, learn, generalize, remember, and produce new information like brain. ANN is called synthetic structures that mimic biological neural networks. In the artificial neural network model, there are 3 layers: the input layer, hidden layer, and output layer [45]. The cell structure of the Artificial Neural Network is shown in Figure 1 [46]. The X input, W weights, ∑ Transfer function, and net-input that will enter the activation function are shown in figure 1.

As stated in Equation 1, in order to reach the output in Figure 1, the threshold value is summed with the inputs.

$$\text{Result} = f = (\textstyle\sum_{i=1}^{n} w_i \, x_i + \vartheta) \qquad (1)$$

Where x_i refers to input values, w_i refers to weight values, and ϕ refers to threshold value.

The activation function provides curved equalization created by the input layers and the output layers.

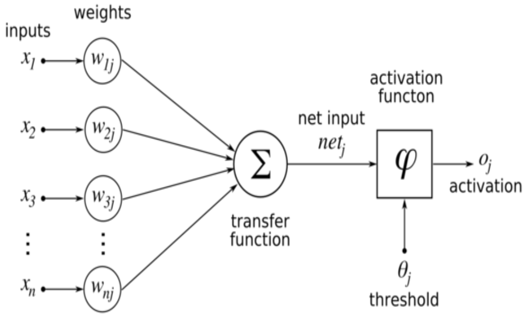


Figure 1. Artificial neural network cell structure

Caution should be exercised when choosing the activation function because this choice significantly affects the performance of ANN. The activation function is unipolar (0 1), but can be selected as bipolar (-1 +1) or linearly qualified. It is the segment that facilitates the introduction and learning of a nonlinear structure in ANN. Sigmoid and ReLU are activation functions.

The most important reason for using the sigmoid activation function is that it compresses the value between 0 and 1. Therefore, it is used in models that find the probability of an event occurring. Although the derivative of the sigmoid function exists, the Sigmoid function is stationary, but its derivative is not stationary. The sigmoid function can cause the model to pause during training. For this reason, it is not recommended for use in machine learning models, which are described as complex. The sigmoid function divides the data into two classes [46]. In Equation 2, the mathematical formula of the Sigmoid function is given. Its values are between 0 and 1.

$$\sigma(X) = \frac{1}{1+e^{-x}} \tag{2}$$

ReLU (Rectified Linear Unit) is a nonlinear function, also known as the transfer function. It is known that the ReLU function takes the value of 0 for the inputs given as negative, whereas it takes the value of x for the inputs given as positive. Today, ReLU is the most used activation function that is generally utilized in Deep Learning and Convolutional Neural Networks [7].

$$R(z) = max(0, z) \tag{3}$$

The mathematical formula of the ReLU activation function is given in Equation 3. If the value obtained as a result of the activation process is negative, it takes the value 1, if it is positive, it takes the value 0.

Deep learning is expressed as the prediction of data sets created by keeping the created and used data together within the framework of the desired results. It is also called the machine learning method because of the multiple layers created. Deep learning (machine learning) is summarized as the sub-branch of artificial intelligence [13, 47]. The word 'deep' in the Deep Learning method refers to having more than one hidden layer. A sample layer structure of the Deep Learning is shown in Figure 2 [48].

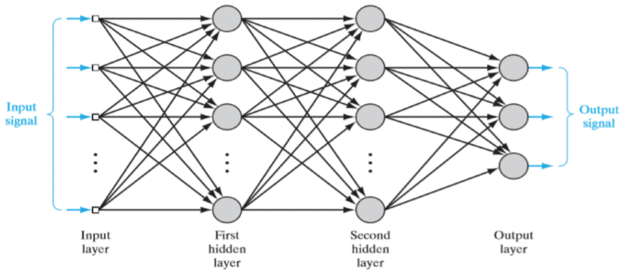


Figure 2. Deep learning sample layer structure

While performing deep learning, three different methods can be used: Supervised, Semi-Supervised, or Unsupervised. In Deep learning methods, together with a large amount of data entering, being able to distinguish between different features are learned. In the learning process, it is determined that the excess data given as input increases success. As the data is processed, it changes its position by passing to the next layer. It is known that as the upper layers are passed, the number of details extracted from the given one increases. It is also known that there are more than one deep model types. These are Multilayer Perceptron, Convolutional Neural Networks, and Recurrent Neural Networks.

Multilayer perceptrons (MLP) have been created as a result of the studies conducted to solve the “exclusive or (XOR)” problem. It has been observed that MLP is effective when classifying and generalizing. Since many inputs will not be enough for a single neuron, more than one neuron is needed for parallel processing. It takes the data from the input layer and transfer to the hidden layer. The intermediate layer can differ as being at least one layer. The exit of each layer becomes the entrance of the next layer. Each neuron is attached to the neurons in the next layer. The output layer determines the output of the network by processing the data received from the previous layer. The number of outputs is equal to the number of elements in the output layer. A training set consisting of sample inputs and outputs is essential for the network to learn.

The usage areas of Deep Learning are increasing day by day. Today, it is mostly used in areas such as face detection, sound detection, and vehicles with auto-pilot feature and called driverless vehicles. In alarm systems, it is thought that since examining the continuous camera recordings leads to loss of time, technologies such as alarm system warnings operating depending on unusual movements provide convenience. Today, this is possible with deep learning methods. Thanks to the deep learning methods used in the health sector, cancer research has gained momentum and it is observed that it eliminates the loss of time. It can be easily diagnosed whether the cells are cancerous or not by recognizing the samples marked as cancerous cells by deep learning algorithms. With this process, fast and successful results are obtained. As in other areas, deep learning methods can be also developed to improve the quality of images in cyber threat analysis.

2.2 Natural Language Processing

In the literature, languages are examined in two categories; these are machine language and the natural language we use in everyday speech. Natural language processing has been developed for computers to understand and interact with the language that people speak. It is also called the process by which computers qualify natural languages.

