Sentiment Analysis with Term Weighting and Word Vectors

Metin Bilgin[1] and Haldun Köktaş[2]
[1]Department of Computer Engineering, Bursa Uludağ University, Turkey
[2]Department of Mechatronic Engineering, Bursa Technical University, Turkey

Abstract: *It is the sentiment analysis with which it is tried to predict the sentiment being told in the texts in an area where Natural Language Processing (NLP) studies are being frequently used in recent years. In this study sentiment extraction has been made from Turkish texts and performances of methods that are used in text representation have been compared. In the study being conducted, besides Bag of Words (BoW) method which is traditionally used for the representation of texts, Word2Vec, which is word vector algorithm being developed in recent years and Doc2Vec, being document vector algorithm, have been used. For the study 5 different Machine Learning (ML) algorithms have been used to classify the texts being represented in 5 different ways on 3000 pieces of labeled tweets belonging to a telecom company. As a conclusion it was seen that Word2Vec, being among text representation methods and Random Forest, being among ML algorithms were most successful and most applicable ones. It is important as it is the first study with which BoW and word vectors have been compared for sentiment analysis in Turkish texts.*

Keywords: *Word2vec, Doc2vec, sentiment analysis, machine learning, natural language processing.*

Received February 16, 2018; accepted July 22, 2018

1. Introduction

While artificial intelligence applications began to develop and find place in each area of technology, studies on natural language have also been effected from this. Natural Language Processing (NLP) is the name given to software-based studies which are conducted on the spoken languages by people. With that respect, it comes out from the combination of language science and computer sciences.

NLP studies being conducted on written texts focuses on information extraction from the text, differentiating sentences into elements, determining meaningful roles of words, revealing the feeling of sentences, content analysis, dependency analysis, and recognition of name of existence. The word natural in its name has been used to mark the first area which forms naturally and which is recalled when it is mentioned about language as not being related with languages which are produced by people.

Among social media messages the ones that are studied most are Tweets due to their popularity, varieties and easiness of having access. Besides the advantages it has got, it also has other aspects such as being restricted with 140 characters, having a unique jargon, having too many writing errors and other features making NLP methods become difficult to implement. Its being restricted with 140 characters reduces the data amount that can be obtained from the message and its having a unique jargon and too many writing errors makes morphological analysis become difficult to be made.

In this study we realized sentiment analysis study with different text representations named also as document classification. Sentiment analysis is the work to reveal the sentiment within a sentence and it has usage area in various fields. For example, a company owner who wish to learn whether customer remarks being too many in number are positive or not or a website owner who wishes to evaluate film remarks automatically, needs to use sentiment analysis studies. No matter what the area of usage is, determining whether a sentence is positive or negative in an automatic way bears significant importance which is present or potential.

This paper contains of six sections, the first being introduction. Section 2 presents related works. Section 3 presents dataset. Section 4 presents methods. Experiment and Results are presented in section 5. Finally, the conclusions of this paper are presented in section 6.

2. Related Works

In this section information has been given about the studies being conducted on sentiment analysis.

In the studies they conducted, Zhang *et al*. [24] have conducted positive and negative labeled sentiment analysis on 100.000 Chinese remarks made on a product being purchased from Amazon by using various classification and learning methods. In the study on which two separate experiments were made as having Lexicon and part of speech basis, combination of Word2Vec and SVMperf has shown success on

lexicon based one with a ratio of 89.95% and it has shown success with ratio of 90.30% on part of speech based one.

In order to investigate how price changes in a company's stocks effected the emotions in tweets sent about the product, Dickinson et al have conducted sentiment analysis by using n-gram and Word2Vec methods. As a result of the study while success was achieved with n-gram method with a ratio of 65.7%, a success was achieved Word2Vec method with a ratio of 75% [9].

Tang *et al*. [21] have conducted sentiment analysis on the tweets being sent. In this study it was tried to reveal how various word embedding methods changed the success of system as being different from the other ones.

Polpinij *et al*. [18] have realized sentiment analysis on hotel remarks in another study. In this study where Word2vec method was used, as learning algorithm SVM has been used.

Sahin [20] has realized Word2Vec and SVM based classification process on Turkish texts. In the study where a data set being composed of 22.729 Turkish documents was used, it was investigated how processing on root or added words effected the system performance.

