ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2023, 24(1), pp. 23-34, DOI:10.18038/estubtda.1175001

RESEARCH ARTICLE

ON THE EFFECTIVENESS OF PARAGRAPH VECTOR MODELS IN DOCUMENT SIMILARITY ESTIMATION FOR TURKISH NEWS CATEGORIZATION

Ali YÜREKLİ *

Department of Computer Engineering, Faculty of Engineering, Eskişehir Technical University, Eskişehir, Turkey

ABSTRACT

News categorization, which is a common application area of text classification, is the task of automatic annotation of news articles with predefined categories. In parallel with the rise of deep learning techniques in the field of machine learning, neural embedding models have been widely utilized to capture hidden relationships and similarities among textual representations of news articles. In this study, we approach the Turkish news categorization problem as an ad-hoc retrieval task and investigate the effectiveness of paragraph vector models to compute and utilize document-wise similarities of Turkish news articles. We propose an ensemble categorization approach that consists of three main stages, namely, document processing, paragraph vector learning, and document similarity estimation. Extensive experiments conducted on the TTC-3600 dataset reveal that the proposed system can reach up to 93.5% classification accuracy, which is a remarkable performance when compared to the baseline and state-of-the-art methods. Moreover, it is also shown that the Distributed Bag of Words version of Paragraph Vectors performs better than the Distributed Memory Model of Paragraph Vectors in terms of both accuracy and computational performance.

Keywords: Turkish news categorization, Text classification, Neural embeddings, Paragraph vectors, Document similarity

1. INTRODUCTION

Text classification, which is the process of assigning text documents to one or more predefined categories [1], is a challenging task in machine learning due to the unstructured nature of textual data. In addition to the mentioned amorphousness, the substantial growth of data stored as text in information systems arises the need for effective and accurate text classification techniques that automatically organize data into categories [2].

The problem of text classification finds applications in a wide variety of domains [3] such as natural language inference, sentiment analysis and question answering. One such common application area of text classification is news categorization, in which the primary goal is to associate a given news article with a predefined news category such as economy, politics, sports, or technology. In today's era of digital media, where users prefer to follow daily news through online platforms [4], the ability to automatically categorize news articles is a necessity to automate and ease the data management and business procedures.

This study focuses on Turkish news categorization and approaches the problem as an ad-hoc retrieval task. We investigate the effectiveness of learning distributed word representations to capture semantic similarities of Turkish news articles. We employ Paragraph Vectors [5] (also known as Doc2Vec in the literature) to create neural embeddings of documents, and then perform top-*k* document retrieval on those embeddings to categorize given news articles. Although Doc2Vec has been successfully utilized for feature extraction and input vector generation in several news classification approaches [6-8], the effectiveness of paragraph vector models has not been well-explored in an ad-hoc retrieval setting.

*Corresponding Author: aliyurekli@eskisehir.edu.tr
Received: 16.09.2022 Published: 29.03.2023

During the experimental phase of the study, we evaluate several models combining two paragraph vector architectures with ad-hoc retrieval. The experiments performed on a well-known Turkish news collection [9] show that the proposed approach can reach up to 93.5% classification accuracy, which results in more accurate predictions than the baseline benchmark methods provided by Kılınç et al. [9]. Furthermore, our results are highly close and comparable to the current state-of-the-art [6].

1.1. Contributions and Organization

In this study, we propose an ad-hoc retrieval system for the task of Turkish news categorization. In order to evaluate the effectiveness of the system, extensive experiments are conducted on a collection of Turkish news articles. The main contributions of the study can be listed as follows:

- An end-to-end news categorization approach based on paragraph vectors and top-k document retrieval is proposed.
- Different modalities of paragraph vector learning architectures are explored.
- All source code and data required to reproduce the experimental results are made publicly available for interested researchers (see Section 4.2).

The rest of the paper is organized as follows. Section 2 provides a brief theoretical background of paragraph vectors and their learning architectures. Section 3 introduces the proposed ad-hoc retrieval approach in the study. Section 4 presents the experimental work and elaborates on the results. Finally, Section 5 includes future research directions and concludes the work.