Directly imported objects cannot be used for the classification of texts. Data is obtained from these extracted objects and the process is carried out on these data. In the classification process, the separation of the texts into classes by pre-processing the texts ensures that successful results are obtained by facilitating the process to be done in the next step. Various categories are produced on the texts whose classification is desired, and various steps such as Lower-Case process, tokenization process, stop words extraction, stemming process, lemmatization process, and BoW process are applied.

The Lower-Case Process is used to eliminate differences by translating all the words in different sentences into lowercase letters in existing texts.

The Tokenize process stays on the whole of the texts in most data sets. Since it is not functional to operate with these texts, the data must be separated into parts and removed. This process is called token (shred) and Tokenizing is done as a string.

On the other side, words that we use a lot in our daily lives and that does not make sense when used alone are called “Stop Words”. The elimination of words such as “and”, “or”, and “at” in the texts in order not to deviate from the accuracy of the text models to be studied is expressed as the Stop Words process.

The Stemming Process is described as Stemming tightening in Turkish. It is one of the most important steps in data preprocessing. It is known as the process of obtaining a root by trying to cut the suffixes in the front and the end in existing words.

Table 1 shows the examination of the words stemming process. As shown in the table, it does not discriminate while descending to the roots of the words. For this reason, it can be successful in some cases. Lemmatization Process, called as the Lemmatization root analysis in Turkish, is a word unitization. It has similar features with the Stemming process. In both processes, the aim is to obtain root. The Lemmatization Process tries to extract a morphological or semantic root from words. If this process is done, more accurate results will be obtained by discarding the excesses that are described as noise.

Table 2 shows steps of the Lemmatization process. The most important feature of the lemmatization process is that it is linguistic. To obtain an accurate lemma, the given words should be analyzed morphologically.

BoW (Bag of Word) is a process used in natural language processing to simplify models. The words in the texts are kept in a bag with the BoW model. While holding together, grammar and sequence errors are not taken into account. It is the most used method for extracting attributes in texts in data sets being used.

Table 1. Stemming process

Word	Suffixes	Stemming
Fly	-es	Fli
Flying	-ing	fly

Table 2. Lemmatization process

Word	Morphological Information	Lemmatization
Flies	3rd person, Singular, Present Simple Tense	fly
Flying	3rd person, noun-verb, Present Tense	fly

We know that Natural Language Processing is examined in two categories; machine and everyday speech languages. Natural language processing is used because computers cannot understand the language that people speak. It is known that the texts written on computers are digitized as 0 and 1. The large data sets used in Natural Language Processing and the texts in these data sets are digitized by various methods by computers. This digitization process is carried out with various algorithms and methods.

Loss function is defined as function that measures the error and success rate of the created model. The last layer of artificial neural networks is known as the layer where the loss function is defined. The Loss function transforms the error rate calculation problems that occur during training into an optimization problem. The Loss function first calculates the difference between the actual values of the estimation made by the training model. Since the prediction will not be good, the difference between the real values and the estimated values will be large if the necessary care is not taken in the model creation. Accordingly, the loss value will be high. In cases were the model is created well, the difference will decrease, and in cases where it is the same, the loss will be 0.

Optimization (Optimizer) Algorithms are methods used to minimize the difference between the output value and the actual value produced at the output layer in the artificial neural network. Based on the size of the data set used in model training, there are 3 different types of algorithms described as slope descent.

Metrics are measurements where data can be counted and numerical values are obtained. The values obtained as a result of these measurements are expressed as total or ratio. Metrics are used when showing all or different elements of a dimension.

Term Weighting is a process performed on the roots obtained by natural language processing. If the term exists in the document, a weighted value of this term is formed. However, if this value is not available, it is expressed as 0.

In terms of the Term Frequency (TF), the weighting process is calculated in the document where the data is found. Weighting is made according to the term in the document. Since it takes value as much as amount of it in the document, the more it passes, the more value it gets.

Inverse Document Frequency (IDF) weighting gives weight according to the number of documents in which a term is mentioned. In the total documents used, the low presence of the same term increases the distinguishing feature, namely the IDF value. Otherwise, if a term is detected more than once in many documents, there is a decrease in the IDF value as its distinctiveness decreases. The result to be reached here is a standardization obtained by multiplying the frequency of the terms used with the IDF value.

$$w_{i,d} = tf_{i,d} \times log(n/df_i) \quad (4)$$

Equation 4 represents the calculation for the term I and the document d. TF is calculated first. The ratio of the number of times the term is mentioned in the document to the most mentioned term is shown in equation 5 below.

$$tf_{i,d} = aza(fr_{i,d} \, / \, df_i) \quad (5)$$

$$\{0,1\} \quad (6)$$

Binary looks for the presence of terms in used documents. It is shown in Equation 6.

$$f_{t,d} \quad (7)$$

Raw Frequency refers to the number of repetitions of terms in documents and the total number of words in the documents. It is shown in Equation 7.

$$log\ (1 + f_{t,d}) \quad (8)$$

Log Normalization is taking its logarithm based on the value of f_ (t, d) calculated in Equation 7. It is shown in Equation 8.

$$0.5 + 0.5\frac{f_{t,d}}{max f_{t,d}} \quad (9)$$

As a result of weighting, Double Normalization generates a value between 0.5 and 1. It is shown in Equation 9. By dividing the value obtained by calculating Raw Frequency to the maximum number of terms in the document, the ratio of the other terms is calculated and standardization is performed in frequency.

$$K + (1 - K)\frac{f_{t,d}}{max f_{t,d}} \quad (10)$$

Double Normalization K is calculated as a smoothing term that can be viewed as scaling with the largest TF value. It is shown in Equation 10.

The Confusion Matrix is used to measure performance in classification algorithms performed on test data and where the accuracy of the actual values is known.

