Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
DOI: https://doi.org/10.35414/akufemubid.1420120
AKÜ FEMÜBİD 24 (2024) 055105 (1181-1189)
AKU J. Sci. Eng. 24 (2024) 055105 (1181-1189)

Turkish Text Classification Based On Wrapper Feature Selection Using Particle Swarm Optimization

Parçacık Sürü Optimizasyonunu Kullanan Sarmalayıcı Öznitelik Seçimine Dayalı Türkçe Metin Sınıflandırma

*Makale Bilgisi / Article Info
Alındı/Received: 15.01.2024
Kabul/Accepted: 16.07.2024
Yayımlandı/Published: xx.xx.xxxx

Ezgi ZORARPACI*

İstanbul Üniversitesi, Bilgisayar ve Bilişim Teknolojileri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

Abstract

The vast majority of the digital era data is stored as text. Text mining is an integral part of data mining. Text classification (TC) is a natural language processing (NLP) operation often needed in text mining. This operation is needed in numerous kinds of research such as information retrieval, document classification, language detection, sentiment analysis, etc. According to the literature, the filter feature selection methods have often been applied to reduce the dimensionality of data in Turkish TC. However, the wrapper-based feature selection methods can provide better classification accuracies than the filter methods. Motivated by this idea, a Turkish TC method based on wrapper feature selection using particle swarm optimization algorithm (PSO) and multinomial naive bayes (MNB) classifier is proposed in this study. TTC-3600 Turkish news texts are used for TC in the experiments. The proposed method achieves a classification accuracy of 94.55% on TTC-3600 Turkish news text dataset by using stemming Tf-Idf features. Hence, it produces competitive accuracies to the cutting-edge Turkish TC methods.

Keywords: *Feature selection; Natural language processing; Text classification; Text mining.*

Öz

Dijital çağ verilerinin büyük çoğunluğu metin olarak depolanmaktadır. Metin madenciliği veri madenciliğinin ayrılmaz bir parçasıdır. Metin sınıflandırma (TC), metin madenciliğinde sıklıkla ihtiyaç duyulan bir doğal dil işleme (NLP) işlemidir. Bu işleme bilgi erişimi, belge sınıflandırma, dil tespiti, duygu analizi vb. birçok araştırmada ihtiyaç duyulmaktadır. Literatüre göre, Türkçe TC'de veri boyutunun azaltması için filtre öznitelik seçme yöntemleri sıklıkla uygulanmaktadır. Ancak sarmalayıcı tabanlı öznitelik seçme yöntemleri, filtre yöntemlerine kıyasla daha iyi sınıflandırma doğruluğu sağlayabilir. Bu fikirden hareketle, bu çalışmada parçacık sürüsü optimizasyon algoritması (PSO) ve çok terimli naive bayes (MNB) sınıflandırıcısını kullanan sarmalayıcı öznitelik seçim yöntemi tabanlı bir Türkçe TC metodu önerilmektedir. Deneylerde TC için TTC-3600 Türkçe haber metinleri kullanılmıştır. Önerilen yöntem, köklerine ayrılmış (stemming) Tf-Idf özniteliklerini kullanarak TTC-3600 Türkçe haber metni veri kümesinde %94,55'lik bir sınıflandırma doğruluğuna ulaşmaktadır. Böylece en son Türkçe TC yöntemleriyle rekabet edebilen sınıflandırma doğrulukları üretmektedir.

Anahtar kelimeler: *Öznitelik seçimi; Doğal dil işleme; Metin sınıflandırma; Metin madenciliği.*

1. Introduction

Digital age data, such as customer reviews, news, social media and countless digital documents, are progressively produced through various sources (Ghareb *et al*. 2016, Borandağ *et al*. 2021). The vast majority of this data is stored as text. Hence, text mining is always hot research field in data mining (Köksal and Yılmaz 2022). At the same time, TC is an NLP task frequently used in text mining and is defined as the automatic assignment of text to a set of predefined categories. This task is used in various NLP tasks such as information retrieval, customer review analysis, document classification, topic detection, author identification, bioinformatics, content management, web page classification, language detection, information filtering, spam detection, document summarization, and sentiment analysis.