2. THEORETICAL BACKGROUND

Neural embedding models have proven to be successful on a variety of downstream natural language processing tasks, including, but not limited to, text classification [10, 11], spam filtering [12, 13], sentiment analysis [14, 15], and named entity recognition [16]. The common intuition behind these models is to extract and learn high-quality representations that are capable of capturing word similarities at a semantic level with good compositionality [17].

Wod2Vec [18] is such an unsupervised neural embedding model that forms a vector mapping at the word level. Based on the hypothesis that words with similar meanings exhibit close distances [19], Word2Vec leverages this mapping to derive both syntactic and semantic similarities between words. Given a sequence of words $w_1, w_2, \ldots, w_T$, the model maximizes the average log probability of the next words within a window size of k using the equation given in (1).

$$\frac{1}{T}\sum_{t=k}^{T-k} \log\, p\,(w_t \mid w_{t-k}, \ldots, w_{t+k}) \tag{1}$$

Let W, U, b, and h denote the word embedding matrix, parameters of the softmax function, and the concatenation (or average) of word vectors from W, respectively. Then, Word2Vec calculates the predictions using the equations shown in (2) and (3).

$$p(w_t \mid w_{t-k}, \ldots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2}$$

$$y = b + Uh\,(w_{t-k}, \ldots, w_{t+k}; W) \tag{3}$$

Paragraph Vectors, introduced by Le and Mikolov [5], is an extension to Word2Vec that learns continuous distributed vector representations for any variable-length textual data (e.g., sentences, paragraphs, or documents). In addition to the word embedding matrix W in Word2Vec, Paragraph

Vectors framework also employs a paragraph embedding matrix D that contains the vector mapping of each paragraph as its columns. As a result of this notable difference, the equation in (3) is re-formulated to include paragraph embeddings as shown in (4).

$$y = b + Uh\ (w_{t-k}, \dots, w_{t+k}; W, D) \tag{4}$$

The training of paragraph vectors can be performed using two architectures, which are the Distributed Memory Model of Paragraph Vectors (PV-DM) and the Distributed Bag of Words version of Paragraph Vectors (PV-DBOW). Both of these architectures consider the semantics and order of words together when learning the embeddings. Therefore, paragraph vector models result in better representations than traditional Bag of Words (BOW) methods [5].

The PV-DM architecture acts as a memory that remembers missing information in the current context. The main idea behind PV-DM is to sample consecutive words from some piece of text and predict target word from these set of words that are regarded as the input. Figure 1 presents an illustration of learning paragraph vectors using PV-DM. Given a phrase (e.g., “Türkiye Büyük Millet Meclisi”), the algorithm learns to predict the target word (“meclisi”) based on the given context words (“türkiye”, “büyük”, and “millet”). This approach resembles the Continuous Bag of Words (CBOW) method in Word2Vec [18].

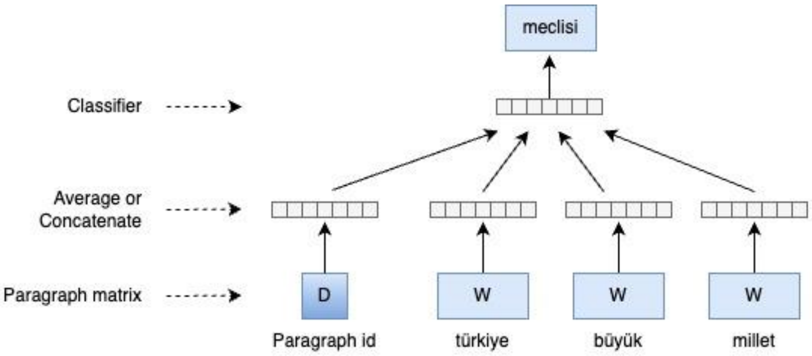

Figure 1. An illustration of learning paragraph vectors using PV-DM. The figure is derived from [5], the only difference is that the sample sentence is given in Turkish language.

On the other hand, the PV-DBOW architecture ignores the context words in the input and predicts words randomly sampled from the paragraph in the output. As illustrated in Figure 2, the output is the predictions of context words (“türkiye”, “büyük”, “millet”, and “meclisi”) for a given document indicated by a paragraph identifier. In contrast to PV-DM, PV-DBOW is inspired from the skip-gram method in Word2Vec [18].