2.3. Flow Chart of The System

In this study, using Deep Learning methods, the classification and analysis of the news, which were real or fake, were done. The flow diagram of the system is shown in Figure 3. After the program is run, it connects to the data set we obtain as ready. The data set consists of real/fake news in English. After this news is classified as real and fake, the pre-processing of text begins. In the context of this study, various text preprocessing were carried out. Space vectors created with TF-IDF, the digitization process of texts was started. Before starting the model training, the training and test data in the data set were parsed. After the model training starts, real values and predictive values analysis were performed. As a result of this analysis, the realization steps of the application were completed by drawing the graph of the values formed.

3. Applications

Almost everyone is exposed to fake news, which is still present today. This type of news can be described as information pollution and they are misleading. In this study, to combat news that does not reflect reality, a model that could make a distinction between real and fake news by using deep learning methods and the text classification process was created. Analyses were made between the trained data and the test data, and these analyses were visualized through graphics.

In the study, Python and Python Libraries were used as the programming language because it is open source. The data (Katharine Jarmul) [49, 50] includes 6336 different news (political, political). The distribution of the data set is 50% fake and 50% real. The fact that the data set is distributed evenly affects the performance criterion positively.

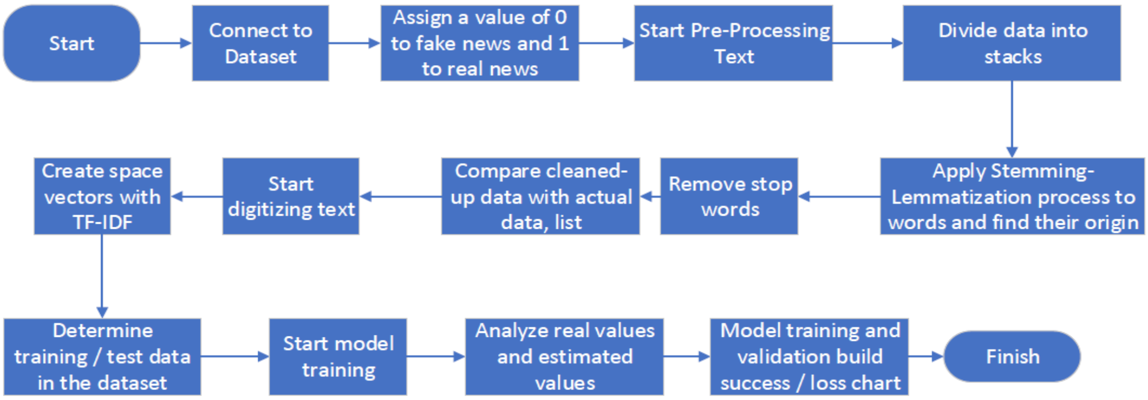

Start
Connect to Dataset
Assign a value of 0 to fake news and 1 to real news
Start Pre-Processing Text
Divide data into stacks
Apply Stemming-Lemmatization process to words and find their origin
Remove stop words
Compare cleaned-up data with actual data, list
Start digitizing text
Create space vectors with TF-IDF
Determine training / test data in the dataset
Start model training
Analyze real values and estimated values
Model training and validation build success / loss chart
Finish

Figure 3. Flow chart of the system

As can be seen in Figure 4 and 5 respectively, real and fake news were determined in the training data, and whereas the value of 1 was assigned to real news, the value of 0 was assigned to fake news.

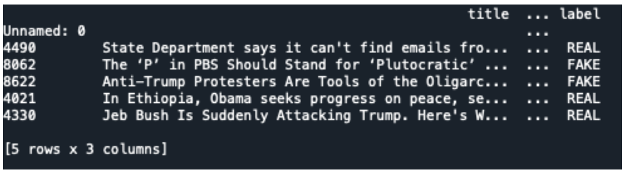
After completing the definition of training data in the data set, the texts were cleaned so that the texts could be classified correctly. The first step in text preprocessing is the fragmentation of data that stacks together. In Figure 6, the raw version of the news headlines is shown without pre-processing.

The visually cleared data are shown in Figure 7. To get more successful results in model training, the parsed texts were involved in the cleaning process. These processes were carried out to convert all letters to lowercase letters, to remove special characters (!, ^.%), to remove numbers, to divide texts into parts (tokenize), to delete Stop Words (ineffective words), and finally to find the morphological root of the words.

After the text preprocessing was completed, the Vector Space Model, which was one of the methods used in digitizing the texts, was used to bring the texts into digital form. Although each word represents a vector, when they are put together, the MxN-sized term-document matrix is formed. In this matrix, M is the number of news and N is the number of terms. The reason for the use of TF-IDF (Term Frequency - Inverse Document Frequency) management is to calculate the term weights, the frequency of the terms used with TF (Term Frequency), and the amount of all news in the data set checked with IDF (Inverse Document Frequency).

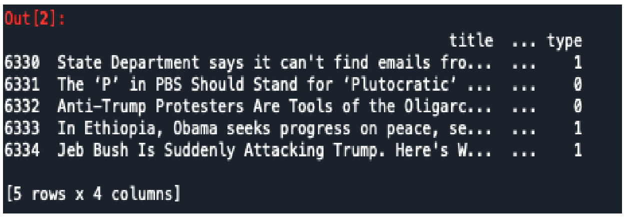
The text digitization process was completed after this stage. Before starting modeling, it is necessary to make the necessary definitions to build the network. To classify news texts, 3 layers were created in the model. Although the first and second layers were hidden layers, the number of nodes in each layer was 32. The third layer was called the output layer. The number of nodes in this layer was 1. ReLU (Rectified Linear Unit) was used as the activation function in the layers called the first and second hidden layers.