In the literature, there are many TC studies in other languages but, few studies on this subject for Turkish language (Kılınç 2016). The reasons are due to language specificity, data availability, and research focus. Turkish is a less commonly studied language compared to English, resulting in fewer resources and less attention from researchers. Limited publicly available Turkish text datasets (Köksal and Akgül 2022) and resources make it challenging for researchers to conduct comprehensive

*Sorumlu Yazar/Corresponding Author: Ezgi Zorarpacı e-posta/e-mail: ezgi.zorarpaci@istanbul.edu.tr

studies. Most research in NLP and TC focuses on widely spoken languages like English.

To remedy this problem, here are some suggestions: (i) creating and releasing more Turkish text datasets for researchers to use in their studies, (ii) encouraging collaboration among researchers to pool resources, share datasets, and collectively work on improving Turkish text classification methods, (iii) making research findings and resources openly accessible to the community to facilitate further studies and advancements in Turkish text classification.

Kılınç *et al*. (2017) have emphasized the lack of a comparison dataset for Turkish TC studies. As a result, they have created TTC-3600 dataset which is a recent publicly accessible and has well-documented news containing 3600 news texts equally apportioned across six classes: economy, arts, culture, health, politics, sports, and technology. They have applied a few commonly used text classification algorithms such as naive bayes (NB), support vector machines (SVM), k-nearest neighbor (K-NN), C4.5, and random forest (RF) on TTC-3600 dataset. Consequently, if the data is stemmed by using the Zemberek library, RF combined with the feature selection method based on feature ranking reaches a classification accuracy of 91.03%.

After creating TTC-3600 dataset, Aci and Çirak (2019) have trained two convolutional neural network (CNN) models using raw and stemming texts of this dataset. Their study achieves f1 scores at 93.3% and 90.1% on stemming and non-stemming data, respectively.

Kuyumcu *et al*. (2019) have used FastText word embedding method to categorize TTC-3600 dataset; without pre-processing stages for the text data. They have attained an accuracy of 93.43% with the training of NB, K-NN, and C4.5 models separately. They have also claimed that their work is the first to categorize a comparison dataset by using FastText.

Doğru *et al*. (2021) have analyzed the impact of Doc2Vec word embedding on CNN, gaussian NB, RF, NB, and SVM text classification models using pre-processed Turkish and English datasets that are TTC-3600 and BBC news texts. CNN generates an f1 score of 94.17% on TTC-3600 dataset, overshadowing the previous studies.

Feature selection, a critical step in text mining, is pivotal in improving model performance, interpretability, and scalability by identifying and retaining the most informative attributes. Recent studies have explored the application of deep learning techniques for feature selection in text mining tasks. Methods such as autoencoders, CNNs, and recurrent neural networks (RNNs) have been employed to automatically learn informative representations from text data. For instance, Xie *et al*. (2019) have proposed a novel feature selection approach based on deep autoencoder networks, achieving the state-of-the-art (SOTA) performance for TC. Hybrid feature selection methods, combining traditional statistical techniques with machine learning algorithms, have emerged as promising approaches for text mining. These methods aim to leverage the strengths of both approaches to achieve superior feature selection performance.

Researchers have explored methods to select features that are relevant to multiple labels simultaneously, improving the efficiency and effectiveness of multi-label TC models. Zhang *et al*. (2023) have proposed a novel group-preserving label-specific feature selection approach for multi-label TC. In low-resource settings where labeled data is scarce, feature selection becomes crucial for building effective text mining models. Researchers have explored feature selection techniques that can adapt to limited training data while maintaining performance (Meetei *et al*. 2021).

These recent studies demonstrate the diverse methodologies and applications of feature selection in text mining, ranging from deep learning approaches to hybrid methods and interpretable feature selection techniques. The advancements in feature selection contribute to enhancing the efficiency, effectiveness, and interpretability of text mining models across various domains and settings.