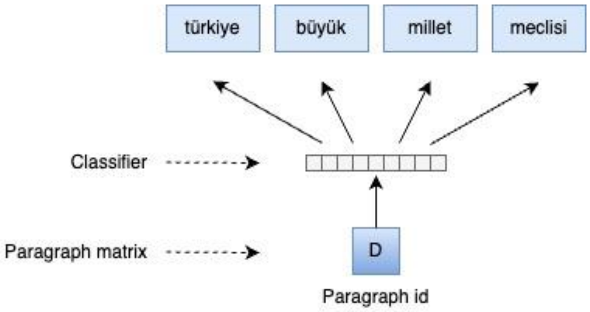

Figure 2. An illustration of learning paragraph vectors using PV-DBOW. The figure is derived from [5], the only difference is that the sample sentence is given in Turkish language.

3. PROPOSED APPROACH

Based on the idea of similar documents in content are likely to point the same point [20], we propose a semantic document similarity estimation approach for Turkish news categorization. Given a collection of news articles, a Doc2Vec model is trained using either PV-DM or PV-DBOW paragraph learning architectures. The learned embeddings are then used to capture syntactic and semantics similarities between news articles. Consequently, top-*k* documents retrieved for an unseen test instance constitute the news category inference for that instance.

The proposed approach consists of three main stages, which are *(i)* document preprocessing, *(ii)* paragraph vector learning, and *(iii)* document similarity estimation. Figure 3 illustrates the overall architecture of the proposed Turkish news categorization approach.

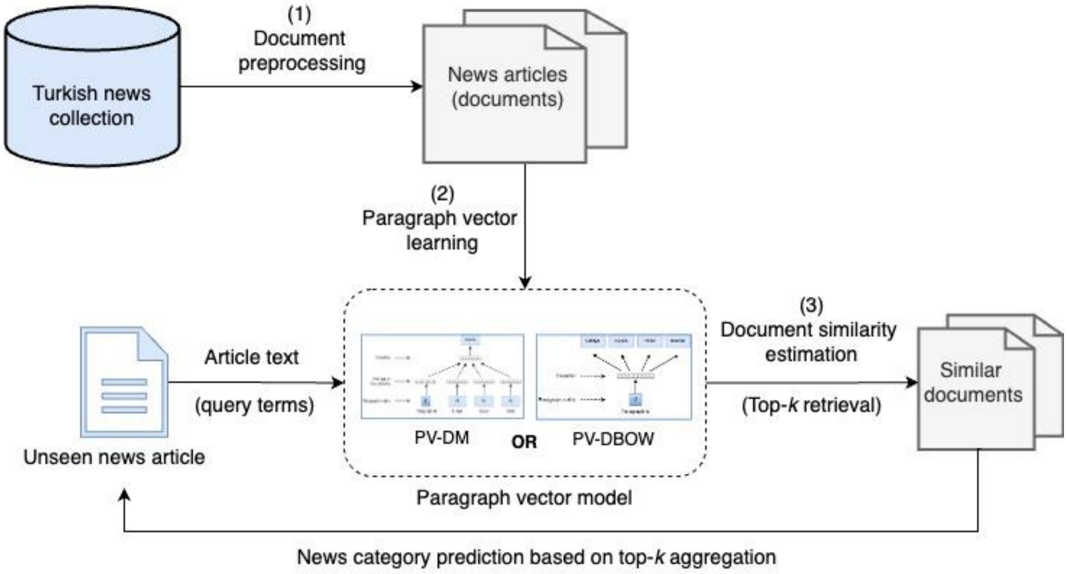
Turkish news collection
(1) Document preprocessing
News articles (documents)
(2) Paragraph vector learning
Unseen news article
Article text
(query terms)
PV-DM
OR
PV-DBOW
Paragraph vector model
(3) Document similarity estimation
(Top-k retrieval)
Similar documents
News category prediction based on top-k aggregation

Figure 3. The overall architecture of the proposed Turkish news categorization approach.

3.1. Document Preprocessing

As practiced in the majority of text classification tasks in the literature [2], we first apply a set of well-known NLP operations on the news collection. The steps we employ for the purpose of homogenizing the data are lowercase conversion, special character (i.e., non-alphanumeric characters like whitespaces and punctuations) elimination, stemming, stop word elimination, and tokenization, respectively. Since stemming is a language-specific process, we use Zemberek-NLP[1] that is a highly favorable toolkit to perform Turkish morphological analysis.