In model training, a layer called Dropout was added to prevent the model from memorizing the data. This layer sets the specified input data to zero.



```
                                                    title ... label
Unnamed: 0                                                ...
4490        State Department says it can't find emails fro...  ...  REAL
8062        The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...  ...  FAKE
8622        Anti–Trump Protesters Are Tools of the Oligarc...  ...  FAKE
4021        In Ethiopia, Obama seeks progress on peace, se...  ...  REAL
4330        Jeb Bush Is Suddenly Attacking Trump. Here's W...  ...  REAL

[5 rows x 3 columns]
```

Figure 4. Identifying real and fake news



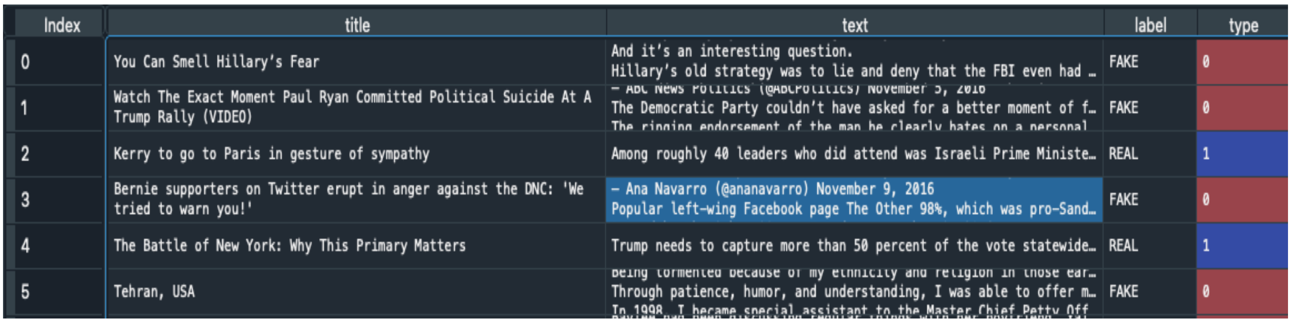
```
Out[2]:
                                                title ... type
6330  State Department says it can't find emails fro...  ...    1
6331  The ‘P’ in PBS Should Stand for ‘Plutocratic’ ...  ...    0
6332  Anti–Trump Protesters Are Tools of the Oligarc...  ...    0
6333  In Ethiopia, Obama seeks progress on peace, se...  ...    1
6334  Jeb Bush Is Suddenly Attacking Trump. Here's W...  ...    1

[5 rows x 4 columns]
```

Figure 5. Value assignment to real and fake news

The purpose of this synchronization is to prevent the model from overfitting the data. The dropout value was chosen between 0 and 1. In the third layer, the Sigmoid function was used as an activation function. These functions were used to describe nonlinear relationships in the model and for binary classification problems. If these functions were not used, there would be a decrease in learning achievement in the model.

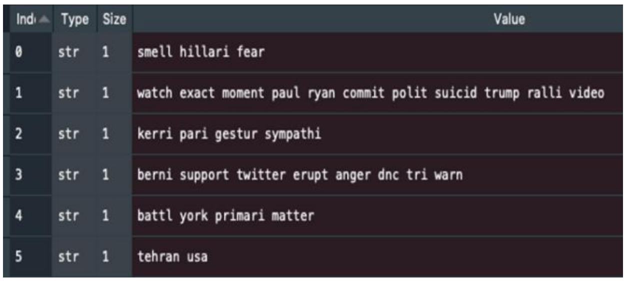
The distinction between fake / real news used in the model is called the binary classification problem. Many loss functions are used in binary classification problems. In this study, binary_crossentopy was used as the loss function. To minimize the value between real and estimated values, the rms probe algorithm was used as an optimizer algorithm, and accuracy values were monitored and measured.

While model training is performed, it is not right to put all the data into training for the training to be more successful. It needs to be broken down into specific pieces and involved in training. For this reason, data are divided into pieces called chunks. The number of times that all data will pass through Artificial Neural Networks is determined by Epoch (cycle). In this study, the data obtained from the model was defined in 20 Epoch and the model training was completed with 128 stacks for the data to be passed. The Scikit-Learn library was used to visualize the analysis between the values created as a result of the training and the actual values.



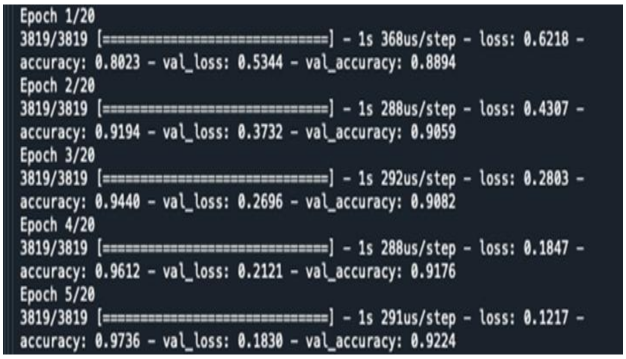
Index	title	text	label	type
0	You Can Smell Hillary’s Fear	And it’s an interesting question. Hillary’s old strategy was to lie and deny that the FBI even had …	FAKE	0
1	Watch The Exact Moment Paul Ryan Committed Political Suicide At A Trump Rally (VIDEO)	– ABC News Politics (@ABCPolitics) November 5, 2016 The Democratic Party couldn’t have asked for a better moment of f… The ringing endorsement of the man he clearly hates on a personal	FAKE	0
2	Kerry to go to Paris in gesture of sympathy	Among roughly 40 leaders who did attend was Israeli Prime Ministe…	REAL	1
3	Bernie supporters on Twitter erupt in anger against the DNC: 'We tried to warn you!'	– Ana Navarro (@ananavarro) November 9, 2016 Popular left-wing Facebook page The Other 98%, which was pro-Sand…	FAKE	0
4	The Battle of New York: Why This Primary Matters	Trump needs to capture more than 50 percent of the vote statewide…	REAL	1
5	Tehran, USA	Being tormented because of my ethnicity and religion in those ear… Through patience, humor, and understanding, I was able to offer m… In 1998, I became special assistant to the Master Chief Petty Off	FAKE	0