Xue *et al*. [23] have conducted sentiment analysis on Chinese microblog site Sina Weibo by using Word2Vec. Different aspect of the study is that it proposes a model to create an sentiment dictionary by using Word2Vec. Later on by using this dictionary it is aimed to determine the emotional tendency. As a result of the study for positive documents success with a rate of 0.94 was achieved and for negative documents success with rate of 0.96, and for neutral documents success with rate of 0.85 was achieved.

Kaur and Gupta [12] have developed a hybrid system for sentiment analysis of Punjabi texts. This model has integrated lexicon, N-gram model and Support Vector Machine.

In order to investigate the impact of training methods on performance with regards to sentiment analysis, Cetin has realized experiments on 2 datasets with various dimensions. As a conclusion it was seen that training methods were more successful [7].

As being different than others, Bilgin and Şenturk [4] has conducted sentiment analysis on twitter with semi-supervised data in both English and in Turkish. In this study with which it was investigated how document numbers effected the outcome, Distributed Bag Of Words (DBOW) and Distributed Memory (DM), being two versions of the Doc2vec algorithm were used. As a conclusion, it was seen that DBOW demonstrated a better performance than DM.

In the studies being conducted as Turkish, mostly traditional method of BoW was used until now. In this study Word2Vec, being the newly developed word vector extraction process and Doc2Vec, document vector extraction method have been used. Our study being among the first publishing applying word vectors on Turkish sentiment analysis [5] has been extended to make comparison with the methods of Document Vectors (Doc2Vec) and BoW (tf-idf). As a conclusion it was seen that in Turkish texts for text representation, Word2Vec method had big superiority when compared with the traditional methods.

3. Dataset

Dataset being used in our study is composed of 3000 pieces of Twitters belonging to a private company in telecom sector. Tweets being used in the study have been manually labeled by a group of students composed of three people getting expert systems lesson in Yildiz University Computer Engineering Division with the aim to be used in text classification and sentiment analysis studies [7].

Each data cluster in our hands has been first labeled by each student. The road that was followed for labeling each tweet was as follows:

- If all three students have assigned the same label, it was deemed to be correct. Ratio of tweets being labeled in this way was 85%.
- If two students have labeled the same and one student has labeled it in a different way, labeling of two people has been deemed as correct. Ratio of tweets labeled in this way was 13%.
- If all three students have assigned different labels, labeling was realized by the academician who was responsible from the lesson. Ratio of tweets labeled in this way was 2%.

As a result of labeling process, in the data cluster we had, 756 were labeled as positive, 1287 were labeled as negative and 957 were labeled as neutral.

3.1. Pre-Processing

Before using language processing algorithms it is required to realize various preliminary process on the texts. The reason for this is to purify the texts from situations which could influence the outcome in a wrong direction. In this study for these processes KNIME program has been used [1].

In pre-processing operation, first of all tokenization process was realized with OpenNLP Whitespace Tokenizer. Afterwards by converting the documents into small letters, with the number filter numbers on the documents have been cleaned. The following process was a bodily process. In this way, words having the same root but looking like different words due to different word endings have been determined and it was enabled for the size of vectors to be reduced. The process which was realized as final, has erased stop words and punctuation marks for Turkish and it

has enabled for unnecessary words not providing any benefits with regards to meaning to be cleaned.

4. Methods

4.1. Representing Text with Bag of Words

One of the most widely used text feature extraction methods in our time is BoW.

Opening out of tf-idf method which is a method of BoW is "term frequency-inverse document frequency". While term frequency denotes how many times a word has passed in a document (in one piece of text input), inverse document frequency, calculates number of documents where a text has passed and number of total documents in data.

In this method all the words in the data are transformed into a list. Afterwards all the words on this list are assigned to each document as a vector. For example if there is a dataset of 10.000 words, each document has a vector of 1 row and 10.000 columns. Each column in this vector defines a word. If the word passes in the document, number of times the word passes in the text is assigned to the column of that word. When this process is done for all documents, illustration of data as per text frequency is obtained.

Inverse Document Frequency (IDF) is calculated by using number of documents where a term has passed and total number of documents in the data with equality (1).

$$\log(n \div N) \tag{1}$$

n=Number of documents that the word has passed
N=Total number of documents

Finally by multiplying the values of *tf* and *idf, tf-idf* illustration of document is obtained. As a conclusion equality of *tf-idf* (2) is calculated as being shown [8].

$$tf - idf = tf * \log(n \div N) \tag{2}$$

4.2. Representing Text with Word Vectors

In this section two different word vector extraction method has been used.