3.2. Paragraph Vector Learning

In the prior stage, where data preprocessing is carried out, the news articles are transformed into logical document representations. The next stage of the proposed approach aims to learn paragraph embeddings of these documents via PV-DM and PV-DBOW architectures.

[1] https://github.com/ahmetaa/zemberek-nlp

During paragraph learning, we instantiate and train several Doc2Vec models varying with respect to notable factors such as the learning architecture and vector size. While models trained using PV-DM utilize both context word vectors and document vectors, the models based on PV-DBOW use only document vectors. Detailed information about the parameters employed at this stage can be found in Section 4.2.

When implementing the paragraph vector learning stage, we use Gensim[2], which is an open-source Python library designed to process unstructured text data using unsupervised machine learning algorithms.

3.3. Document Similarity Estimation

During document similarity estimation, we assess how similar two news articles are by the cosine similarity of their corresponding paragraph embeddings [21]. Suppose *A* and *B* are two paragraph vectors representing two news articles from the news collection. Then, the cosine of the angle θ between these vectors corresponds to their measure of similarity. The mathematical equation of cosine similarity between *A* and *B* is presented in (5).

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5}$$

The ability to estimate document similarities also allows us to employ top-*k* document retrieval for a given news article. When an unseen test instance is given to the system, the *k* most similar documents, both syntactically and semantically, are retrieved from the collection. Afterwards, news categorization is performed by aggregating the actual categories of the retrieved documents.

4. EXPERIMENTAL WORK

This section presents the experimental work performed during the study. First, we describe the Turkish news collection and evaluation metrics to create and evaluate classification models. Then, we explain our full-factorial setup aiming to explore the factors affecting overall performance of the proposed system. Finally, we present our results including our discussions.

4.1. Dataset and Evaluation Metrics

In this study, we use the public TTC-3600 dataset [9] to develop and evaluate models for Turkish news categorization. As its name implies, the collection consists of 3600 Turkish news articles and their corresponding categories. In the whole collection, there exist exactly 600 news articles for six major news categories, which are culture, economy, health, politics, sports, and technology. Accordingly, TTC-3600 can be considered an instance of balanced datasets.

Besides publicity and balance, there exists one more factor promoting TTC-3600 as a valuable dataset with high research potential. During the publication of the dataset, Kılınç et al. [9] had also provided a baseline presenting the performances of a set of well-known classifiers along with data preprocessing and feature selection aspects. For all these reasons, TTC-3600 has been widely used as a benchmark dataset in several studies [6, 22-25] concerning Turkish text classification.

For the purpose of training document similarity models and evaluating these models in terms of their classification performance, we split TTC-3600 into train and test sets using the stratified k-fold cross

[2] https://radimrehurek.com/gensim/index.html

validation method with k=10. This technique preserves the percentage of samples for each class and reduces experimental variance [26]. Consequently, for each fold, 90% of the data is used for training and the remaining 10% for testing. Table 1 shows the distribution of news categories in TTC-3600 to the train and test sets on the fold basis.

Table 1. The distribution of news categories obtained by stratified k-fold cross validation with k=10.

Category	Number of train instances (per fold)	Number of test instances (per fold)
Culture	540	60
Economy	540	60
Health	540	60
Politics	540	60
Sports	540	60
Technology	540	60

During the experimental evaluation phase of the study, we employ the accuracy metric to measure the classification performance of the proposed models. As given in (6), accuracy can simply be defined as the ratio of correct predictions to total predictions. It is noteworthy that Kılınç et al. [9] also uses accuracy as the primary evaluation metric in their Turkish news classification benchmark.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{6}$$

4.2. Experimental Setup

Classification accuracy of the proposed approach depends on three main factors: *(i)* the training algorithm of the paragraph vectors, *(ii)* the dimensionality of feature vectors (i.e., vector size), and *(iii)* the number of documents retrieved to categorize a given news article. Table 2 presents these factors along with their notations, scope, and set of intended options.

Table 2. The factors affecting classification accuracy of the proposed approach.