Figure 6. Display of the headlines before text preprocessing



Index	Type	Size	Value
0	str	1	smell hillari fear
1	str	1	watch exact moment paul ryan commit polit suicid trump ralli video
2	str	1	kerri pari gestur sympathi
3	str	1	berni support twitter erupt anger dnc tri warn
4	str	1	battl york primari matter
5	str	1	tehran usa

Figure 7. Data cleared as a result of text preprocessing



```
Epoch 1/20
3819/3819 [==============================] - 1s 368us/step - loss: 0.6218 -
accuracy: 0.8023 - val_loss: 0.5344 - val_accuracy: 0.8894
Epoch 2/20
3819/3819 [==============================] - 1s 288us/step - loss: 0.4307 -
accuracy: 0.9194 - val_loss: 0.3732 - val_accuracy: 0.9059
Epoch 3/20
3819/3819 [==============================] - 1s 292us/step - loss: 0.2803 -
accuracy: 0.9440 - val_loss: 0.2696 - val_accuracy: 0.9082
Epoch 4/20
3819/3819 [==============================] - 1s 288us/step - loss: 0.1847 -
accuracy: 0.9612 - val_loss: 0.2121 - val_accuracy: 0.9176
Epoch 5/20
3819/3819 [==============================] - 1s 291us/step - loss: 0.1217 -
accuracy: 0.9736 - val_loss: 0.1830 - val_accuracy: 0.9224
```

Figure 8. The result of the study

4. Results and Discussion

The result of the study is shown in Figure 8. The news in the dataset used was determined to be fake and real. Before the training, the value of 0 (zero) was assigned to fake news and the value of 1 (one) was assigned to real news. The above figure shows the first 5 stages of 20 epoch training. It is seen that the loss values are decreasing here. It is observed that as the training increases, the loss value decreases. On the other hand, the increase in value accuracy rates as a result of the training can be seen, and the software can distinguish the news as fake or real. As the training increases, the val_accuracy value increases, that is, the proximity to 1 increases, and the distinguishing of true news from false news can be ensured.

Deep learning contributes to the development of artificial intelligence through the use of machine learning with many applications. The technology, which develops day by day, allows new technological developments to occur to prevent the increase of false information in the digital environment. Natural language processing is a field developed to transform the information, which is increasing day by day in the digital environment, into a language that can be understood by machines. The increase in data day by day makes the processing of these data difficult. Natural language processing categorizes the texts and classifies them according to their characteristics, which allows us to quickly respond to the desired data.

In this study, using deep learning methods, news described as real or fake were classified by the TF-IDF (Term Frequency - Inverse Document Frequency) method. Keras library was used for model training and Pandas library was used to process data. Training and test data were determined just before starting the model training. A 3-layer model was created for model training. The number of training for the model was determined as 20.

As a result of the study, as shown in Figure 9, while the training loss of the model decreases in every step of the training, the success of the trained model increases. This increase is provided by the success of the model in training. Accordingly, when it comes to the 20th Epoch, while the decrease in the training loss of the model approaches to zero, the inversely proportional training loss increases.

In Figure 10, when we set the number of steps for our model training as 200, it is seen that the training loss decreases to zero, and accordingly, the loss of validation increases as the model is trained.

As a result of the statistical information we obtained in model training, the training of the model was calculated as 92.00% and the margin of error was 8.00%.

As seen in Figure 11, the number of epoch used for our model training was chosen as 10000. As a result of the statistical information we obtained in model training, the training of the model was calculated as 93.88% and the margin of error was 6.12%. A Confusion matrix was obtained by using the Scikit-Learn library to evaluate the model used.