4.2.1. Word2Vec

Word2Vec, is a word vector finding algorithm which is developed by Mikolov *et al*. [16] In fact this method is composed of two pieces of algorithms which are named as Continuous BoW (CBOW) and Skip-Gram.

Word2Vec, is a deep learning method where 3 pieces of layers are used. For CBOW algorithm in the first layer there are word inputs and from these inputs it is worked on to predict the context. Skip-Gram is completely the opposite of this. Meaning that while input is context, algorithm gives the words as output [16].

Afterword vectors are extracted, in order to obtain document vector another process is also needed. In this step by obtaining an arithmetic average of each column, document vectors are formed.

4.2.2. Doc2Vec

As being different from Word2Vec, in Doc2Vec method, it is obligatory for the documents to be labeled. Because while Word2Vec works on the relations between words and other words, Doc2Vec matches the words with these labels. Apart from this as a method, it is very similar to Word2Vec because what is being done here is to assign an ID to the documents in addition during the training stage. For this reason Doc2Vec is divided into two algorithms just like Word2Vec. These are DBOW and DM algorithms. At DM, in addition to the words in CBOW, paragraph ID is also given at the training. In DBOW as being similar to Skip-Gram, it is worked on to predict words from paragraph ID (from its label) [14].

Vector extraction processes have been realized in KNIME environment and the parameters being used have been shown in Table 1. Same parameters have been used for both Word2Vec and Doc2Vec.

Table 1. Word2vec and Doc2vec parameters.

Parameter	Value
size	100
Window size	5
Minimum Word frequency	5
Negativesampling rate	5

4.3. Learning Algorithms

Sentiment analysis is the process of forming a model by the system when data are provided as being labeled like positive and negative and assigning a label to the data when it does not have labels. For this purpose, ML, lexicon or their mixture is used [17]. In our study ML approach has been used.

ML is the name given to the work of finding a pattern from out of the existing data and to teach these patterns to the computer software. The way of realizing this process is determined by the features of data being owned. If data is composed of input and target value, this is called advisory learning and it is required for convenient algorithms to be used for this. If target values are not determined, meaning that there is no information as to which particulars the data in hand corresponds to, then non-advisory learning methods should be implemented [2]. Mainly there are 3 pieces of ML methods. These are regression, classification and clustering. Target values of regression process are specific and these values are real (continuous). In classification even if target values are specific, as it can be understood from the name of the method, these are discrete values. In the clustering process target values are not specific. Software groups the data according to their common features, meaning that it tries to form its own classes [10].

In this study during the learning stage, five pieces of algorithms were used as being Naive Bayes (NB), Support Vector Machines (SVM), k Nearest Neighbor (kNN), Decision Trees (DT) and Random Forest (RF).

NB, is a simple learning method based on Bayes theorem which is a conditional probability theorem.

SVM method is composed of linear regressions and it is quite useful for high dimensional data. Basically SVM finds the plane which will create maximum discreteness among classes.

kNN algorithm is the process of finding the closest element in training data by using cosine similarity of test data class. K parameter shows how many neighbors of training data will be considered regarding test data.

DT is an inductive learning method which is realized by dividing the data. In this method having internal nodes and leaves, internal nodes contain questions where data had branching. Data passing through this question reach to the leaves representing the classes. Afterward, as per the value of this leaf, the classification process is applied to the data [22].

RF is a method which is formed by bringing together the outcomes of more than one decision tree. It has been developed to increase the success of decision trees [6, 7, 8, 9, 10, 11].

5. Experiment and Results

After the cluster in our hands is represented with tf-idf, Word2Vec and Doc2Vec, by being given to 5 ML algorithm with advisory learning, results were obtained. Information relating train-test stages which we have used during the trial stage, metrics which we have used, and the results we obtained have been detailed in this section.

5.1. Train and Test

For training and test processes, Weka, being a data mining application containing ML algorithm and measurement metrics, has been used [11].

To increase verification of outcomes, by assuming k=10 cross-validation method has been used. Cross-validation, is the name given to the process of continuously changing the training and testing data. For example for k=10 parameter first of all data is divided into 10 parts. One piece is used for testing and remaining 9 parts are used for training. This process is repeated for all other parts. By getting the average of outcomes received, it is recorded as the system success.