Factor	Notation	Scope	Options
Paragraph vector architecture	*PV*	Training	PV-DM, PV-DBOW
Vector size	*VS*	Training	40, 80, …, 400
Top-*k* documents	*K*	Classification	1, 2, …, 20

As illustrated in Table 2, two of the factors affecting classification accuracy (*PV* and *VS*) are defined during the learning phase of paragraph embeddings. Therefore, tuning their corresponding options requires training of new models from scratch. However, *K* can be determined dynamically when classifying the test instances. In other saying, top-*k* document retrieval does not require any additional model training.

Since we search for the best-performing configuration of the proposed system, we follow a full-factorial experimental design that models and evaluates all possible combinations of above-mentioned factors. In total, 20 different models (i.e., 2 options for *PV* and 10 options for *VS*) are trained. Each of these models is tested with 20 choices of *K* varying from 1 to 20. The ranges for both *K* and *VS* values are determined empirically. Consequently, the proposed approach is subjected to 400 extensive experiments.

4.2.1. Technical details and code availability

All experiments are performed on a MacBook Pro with the M1 chip and 16GB RAM. In this environment, the final model (i.e., the best-performing configuration) can be trained and tested within less than a few minutes. The source code required to reproduce the experimental results is publicly available on GitHub[3]. In addition to the source code, the repository also contains a script to perform all the above-mentioned experiments from scratch.

4.3. Results and Discussions

4.3.1. Offline evaluations

In line with our experimental setup, we train and evaluate a total number of 200 distinct system configurations to explore the effectiveness of paragraph vector models over document retrieval for Turkish news categorization. Figure 4 presents the classification accuracy of all the configurations as a 3-D scatter plot, in which the x-axis represents *VS*, the y-axis represents *K*, and the z-axis represents accuracy, respectively. In addition, plus signs (+) colored in blue indicate the data points belonging to PV-DBOW models, and asterisk signs (*) colored in green show the data points belonging to PV-DM models.

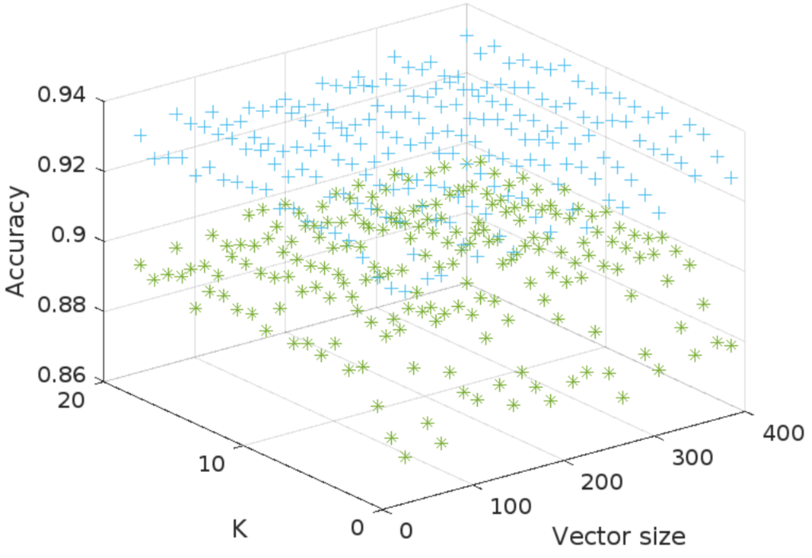

Figure 4. Data points showing the classification accuracies of PV-DM and PV-DBOW models regarding *VS* and *K*.

As illustrated in Figure 4, the data points of PV-DBOW models are located substantially higher on the z-axis than the data points of PV-DM models. While the accuracy of PV-DBOW models varies between 91.5% and 93.5%, the accuracy of PV-DM models is in the range of 87% to 90.5%. Accordingly, it can be concluded that PV-DBOW performs better than PV-DM in detecting similarities between Turkish news articles.

The choice of *VS* and *K* has a notable impact on news classification accuracy of the proposed approach. Particularly, the ranges in which PV-DM and PV-DBOW architectures achieve the best performance

[3] https://github.com/aliyurekli/turkish-news-categorization

(i.e., paragraph vector learning and document retrieval) also differ from each other. PV-DBOW achieves the highest accuracy of 93.51% with *VS*=360 and *K*=6, while PV-DM reaches 90.48% with *VS*=200 and *K*=12. The optimal *VS* option is in the range of 200-240 for PV-DM and 360-400 for PV-DBOW. Additionally, the optimal K value to be used in determining the news category is between 8-12 for PV-DM and 4-6 for PV-DBOW. Accordingly, in our setting, PV-DBOW makes more accurate predictions by choosing from a narrow document pool with larger vector sizes, while PV-DM succeeds by accessing more documents based on smaller vector representations.