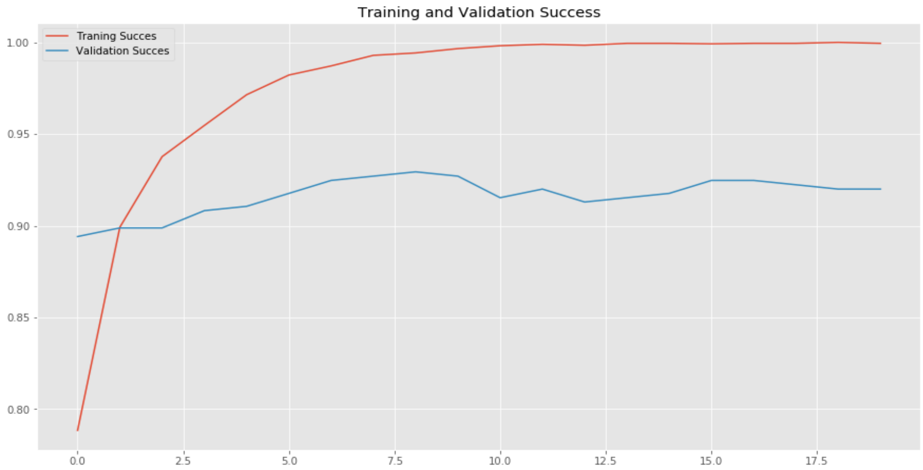


Figure 9. Model training success and loss in 20 Epoc

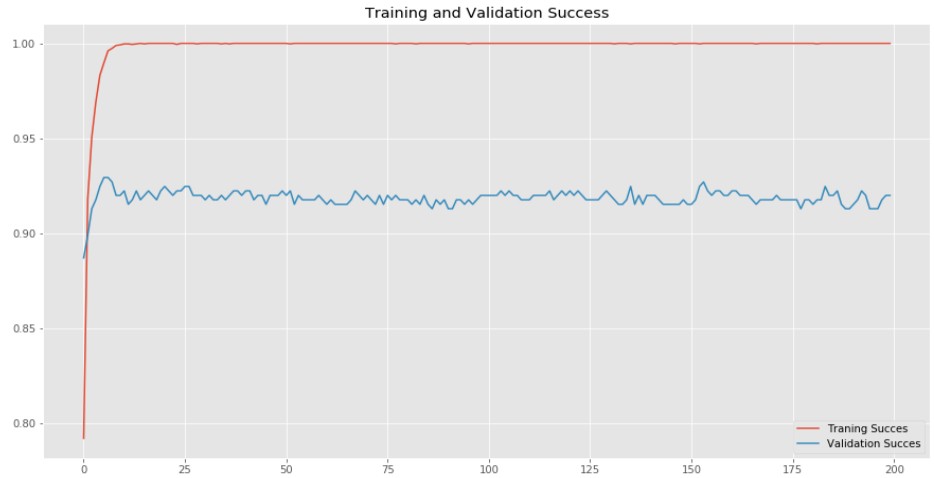


Figure 10. Model training success and loss in 200 Epoc

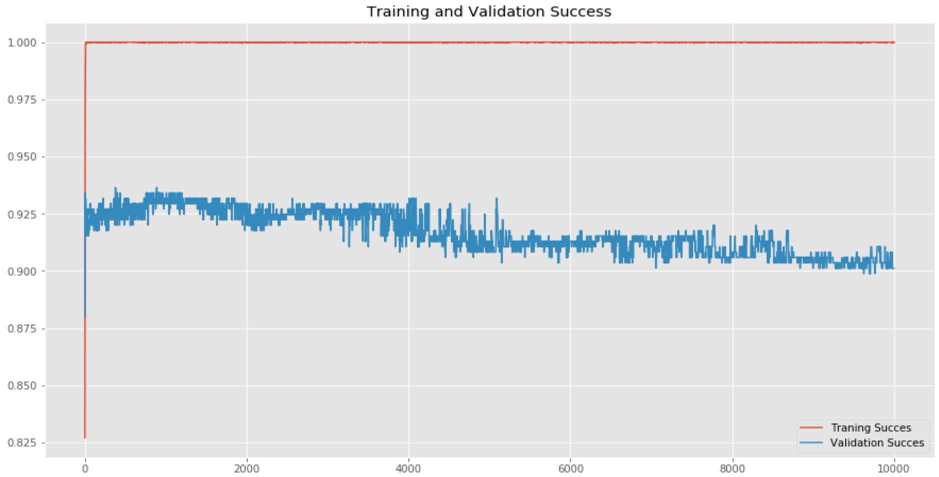


Figure 11. Model training success and loss in 10000 Epoc

The Confusion Matrix is given in Figure 12. It facilitated the observation of the result of the real values and performed on test data. It is seen that 79 of the 929 fake news selected as test data are real, and 85 of the 998 real news are determined as fake. Table 3 contains the success and error rates of the experiments conducted during the software development process. 8 experiments were carried out during the study process.

Experiments were carried out on the same data sets. Training and test data were not changed and necessary studies were carried out on the same model. The difference between experiments is that the epoch values during training were different. It is seen that as the Epoch value increases, the rate of success from training increases. While the success rate of 20 epochs in Experiment 1 was 91.06%, the error rate was 8.94 %. In Experiment 2, while the success rate of 200 epoch was 92.00 %, the error rate was 8.00 %. In Experiment 3, 500 epochs result in 92.89 % success rate and 7.11 % error rate. It was observed that as the number of epochs increased, the success rate increased and the error rate decreased.

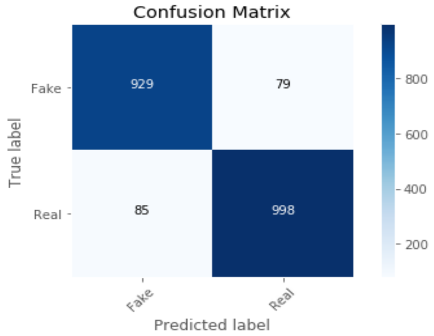


Figure 12. Confusion matrix

Table 3. Experiments and their results

Experiment No.	Epoch Value	Success Rate	Error Rate	Times (Sec)
1	20	91.06%	8.94%	24
2	200	92.00%	8.00%	241
3	500	92.89%	7.11%	595
4	1000	93.00%	7%	1216
5	2000	93.15%	6.85%	2532
6	3000	93.46%	6.54%	3795
7	5000	93.65%	6.35%	6290
8	10.000	93.88%	6.12%	12730

Table 4. Evaluation of the studies in the literature

Study No.	Technique Used	Algorithm Used	Success Rate	Fail Rate
Bilgin et al., 2019	Deep Learning	DSA algorithm	96.00	4.00
Bilgin, 2019	Deep Learning	Random Forest Algorithm	94.40	5.60
This Study	**Deep Learning**	**Tf-Idf Algorithm**	**93.88**	**6.12**
Süzen, 2019	Deep Learning	Convolutional Neural Networks (CNN)	93.70	6.30
Yücel et al., 2018	Machine Learning	Sequential Minimal Optimization Algorithm	88.00	12.00
Arı et al., 2017	Deep Learning	Convolutional Neural Networks (CNN)	87.30	12.70
Özmen et al., 2019	Deep Learning	it is not specified.	81.20	18.80

The data used in Table 4 were obtained as a result of the literature review and they were brought together to compare the algorithm and methods used. Based on these data, it was concluded that the study with the highest success was realized with the deep learning technique and the DSA algorithm [4]. It is observed that the lowest success percentage was obtained in the study performed with deep learning techniques [28]. Compared to other algorithms, the success rate of the method applied within the scope of the study is over 90%. The study has a low success rate of 2.12 percent according to the DSA algorithm. There is a 0.52 percent success rate reduction compared to the random Forrest algorithm. A noticeable increase in success has been achieved compared to the other algorithms in Table 4.

The abundance of information available today increases day by day with the advancement of technology. It is also easier for people to access information thanks to the internet. More research is carried out to ensure the accuracy of the information obtained. Therefore, the importance of deep learning, which is one of the sub-branches of artificial intelligence, is increasing day by day. Natural language processing used in deep learning techniques enables the processing of existing texts. Accordingly, classification of texts in different categories according to their subjects and processes becomes easier. We carried out our study by taking into account the access to the real news and the speed of reaching it accordingly.