In addition, researchers have paid attention to the feature selection based Turkish TC studies in the literature. For instance, a recent filter feature selection method, named trigonometric comparison measure (TCM) considering relative document frequencies have been proposed for TC. The proposed method has been compared to eight well-known filter feature selection methods including balanced accuracy measure (ACC2), IG, chi-squared (CHI), odds ratio (OR), gini index (Gini), deviation from a poisson distribution (DP), distinguishing feature selector (DFS), and normalized difference measure (NDM). The proposed method has been evaluated on ten datasets including TTC-3600 by using MNB and SVM classifiers (Kim and Zzang 2019). Heyong and Ming (2019) have developed supervised hebb rule based feature selection (HRFS) for TC. MNB classifier achieves a classification accuracy of 90% with 500 features selected by HRFS. Parlak (2023) has emphasized that pre-processing is one of the key components to improve the performance of TC. In the

study, Gini, CHI, OR, extensive feature selector (EFS), and NDM are used for feature selection. According to the experimental results, SVM reaches an f1 score of 78.1% with 1000 features selected by EFS on TTC-3600 dataset.

Zorarpaci (2023) has made a performance evaluation to measure the success of density peaks clustering algorithm (DPC), a new semi-supervised machine learning method, for Turkish TC. To improve the performance of DPC, it has been proposed to use IG filter feature selection method, which eliminates irrelevant Tf-Idf (term frequency-inverse document frequency) features in this study. TTC-3600 benchmark dataset has been used to evaluate the study's contribution to the literature, analyze the findings, and compare the results with the existing results. The proposed method reaches an accuracy of 99.69% for the categorization of TTC-3600 dataset.

According to the literature, well-known filter feature selection methods have frequently been employed for dimensionality reduction in Turkish TC. On the other hand, the wrapper based feature selection methods can result in better classification accuracies as compared to the filter methods. The wrapper methods employ a certain machine learning algorithm to assess the feature subsets, which can lead to selecting the most relevant features for that particular model. These methods can capture interactions between features, which can be beneficial in complex datasets. Since the wrapper methods directly consider the performance of the model, they can potentially provide better feature subsets tailored to the model's performance. However, the wrapper methods are computationally expensive as they need to train and evaluate the model for each subset of features, making them slower and resource-intensive.

The filter methods are computationally less expensive since they evaluate the features based on their intrinsic properties without involving model training. The filter methods might not consider feature interactions or the effects of features on the model's performance, potentially missing important relationships between features. Since the filter methods do not consider the model's performance, they may select subsets of features that are not necessarily optimal for a specific model. The filter methods may not always select the most relevant features for a specific model.

Consequently, a Turkish TC based on wrapper feature selection is proposed in this study. In the proposed wrapper feature selection method, PSO is used to discover the possible feature subsets and MNB is utilized to assess the qualities of these feature subsets. The combination of global and local search capabilities, adaptability, efficiency, and flexibility makes PSO a powerful optimization technique for wrapper feature selection in machine learning tasks.

On the other hand, MNB classifiers are computationally efficient, making them suitable for large-scale TC tasks with high-dimensional feature spaces. Due to their simplicity and efficiency, MNB classifiers can scale well with the size of the dataset and the number of classes. They can handle a large number of features (words or n-grams) without significantly increasing computational complexity or memory requirements. Despite their simple assumptions and independence assumptions, MNB classifiers often perform surprisingly well in practice, especially in TC tasks.

Based on these reasons, MNB is also applied to classify Turkish news texts (i.e., TTC-3600) using the optimal feature subset specified by the proposed wrapper feature selection method.

The rest of the paper is organized as follows. Section 2 introduces the methods used in the proposed approach. Section 3 describes the proposed wrapper feature selection method in detail. In Section 4, the datasets used in the experiments, the experimental setup, and the experimental results of Turkish TC are given. In Section 5, the paper is concluded.

2.Materials and Methods

In this part of the paper, the explanations related to the methods used in this study are presented as subsections.