For further insights on Turkish news categorization, we examine classification accuracy of the proposed system in terms of the news categories in the TTC-3600 dataset. Using the best-performing configuration of the proposed retrieval approach (i.e., settings chosen as *PV*=PV-DBOW, *VS*=360, and *K*=6), the classification accuracy of each news category is measured. As illustrated in Figure 5, news articles belonging to "Sports" category can be predicted accurately as high as 98%. Similarly, predictions in "Health", "Politics", and "Culture" categories are also satisfactory (i.e., approximately 96%). On the other hand, it seems that it is more difficult to categorize "Economy" and "Technology" news correctly when compared to the other categories. The accuracies for these two categories drop down to 87% and 88%, respectively.

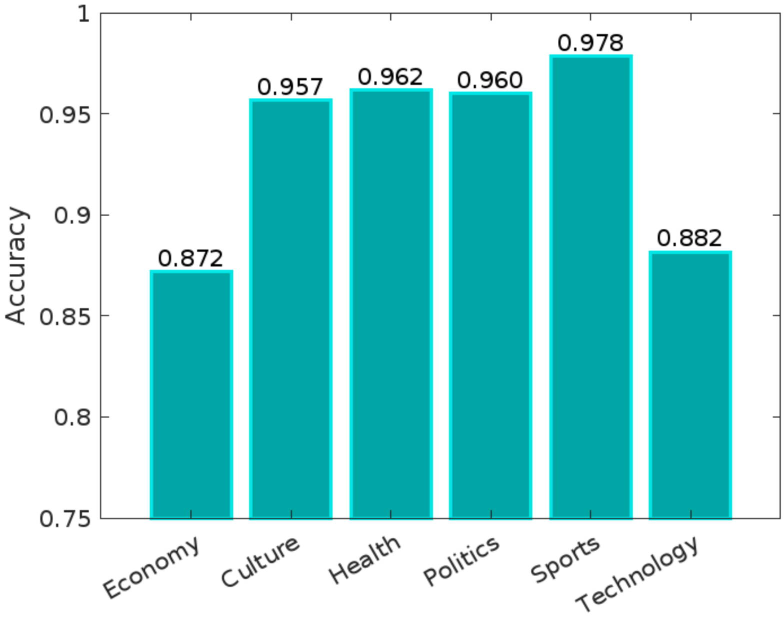

Figure 5. Classification accuracy achieved with the best-performing configuration of the system for news categories in the TTC-3600 dataset.

The accuracy analysis on the basis of news categories shows us that different categories achieve different degrees of classification performance. This observation indicates that documents from particular categories have varying characteristics in terms of semantics and structure. Nevertheless, uncovering the underlying reasons for the obvious accuracy decline in "Economy" and "Technology" might be a potential step towards overall performance improvement.

In addition to classification accuracy, we also evaluate paragraph vector learning architectures in terms of their training time. Figure 6 presents the time elapsed (in seconds) for model training according to the vector sizes varying from 40 to 400. As shown in the plot, PV-DBOW operates approximately 22% faster than PV-DM. Furthermore, as the vector size increases, the model training time for both paragraph vector learning architectures naturally increases.