5.2. Metrics

Metrics which we use to define trial outcomes and which are also preferred in similar studies are Accuracy, Precision, Recall, F-Measure and Kappa statistics [13, 15, 19].

5.3. Results

By using 3 different text representation method and for 5 different ML algorithm system has been operated. Results belonging to 5 different metrics are given in Tables 2, 3, 4, 5, and 6.

As it can be seen in Table 2, highest values for CBOW method of the Word2Vec method were reached with kNN and RF algorithms. NB has been the algorithm reaching to the lowest value.

Table 2. CBOW Results.

Algorithm	Accuracy (%)	Precision	Recall	F-Measure	Kappa
NB	83.64	0.843	0.836	0.838	0.7495
SVM	94.45	0.950	0.945	0.944	0.9139
kNN	99.83	0.998	0.998	0.998	0.9974
DT	97.88	0.979	0.979	0.979	0.9674
RF	99.76	0.998	0.998	0.998	0.9964

As it can be seen in Table 3, highest values for Skip-Gram method of Word2Vec method were reached with kNN and RF algorithms. NB has been the algorithm reaching to the lowest value.

Table 3. Skip-Gram Results.

Algorithm	Accuracy (%)	Precision	Recall	F-Measure	Kappa
NB	83.50	0.871	0.835	0.826	0.7552
SVM	98.21	0.983	0.982	0.982	0.9728
kNN	99.86	0.999	0.999	0.999	0.9979
DT	91.06	0.925	0.911	0.912	0.8643
RF	99.93	0.999	0.999	0.999	0.999

As it can be seen in Table 4, highest values for DBOW method of Doc2Vec method were reached with SVM and RF algorithms. kNN has been the algorithm reaching to the lowest value.

Table 4. DBOW Results.

Algorithm	Accuracy (%)	Precision	Recall	F-Measure	Kappa
NB	39.13	0.385	0.391	0.385	0.0546
SVM	44.13	0.443	0.441	0.341	0.0493
kNN	36.16	0.360	0.362	0.361	0.0167
DT	39.76	0.364	0.398	0.365	0.0293
RF	43.62	0.399	0.436	0.375	0.0673

As it can be seen in Table 5, highest values for DM method of Doc2Vec method were reached with SVM and RF algorithms. DT has been the algorithm reaching to the lowest value.

Table 5. DM Results.

Algorithm	Accuracy (%)	Precision	Recall	F-Measure	Kappa
NB	36.16	0.403	0.362	0.308	0.0455
SVM	40.93	0.344	0.409	0.323	0.0037
kNN	35.12	0.349	0.351	0.350	-0.0005
DT	34.35	0.345	0.344	0.344	-0.0072
RF	41.39	0.360	0.414	0.333	0.0167

As it can be seen in Table 6, highest values for tf-idf method were reached with SVM and RF algorithms. kNN has been the algorithm reaching to the lowest value.

Table 6. tf-idf Results.

Algorithm	Accuracy (%)	Precision	Recall	F-Measure	Kappa
NB	53.21	0.526	0.532	0.525	0.2669
SVM	54.41	0.547	0.544	0.545	0.3029
kNN	38.94	0.528	0.389	0.323	0.1139
DT	47.94	0.475	0.479	0.476	0.1908
RF	56.48	0.565	0.565	0.560	0.3204

When Tables 2, 3, 4, 5, and 6 is investigated it is seen that among 3 different representation methods highest values were reached by Word2Vec. Results obtained by tf-idf ranked as the second while those of Doc2Vec were at the last row within all metrics. While CBOW and Skip-gram reached to similar values, DBOW, succeeded to reach to higher values than DM..

Verification values belonging to ML algorithms being used are given in Figures 1, 2, and 3.

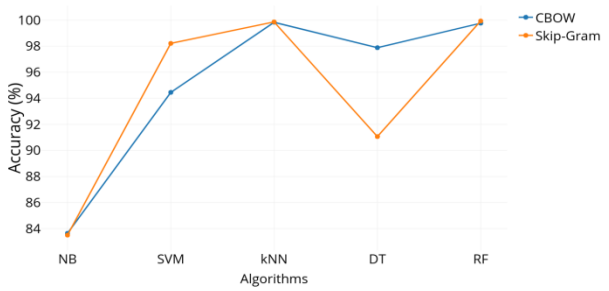

Figure 1. Word2Vec accuracy graph.