2.1 MNB

MNB is an NB algorithm for multinomially distributed data. Equation (1) is simply the multinomial distribution (McCallum and Nigam 1998):

$$P(d_i|c_j;\theta) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j;\theta)^{N_{it}}}{N_{it}!} \quad (1)$$

where d_i represents the i^{th} document, N_{it} is the number of occurrences of feature w_t in document d_i (i.e., i^{th} text). The parameters of the generative component for each class are the probabilities for each feature and are indicated by $\theta_{w_t|c_j} = P(w_t|c_j;\theta)$, where $0 \leq \theta_{w_t|c_j} \leq 1$ and $\sum_t \theta_{w_t|c_j} = 1$. Bayes-optimal estimations of these parameters are computed using a set of labeled training data. After the parameter estimation step, the test data is classified by computing the posterior probability for each class. The class label with the highest probability is assigned to each test instance (McCallum and Nigam 1998).

2.2 PSO

PSO (Kennedy and Eberhart 1995) is a swarm-based stochastic optimization algorithm. It mimics the social behaviors of animals such as insects, birds, fishes, etc. The swarms follow a cooperative path to discover the food. Each particle in the swarm sustains altering the search model based on its own and other particles' experiences (Wang *et al*. 2018). In Figure 1, an entire presentation of PSO algorithm is given.

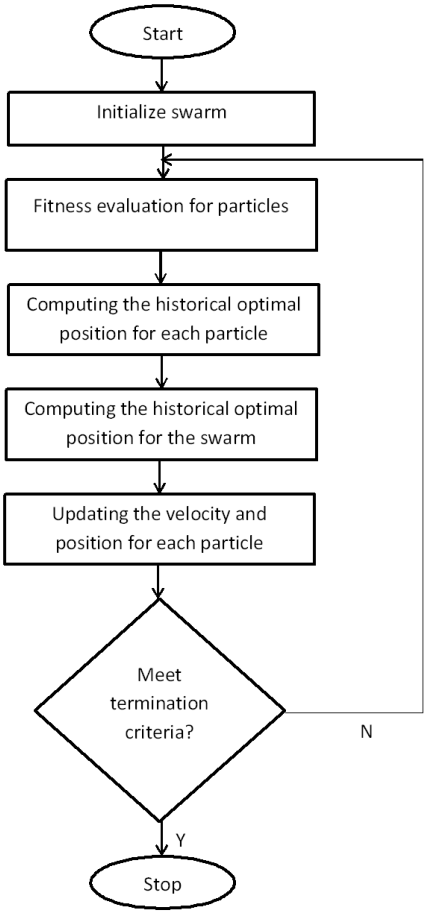

Figure 1. Flowchart of PSO (Wang *et al*. 2018).

Let be the size of swarm is N, the position vector for each particle is represented as $X_i = (x_{i_1}, x_{i_2}, \dots, x_{i_d})$, where d is the number of parameters for the problem in hand. Velocity vector is represented as $V_i = (v_{i_1}, v_{i_2}, \dots, v_{i_d})$. Particle's historical optimal position is given as $P_i = (p_{i_1}, p_{i_2}, \dots, p_{i_d})$. The historical optimal position for the swarm is given as $P_g = (p_{g_1}, p_{g_2}, \dots, p_{g_d})$. For any maximization problem, the updating formula for each particle's optimal position is presented in Equation (2) (Wang *et al*. 2018).

$$p_{i,t+1}^d = \begin{cases} x_{i,t+1}^d, & if\ f(X_{i,t+1}) > f(P_{i,t}) \\ p_{i,t}^d, & otherwise \end{cases} \quad (2)$$

The updating formulas of velocity and position are given in Equation (3) and Equation (4), respectively.

$$v_{i,t+1}^d = \omega * v_{i,t}^d + c_1 * rand * \left(p_{i,t}^d - x_{i,t}^d\right) + c_2 * rand * \left(p_{g,t}^d - x_{i,t}^d\right) \quad (3)$$

$$x_{i,t+1}^d = x_{i,t}^d + v_{i,t+1}^d \quad (4)$$

According to Equation (3), ω, c_1, and c_2 are inertia weight, invidual weight, and social weight, respectively. For each iteration, the updating of the position of each particle X_i is shown in Figure 2.