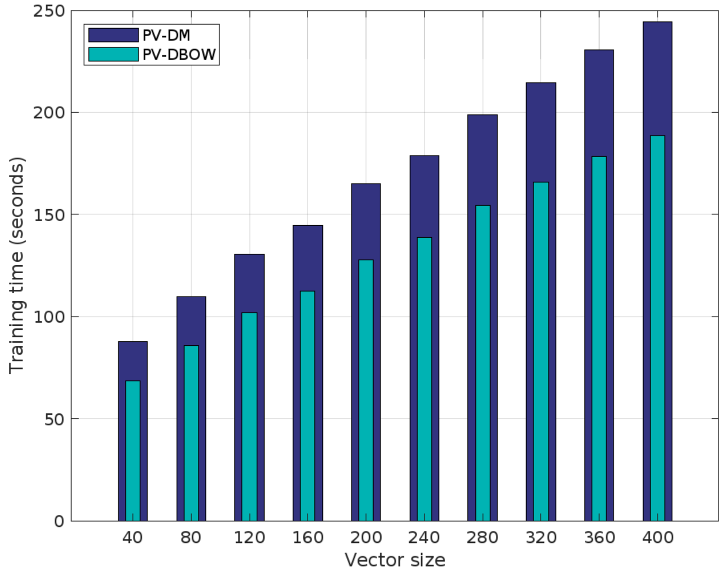

Figure 6. Comparison of paragraph vector learning algorithms in terms of training time with respect to varying vector size.

4.3.2. Comparison with the baseline and the state-of-the-art

In order to explore the effectiveness of the proposed system, we compile and compare a number of Turkish news categorization studies experimenting on the TTC-3600 dataset as the baseline and the state-of-the-art.

The baseline comprises of several classifiers evaluated in the Turkish news classification benchmark by Kılınç et al. [9]. The authors employ five well-known techniques, which are Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (J48), and Random Forest (RF), on TTC-3600. The models have been diversified with the options of preprocessing, stemming, correlation-based feature selection (CFS), and attribute ranking-based feature selection (ARFS). In studies following this benchmark work, the state-of-the-art results have been obtained by utilizing convolutional neural networks (CNN) on neural embeddings of Doc2Vec or Word2Vec models [6, 22].

Table 3. Comparison of the proposed approach with the baseline and state-of-the-art methods.

	Study	**Model**	**Accuracy (%)**
Baseline	Kılınç et al. [9]	KNN (stemming + CFS)	74.97
		J48 (stemming + ARFS)	79.39
		SVM (no stemming or feature selection)	86.03
		NB (stemming + ARFS)	87.19
		RF (stemming + ARFS)	**91.03**
State-of-the-art	Acı et al. [22]	CNN (Word2Vec + stemming)	93.30
	Dogru et al. [6]	CNN (Doc2Vec + stemming)	**94.17**
Our work		PV-DM (*VS*=200, *K*=12)	90.48
		PV-DBOW (*VS*=360, *K*=6)	**93.51**

Table 3 presents the classification accuracies on the TTC-3600 dataset of the approaches included in our comparison. As given in the table, our system provides better predictions than baseline methods and performs close to the state-of-the-art approaches that use CNN classifiers on Word2Vec or Doc2Vec embeddings. Therefore, the proposed model can be considered as an effective and alternative solution to the task of Turkish news categorization.

5. CONCLUSION

News categorization is a text classification task that automatically associates news articles with a predefined news category. Due to the idiosyncratic characteristics of natural languages, it is essential to develop elegant categorization techniques that take language-specific requirements and situations into consideration.

In this work, we focus on Turkish news categorization and investigate the effectiveness of paragraph vector models for document similarity estimation. The proposed approach utilizes the neural embeddings of Turkish news articles to retrieve documents with semantically and syntactically similar content. For unseen articles, the predictions are generated by aggregating category labels of top-*k* similar documents. The experiments performed on the TTC-3600 dataset show that our approach outperforms the baseline methods and performs highly close to the state-of-the-art. Specifically, learning paragraph vectors with PV-DBOW can reach up to 93.5% classification accuracy. On the other hand, the performance is bounded by 90.5% accuracy for the models trained with PV-DM. Our analyses also reveal that classification performance may vary based on news categories. For example, a sports news is more likely to be categorized correctly than an economy news.

In the near future, we are planning to refine the proposed system to hybridize PV-DM and PV-DBOW architectures. As emphasized by Le and Mikolov [5], the combination of these algorithms usually results in more consistent neural embeddings, which might also be a possible way to improve the overall classification accuracy of the proposed approach. Furthermore, our document similarity estimation method can be adapted for novel similarity measures [27].

ACKNOWLEDGEMENTS

The author would like to thank the editor and anonymous reviewers.

CONFLICT OF INTEREST

The author stated that there are no conflicts of interest regarding the publication of this article.

REFERENCES

[1] Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D. Text classification algorithms: A survey. Information ,2019; 10(4): 150.