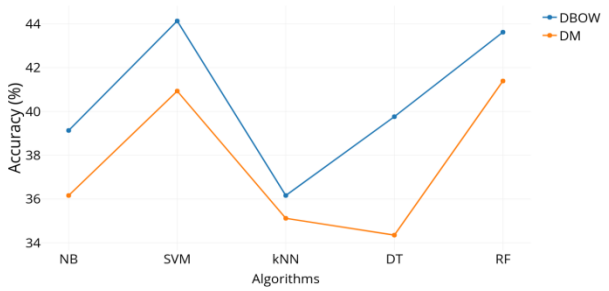

Figure 2. Doc2Vec accuracy graph.

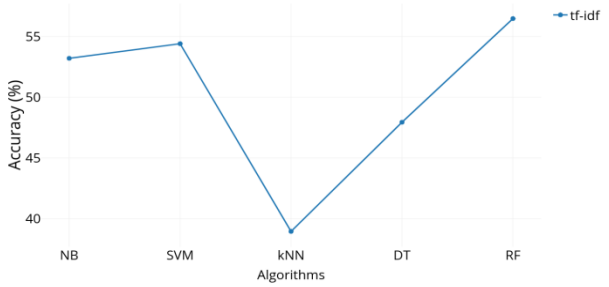

Figure 3. TF-IDF accuracy graph.

6. Conclusions

In this study sentiment analysis, being the most popular application of text classification in recent years has been realized on Turkish Tweets.

As a result of values being obtained in the study being realized, Word2Vec method has reached to higher values of accuracy, precision, recall, f-measure and kappa when compared with other two methods (tf-idf and Doc2Vec). Even though it produces similar values for CBOW and Skip-Gram being the two Word2Vec methods for NB, kNN and RF, in SVM, CBOW has obtained higher values and for DT Skip-Gram has obtained higher values.

Even though tf-idf method is lower than Word2Vec, it has succeeded to obtain better results than Doc2Vec. Doc2Vec has remained behind other two methods in all metrics. In values that are obtained for 5 different ML algorithm, DBOW, has succeeded to reach to higher values than DM.

The study being realized has produced similar results as obtained in the studies conducted in other languages for tf-idf and Word2Vec. When Table 2-6 is investigated, RF algorithm reaching to higher values has produced better results than other ML algorithms. Even though its calculation costs are higher when compared with other ML algorithms, its containing more detailed processes can be shown as the reason for its reaching to higher values. As Doc2Vec method is newer when compared with other two methods and as its development process still continues, these particulars could be shown as the reasons for its remaining behind when compared with others. This situation can also be shown as the reason of high success of Word2Vec in which more advanced calculation method is used with respect to tf-idf.

CBOW and Skip-Gram have produced similar results for different algorithms. Same situation is not valid for Doc2Vec. It could be stated that DBOW method is relatively better than DM as per the results being obtained.

Primary one of our study plans for the future is related with proposal of a new method for the formation way of vectors that are created for Word2Vec and with the testing of this method. Expansion of the data set we have formed can be specified as another future plan we have. Investigation of representation methods of used text with respect to model formation processes is among our future plans. Furthermore, realizing studies that would increase the performance of Doc2Vec is also among the targets.

References

[1] Berthold M., Cebron N., Dill F., Gabriel T., Kötter T., Meinl T., and Wiswedel B., "KNIME-the Konstanz İnformation Miner: Version 2.0 And Beyond," *AcM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 26-31, 2009.

[2] Bhavitha B., Rodrigues A., and Chiplunkar N., "Comparative Study of Machine Learning Techniques in Sentimental Analysis," *in Proceedings of International Conference on Inventive Communication and Computational Technologies*, Tamilnadu, pp. 216-221, 2017.

[3] Breiman L., "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[4] Bilgin M. and Şentürk İ., "Sentiment Analysis on Twitter Data with Semi-Supervised Doc2Vec," *in Proceedings of 2nd International Conference on Computer Science and Engineering*, Antalya, pp. 661-666, 2017.