Before starting our studies, various literature studies were reviewed and 6 different studies were selected by referring to 50 different studies. By making various comparisons over the data sets used, the common features of these selected studies were analyzed taking into account the differences in the methods used and the success achieved.

Based on the study we carried out, the data set, working model, and methods used in word analysis will change depending on the technological developments and more successful results can be obtained.

The fact that the information that needs to be confirmed is increasing day by day leads to an increase in the number of these types of studies. By adding a Turkish data set to the study we have carried out, an interface can be created and a website that is constantly updated and renews itself can be created using the necessary web services.

5. Conclusions

In this study, whether the news found on the internet is real or fake was analyzed using deep learning methods and the TF-IDF algorithm. In this way, by identifying and detecting real news with high success rates, it can be ensured that people can distinguish between real and fake news; in addition, by filtering fake news before news is added to search engines, only real news can be shown to the user. Considering this, a system was developed that can easily and quickly reveal the accuracy of news without having to be read by the user. With this system, real news can be shown to the user at a rate of 93.88%. Using the developed method, social media analysis, comment analysis, and twitter analysis can be carried out on web pages and accuracy analysis can be performed in different areas. In addition, by analyzing comments made on applications used in mobile platforms, this method can also be improved in a way to detect fake comments.

Declaration

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors also declare that this article is original, has been prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

T. Korkmaz, A. Çetinkaya and H. Aydın have planned and designed the research. M. A. Barışkan contributed to the process of working stages. T. Korkmaz, A. Çetinkaya and H. Aydın have carried out data collection and analysis. All the authors discussed the results and contributed to the final form of the article.

References

1. Arroyo-Fernández, I., Méndez-Cruz, C. F., Sierra, G., Torres-Moreno, J. M, and Sidorov, G., *Unsupervised sentence representations as word information series:*

Revisiting TF–IDF. Computer Speech & Language, 2019. **56**: p. 107-129.

2. Kim, D., Seo, D., Cho, S., and Kang, P., *Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec*. Information Sciences, 2019. **477**: p. 15-29.
3. Koyun, A., and Afşin, E., *Derin öğrenme ile iki boyutlu optik karakter tanıma.* Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2017. **10**(1): p. 11-14.
4. Yapıcı, M. M., Tekerek, A., and Topaloğlu, N., *Literature review of deep learning research areas.* Gazi Mühendislik Bilimleri Dergisi (GMBD), 2019. **5**(3): p. 188-215.
5. Patil, T., Pandey, S., and Visrani, K., *A review on basic deep learning technologies and applications*. In Data Science and Intelligent Applications, 2020. pp. 565-573. Springer.
6. Calisan, M., and Sakar, C. O., *Classification of short-texts by utilizing an external knowledge source*. Journal of Science and Engineering, 2017. **19**(57).
7. Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G., *A survey of deep learning techniques for autonomous driving*. Journal of Field Robotics, 2020. **37**(3), 362-386.
8. Küçük, D., and Arıcı, N., *Doğal dil işlemede derin öğrenme uygulamaları üzerine bir literatür çalışması.* Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi, 2018. **2**(2): p. 76-86.
9. Acı, Ç., and Çırak, A., *Türkçe haber metinlerinin konvolüsyonel sinir ağları ve word2vec kullanılarak sınıflandırılması.* Bilişim Teknolojileri Dergisi, 2019. **12**(3): p. 219-228.
10. Ertam, F., *Sosyal medya verileri için etkili bir sınıflandırma yaklaşımı*. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2017. **29**(2): p. 67-73.
11. Sjarif, N. N. A., Azmi, N. F. M., Chuprat, S., Sarkan, H. M., Yahya, Y., and Sam, S. M., *SMS spam message detection using term frequency-inverse document frequency and random forest algorithm.* Procedia Computer Science, 2019. **161**, p. 509-515.
12. MATLAB Deep Learning. [cited 2020 20 May]; Available from: https://www.mathworks.com/solutions/deep-learning.html
13. Şeker, A., Diri, B., and Balık, H. H., *Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme.* Gazi Mühendislik Bilimleri Dergisi, 2017. **3**(3): p. 47-64.
14. Hark, C., Uçkan, T., Seyyarer, E., and Karcı, A., *Metin özetlemesi için düğüm merkezliklerine dayalı denetimsiz bir yaklaşım.* Bitlis Eren Üniversitesi Fen Bilimleri Dergisi. 2019. **8**(3): p. 1109-1118.
15. Güldal, H., and Çakici, Y., *Ders yönetim sistemi yazılımı kullanıcı etkileşimlerinin sınıflandırma algoritmaları ile analizi*. Journal Of Graduate School Of Social Sciences, 2017. **21**(4).
16. Salur, M. U., Aydin, İ., and Karaköse, M. *gömülü derin öğrenme ile tehdit içeren nesnelerin gerçek zamanda tespiti*. Dümf Mühendislik Dergisi, 2019. **10**(2): p. 497-509.
17. Machine learning platform. [cited 2020 7 May]; Available from: https://www.tensorflow.org
18. Bayram, F., *Derin öğrenme tabanlı otomatik plaka tanıma.* Politeknik Dergisi, 2020. **23**(4): p. 955-960,
19. Sakarya, Ş., and Yılmaz, Ü. *Derin öğrenme mimarisi kullanarak bist30 indeksinin tahmini*. European Journal Of Educational And Social Sciences, 2019. **4**(2): p. 106-121.
20. PyTorch Deep learning library. [cited 2020 20 May]; Available from: https://pytorch.org
21. Yıldırım, S., and Yıldız, T., *Türkçe için karşılaştırmalı metin sınıflandırma analizi.* Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 2017. **24**(5): p. 879-886.
22. Akdoğan, Ö., and Özel, S. A., *Effects of feature extraction techniques on classification of turkish texts*. Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 2019. **34**(3): p. 95-108.
23. Tezgider, M., Yıldız, B., and Aydın, G., *Improving word representation by tuning word2vec parameters with deep learning model.* In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018. pp. 1-7. IEEE.
24. Çınar, A., *Veri madenciliğinde siniflandirma algoritmalarinin performans değerlendirmesi ve r dili ile bir uygulama*. Öneri Dergisi, 2019. **14**(51): p. 90-111.
25. Özmen, E. P., and Özcan, T., *Dolandiricilik tespiti üzerine melez siniflandirma ve regresyon ağaci uygulamasi*. Yönetim Bilişim Sistemleri Dergisi, 2019. **5**(2): p. 12-20.
26. Dündar, E. B., and Alpaydın, E., *Learning word representations with deep neural networks for turkish*. In 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019. pp. 1-4. IEEE.
27. Yücel, A., and Köylü, M. K., *Spam içerikli e-postalarin tespiti için bir metin madenciliği uygulamasi: terimlerin gama ilişki katsayisina dayali polarizasyonu.* Uluslararası Yönetim Bilişim Sistemleri Ve Bilgisayar Bilimleri Dergisi, 2018. **2**(2): p. 95-104.
28. Bilgin, M., *Türkçe metinlerin siniflandirma başarisini artirmak için yeni bir yöntem önerisi.* Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 2019. **24**(1): p. 125-136.
29. Demir, Ö., Baban Chawai, A., and Doğan, B., *Türkçe metinlerde sözlük tabanli yaklaşimla duygu analizi ve görselleştirme*. PORTA. 2019; **1**(2): p. 58-66.
30. Ertam, F., *Deep learning based text classification with Web Scraping methods*. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018. pp. 1-4. IEEE.
31. Thakkar, A., and Chaudhari, K., *Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks*. Applied Soft Computing, 2020. 106684.
32. Göker, H., and Tekedere, H., *Fatih projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesi.* Bilişim Teknolojileri Dergisi, 2017. **10**(3): p. 291-299.
33. Safali, Y., Nergiz, G., Avaroğlu, E., and Doğan, E., *Deep learning based classification using academic studies in doc2vec model.* In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019. pp. 1-5. IEEE.
34. Karakoç, E., and Yılmaz, B., *Deep learning based abstractive turkish news summarization*. In 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019. pp. 1-4. IEEE.
35. Doğan, O., *Türkiye'de veri madenciliği konusunda yapılan lisansüstü tezler üzerine bir araştırma*. Gazi Universitesi İktisadi ve Idari Bilimler Fakültesi Dergisi, 2017. **19**(3): p. 929-951
36. Pirana, G., Sertbaş, A., and Ensari, T., *Sentence classification with deep learning method for virtual assistant applications.* In 2019 3rd International Symposium On Multidisciplinary Studies And Innovative Technologies (Ismsit), 2019. pp. 1-5. IEEE.
37. Ayata, D., Saraçlar, M., and Özgür, A., *Turkish tweet sentiment analysis with word embedding and machine*