[2] Uysal AK, Gunal S. The impact of preprocessing on text classification. Information Processing & Management, 2014; 50(1): 104-112.

[3] Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Computing Surveys, 2021; 54(3): 1-40.

[4] Skogerbø E, Winsvold M. Audiences on the move? Use and assessment of local print and online newspapers. European Journal of Communication, 2011; 26(3): 214-229.

[5] Le Q, Mikolov T. Distributed representations of sentences and documents. In: 31st International Conference on Machine Learning (ICML 2014); Beijing; China; 2014; pp. 1188-1196.

[6] Dogru HB, Tilki S, Jamil A, Hamed AA. Deep learning-based classification of new texts using doc2vec model. In: 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA); Riyadh, Saudi Arabia; 2021; pp. 91-96.

[7] Trieu LQ, Tran HQ, Tran MT. News classification from social media using twitter-based doc2vec model and automatic query expansion. In: Proceedings of the Eight International Symposium on Information and Communication Technology (SoICT 2017); Nha Trang, Vietnam; 2017; 460-467.

[8] Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2vec. Information Sciences, 2019; 477: 15-29.

[9] Kılınç D, Özçift A, Bozyigit F, Yıldırım P, Yücalar F, Borandag E. TT-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science, 2017; 43(2): 174-185.

[10] Guo B, Zhang C, Liu J, Ma X. Improving text classification with weighted word embeddings via a multi-channel TextCNN model. Neurocomputing, 2017; 363: 366-374.

[11] Pittaras N, Giannakopoulos G, Papadakis G, Karkaletsis V. Text classification with semantically enriched word embeddings. Natural Language Engineering, 2021; 27(4): 391-425.

[12] Fahfouh A, Riffi J, Mahraz MA, Yahyaouy A, Tairi H. PV-DAE: A hybrid model for deceptive opinion spam based on neural network architectures. Expert Systems with Applications, 2020; 157: 113517.

[13] Madisetty S, Desarkar MS. A neural network-based ensemble approach for spam detection in Twitter. IEEE Transactions on Computational Social Systems, 2018; 5(4): 973-984.

[14] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015); Santiago, Chile; 2015; pp. 959-962.

[15] Tang D, Wei F, Qin B, Yang N, Liu T, Zhou M. Sentiment embeddings with applications to sentiment analysis. IEEE Transactions on Knowledge and Data Engineering 2015; 28(2): 496-509.

[16] Unanue IJ, Borzeshi EJ, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of Biomedical Informatics, 2017; 76: 102-109.

[17] Ai Q, Yang L, Guo J, Croft WB. Analysis of the paragraph vector model for information retrieval. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR 2016); New York, USA; 2016. pp. 133-142.

[18] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations (ICLR 2013); Scottsdale, Arizona, USA; 2013.

[19] Sahlgren M. The distributional hypothesis. Italian Journal of Disability Studies, 2008; 20: 33-53.

[20] Benedetti F, Beneventano D, Bergamaschi S, Simonini G. Computing inter-document similarity with context semantic analysis. Information Systems, 2019; 80: 136-147.

[21] Yürekli A, Kaleli C, Bilge A. Alleviating the cold-start playlist continuation in music recommendation using latent semantic indexing. International Journal of Multimedia Information Retrieval, 2021; 10(3): 185-198.

[22] Acı Ç, Çırak A. Turkish news categorization using convolutional neural networks and word2vec. Bilişim Teknolojileri Dergisi, 2019; 12(3): 219-228 (in Turkish with an abstract in English).

[23] Borandağ E, Özçift A, Kaygusuz Y. Development of majority vote ensemble feature selection algorithm augmented with rank allocation to enhance Turkish text categorization. Turkish Journal of Electrical Engineering and Computer Sciences, 2021; 29(2): 514-530.

[24] Cimen E. A random subspace based conic functions ensemble classifier. Turkish Journal of Electrical Engineering and Computer Sciences, 2020; 28(4): 2165-2182.

[25] Wang H, Hong M. Supervised Hebb rule based feature selection for text classification. Information Processing & Management, 2019; 56(1): 167-191.

[26] Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter, 2010; 12(1): 49-57.

[27] Eminagaoglu M. A new similarity measure for vector space models in text classification and information retrieval. Journal of Information Science, 2022; 48(4): 463-476.