[5] Bilgin M. and Köktaş H., "Word2Vec Based Sentiment Analysis for Turkish Texts," *in Proceedings of International Conference on Engineering Technologies*, Konya, pp. 106-109, 2017.

[6] Cutler A., Cutler D., and Stevens J., *Random Forests*, Ensemble Machine Learning, 2012.

[7] Çetin M. and Amasyali M., "Supervised and Traditional Term Weighting Methods for Sentiment Analysis," *in Proceedings of 21st Signal Processing and Communications Applications Conference*, Haspolat, pp. 1-4, 2013.

[8] Dadgar S., Araghi M., and Farahani M., "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification," *in Proceedings of International Engineering and Technology*, Coimbature, pp. 112-116, 2016.

[9] Dickinson B. and Hu W., "Sentiment Analysis of Investor Opinions on Twitter," *Social Networking*, vol. 4, no. 3, pp. 62-71, 2015.

[10] Fayyad U., Piatetsky-Shapiro G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," *Al magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[11] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I., "The Weka Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[12] Kaur A. and Gupta V., "A Novel Approach for Sentiment Analysis of Punjabi Text Using SVM," *The International Arab Journal of Information Technology*, vol. 14, no. 5, pp. 707-712, 2017.

[13] Khanna S. and Agarwal S., "An Integrated Approach towards the Prediction of Likelihood of Diabetes," *in Proceedings of International Conference on Machine Intelligence and Research Advancement*, Katra, pp. 294-298, 2013.

[14] Le Q. and Mikolov T., "Distributed Representations of Sentences and Documents," *in Proceedings of International Conference on Machine Learning*, Beijing, pp. 1188-1196, 2014.

[15] Mahardhika Y., Sudarsono A., and Barakbah A., "An Implementation of Botnet Dataset to Predict Accuracy Based on Network Flow Model," *in Proceedings of International Electronics Symposium on Knowledge Creation and Intelligent Computing*, Surabaya, pp. 33-39, 2017.

[16] Mikolov T., Chen K., Corrado G., and Dean J., "Efficient Estimation of Word Representations in Vector Space," *in Proceedings of International Conference on Learning Representations*, Arizona, pp. 1-12, 2013.

[17] Prabhat A. and Khullar V., "Sentiment Classification on Big Data Using Naïve Bayes And Logistic Regression," *in Proceedings of International Conference on Computer Communication and Informatics*, Coimbatore, pp. 1-5, 2017.

[18] Polpinij J., Srikanjanapert N., and Sopon P., "Word2Vec Approach for Sentiment Classification Relating to Hotel Reviews," *in Proceedings of 13th International Conference on Computing and Information Technology*, Bangkok, pp. 308-316, 2017.

[19] Raut, M. and Barve S., "A Semi-Automated Review Classification System Based on Supervised Machine Learning," *in Proceedings of 1st International Conference on Intelligent Systems and Information Management*, Aurangabad, pp. 127-133, 2017.

[20] Şahin G., "Turkish Document Classification Based on Word2Vec and SVM Classifier," *in Proceedings of 25th Signal Processing and Communications Applications Conference*, Antalya, pp. 1-4, 2017.

[21] Tang D., Wei F., Yang N., Zhou M., Liu T., and Qin B., "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," *in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, pp. 1555-1565, 2014.

[22] Vijayan V., Bindu K., and Parameswaran L., "A Comprehensive Study of Text Classification Algorithms," *in Proceedings of International Conference on Advances in Computing, Communications and Informatics*, Udupi, pp. 1109-1113, 2017.

[23] Xue B., Fu C., and Shaobin Z., "A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec," *in Proceedings of IEEE International Congress on Big Data*, Anchorage, pp. 358-363, 2014.

[24] Zhang D., Xu H., Su Z., and Xu Y., "Chinese Comments Sentiment Classification Based on Word2vec and Svmperf," *Expert Systems with Applications*, vol. 42, no. 4, pp.1857-1863, 2015.

Metin Bilgin received the Ph.D. degree in Computer Engineering from Yıldız Technical University in 2015. He is currently assistant professor in the Department of Computer Engineering, Bursa Uludağ University, Turkey. His current research interests include machine learning, natural language processing and text classification.

Haldun Köktaş is currently pursuing MSc at Mechatronics Engineering Department in Bursa Technical University. His research interests are machine learning, Natural language processing, mechanical design of robots and active exoskeletons.