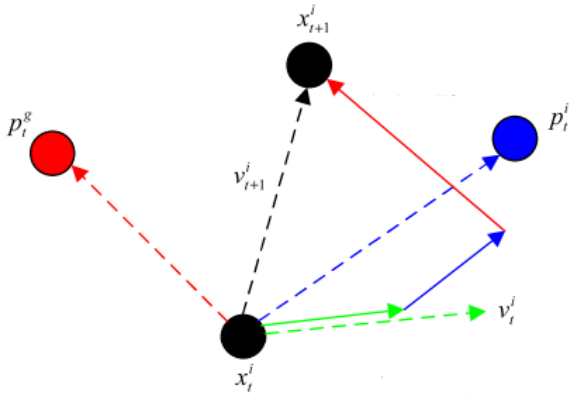

Figure 2. The position update of the particle (Wang *et al*. 2018).

3. The Proposed Method

In this section, the proposed method is explained in detail. To create features for the classification of the texts, Tf-Idf vectorization is used. Term frequency (Tf) indicates the number of occurrences for a word in the document. Tf value of the word t in a document d is computed as given in Equation (5) (Köksal and Akgül 2022).

Document frequency is acquired by using the instances of a word in corpus D. Idf (Inverse document frequency) expressed in Equation (6) is obtained through the inversion of this value. Consequently, Tf-Idf value is computed from Tf and Idf values as given in Equation (7).

$$Tf\ (t,d) = log\ (1 + freq(t,d)) \quad (5)$$

$$Idf\ (t,D) = \log\ (N\ /\ (count\ (d \in\ D: t \in\ D) + 1)) \quad (6)$$

$$Tf - Idf\ (t,D) = Tf\ (t,d) \times\ Idf\ (t,D) \quad (7)$$

The proposed TC consists of two phases, the wrapper feature selection and the categorization of the text data, respectively. The most general form of the proposed

method can be given as in Figure 3. According to Figure 3, high-dimensional text training data is processed with PSO-MNB (i.e., the wrapper feature selection) to eject redundant and indiscriminate Tf-Idf features from the data. After applying the wrapper feature selection, this reduced data with the best selected Tf-Idf features is utilized to classify news texts with MNB classifier effectively. MNB classifier is also used to compute the fitness values of the particles in the swarm. To compute the fitness value of each particle X_i in the swarm, the features selected by X_i are taken into consideration and the other features are not included for classification with MNB classifier. Classification is made for these selected features by using MNB with 2-fold cross-validation and the classification accuracy of MNB is assigned as the fitness value of X_i, $f(X_{i,t})$. Thanks to the proposed wrapper feature selection, the best discriminative feature subset with a higher classification accuracy is determined.

4. Experimental Results

4.1 Dataset

In the experiments, the publicly available TTC-3600 Turkish news dataset, which is used as a benchmark dataset for Turkish TC in previous studies in the literature, is used. TTC-3600 is an evenly distributed dataset with six news categories: culture, economy, health, politics, sports, and technology. The dataset includes 3600 texts of news in total. This dataset is publicly available and can be downloaded from (Int.Ref-1)

4.2 Experimental Setup

In this study, the well-known WEKA package is employed to implement the proposed wrapper feature selection and classification methods. The experiments are performed on a system equipped with a 2.6 GHz Intel Core i7-9750h CPU and 16 GB of RAM. The accuracy, f1 score, precision, and recall values are utilized as the classification performance metrics for the algorithms. The mathematical formulas for the accuracy, f1 score, precision, and recall are given in Equation (8), Equation (9), Equation (10), and Equation (11), respectively. 10-fold cross-validation is used to obtain the classification results in the experiments.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad (8)$$

$$F1\_score = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

In Equations (8-11), TP and TN stand for true positives and true negatives while FP and FN represent false positives and false negatives. For the proposed wrapper feature selection (i.e., PSO-MNB), the default WEKA data mining tool parameters (i.e., 0.33, 0.34, and 0.33 respectively) are used for inertia weight (ω), individual weight (c_1), and social weight (c_2) in PSO algorithm. Because the default parameters generally produce close to optimal results. 20–50 particles have been suggested for PSO in the literature. In the experiments, the best results have been achieved when the population size (N) is set to 50. The number of maximum iterations is set to 500 for PSO algorithm since there is no significant progress for the higher number of iterations.