learning. In 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017. pp. 1-4. IEEE.

38. Nasim, Z., and Haider, S., *Cluster analysis of urdu tweets.* Journal of King Saud University-Computer and Information Sciences, 2020.
39. Bilgin, M., and Şentürk, İ. F., *Danışmanlı ve yarı danışmanlı öğrenme kullanarak doküman vektörleri tabanlı tweetlerin duygu analizi.* Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 2019. **21**(2): p. 822-839.
40. Alpkoçak, A., Tocoglu, M. A., Çelikten, A., and Aygün, İ., *Türkçe metinlerde duygu analizi için farklı makine öğrenmesi yöntemlerinin karşılaştırılması.* Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen Ve Mühendislik Dergisi, 2019. **21**(63): p. 719-725.
41. Süzen, A. A., *LSTM derin sinir ağları ile üniversite giriş sınavındaki matematik soru sayılarının konulara göre tahmini.* Engineering Sciences, 2019. **14**(3): p. 112-118.
42. Altan, G., *DeepGraphNet: grafiklerin sınıflandırılmasında derin öğrenme modelleri.* Avrupa Bilim ve Teknoloji Dergisi, 2019. (Özel Sayı): p. 319-327.
43. Doğan, F., and Türkoğlu, İ., *Derin öğrenme modelleri ve uygulama alanlarına ilişkin bir derleme.* Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 2019. **10**(2): p. 409-445.
44. Chouseinoglou, O., and Şahin, İ., *Metin madenciliği, makine ve derin öğrenme algoritmaları ile web sayfalarının sınıflandırılması.* Yönetim Bilişim Sistemleri Dergisi, 2019. **5**(2): p. 29-43.
45. Partal, T., Kahya, E., and Ciğizoğlu, K., *Yağış verilerinin yapay sinir ağları ve dalgacık dönüşümü yöntemleri ile tahmini.* İtüdergisi/D, 2008. **7**(3): p. 73-85.
46. Arı, A., and Berberler, M. E., *Yapay sinir ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı.* Acta Infologica, 2017. **1**(2): p. 55-73.
47. Özcan, B., Kumru, P. Y., and Fığlalı, A., *Forecasting operation times by using artificial intelligence.* International Advanced Researches and Engineering Journal, 2018. **2**(2): p. 109-116.
48. Al Qasem, O., Akour, M., and Alenezi, M. *The influence of deep learning algorithms factors in software fault prediction.* IEEE Access, 2020. 8, 63945-63960.
49. Fake News Dataset. [cited 2020 15 May]; Available from: https://www.datacamp.com/community/tutorials/scikit-learn-fake-news
50. Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., and Afroz, S., *A benchmark study on machine learning methods for fake news detection.* 2019. arXiv preprint arXiv:1905.04749.