In PSO-MNB, 2-fold cross-validation is used for the model evaluation to measure the qualities of the candidate feature subsets. A wrapper feature selection method (i.e., PSO-MNB) employs an optimization method (i.e., PSO) that searches for the candidate feature subsets. As a result, a pre-determined classifier (i.e., MNB) is built and evaluated on the dataset as many as the number of iterations of the optimization method. This process is time-consuming and computationally expensive. To shorten the runtime of the model evaluation during the determination of the qualities of feature subsets, the minimum number of cross-validation (i.e., 2-fold) is preferred in this study.

On the other hand, the default parameters in the WEKA data mining tool are used for all machine learning methods in the experiments.

4.3 Classification Results

From the observed results in Table 1, it is clear that the proposed wrapper feature selection based method (i.e., PSO-MNB) has the highest classification result among the algorithms both on stemming and non-stemming data in terms of TC. In addition, the best TC result (an accuracy of 94.55%) is obtained by the proposed method on stemming data. At the same time, the proposed method obtains f1 score of 94%, 94% of precision, and 94% of recall on stemming data. In addition, the proposed method achieves f1 score of 93.4%, 93.4% of precision, and 93.4% of recall on non-stemming data.

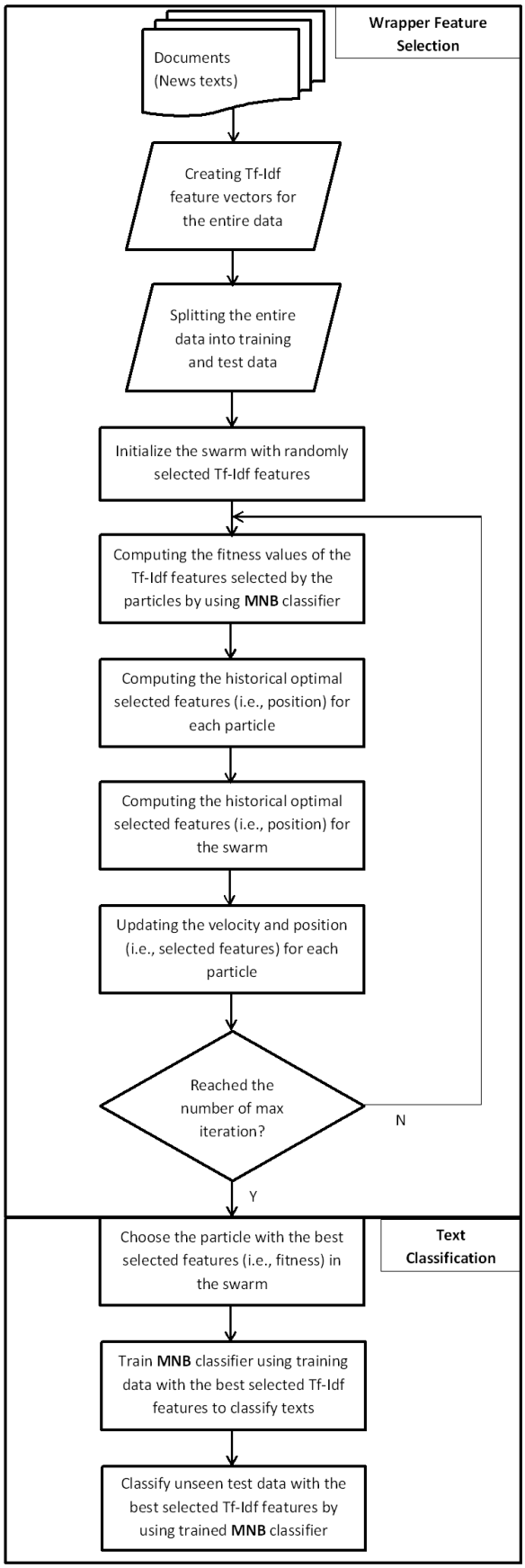

Figure 3. The basic steps of the proposed method.

The number of original stemming Tf-Idf features is 5692. However, the number of Tf-Idf features is reduced to 2949 for stemming data by applying PSO-MNB. On the other hand, the number of original non-stemming Tf-Idf features is 7507 while the number of Tf-Idf features is decreased to 3947 when PSO-MNB is applied to non-stemming data. Besides, the proposed method produces a classification accuracy of 93.44% for Turkish TC on non-stemming data. According to Table 1, MNB classifier has the highest classification result compared to the other classification algorithms for Turkish TC on non-stemming and stemming data following the proposed method. From this point of view, PSO-MNB wrapper-based feature selection method increases the accuracy performance of MNB classifier by 1.39% and 2.64% on non-stemming and stemming data, respectively. After the proposed method and MNB classifier, the best Turkish TC results (87.58% on non-stemming data and 89.25% on stemming data) are attained by RF. Consequently, the proposed method performs better than the widely used classification methods on non-stemming and stemming versions of TTC-3600 Turkish TC dataset.

4.4 Comparison with SOTA

In this section, a comparison of the proposed method with SOTA is given. Therefore, firstly, the comparison of the proposed method with the existing feature selection based Turkish TC is made and the results of the previous studies are delivered in Table 2.

According to Table 2, the best classification accuracy (99.69%) on TTC-3600 dataset for Turkish TC is obtained by IG-DPC (Zorarpaci 2023). This study utilizes IG to select the discriminative Tf-Idf features and then applies DPC which is semi-supervised machine learning algorithm by using the selected Tf-Idf features. Following IG-DPC, the second best classification result (accuracy of 94.55%) is accomplished by the proposed wrapper feature selection based method (i.e., PSO-MNB).

Kılınç *et. al* (2017) have used TTC-3600 dataset for Turkish TC. In this study, Tf-Idf feature weighting schema is proposed for feature extraction. The third classification accuracy (91.03%) is achieved by this method which is based on attribute ranking feature selector and RF classifier (Kılınç *et al*. 2017).

Heyong and Ming (2019) have tested their method on some datasets (i.e., CarF, CarR, CNAE, IMDB, and KDC) for English TC and TTC-3600 dataset for Turkish TC. They have used Tf-Idf method to extract the features. They have applied HRFS feature selection method to the datasets. MNB classifier is employed during the categorization process. They have reached a classification accuracy of 90% on TTC-3600 dataset in this study.

Parlak (2023) has implemented Tf-Idf feature extraction method for TC. In this study, BBC 20Newsgroups and TTC-3600 datasets are utilized for English and Turkish TC, respectively. In this study, EFS is used for feature selection and SVM is used for TC. According to this study, 78.1%

classification result is obtained by the proposed method on TTC-3600 dataset.

Table 2. Classification performances of the existing feature selection based Turkish TC methods on TTC-3600 dataset.

Reference	Feature Selection	Classifier	Accuracy (%)
(Kılınç *et al.* 2017)	Attribute ranking-based	RF	91.03%
(Heyong and Ming 2019)	HRFS	MNB	90%
(Parlak 2023)	EFS	SVM	78.1%
(Zorarpaci 2023)	IG	DPC	**99.69%**
The proposed method	PSO+MNB	MNB	94.55%

From the results in Table 2, it is clear that pre-processing based Turkish TC methods result in pretty good results and the proposed method generates competitive results to the existing pre-processing based Turkish TC studies.

Table 3. Classification performances of the cutting-edge Turkish TC methods on TTC-3600 dataset.

Reference	Method	Accuracy (%)
(Kılınç 2016)	C4.5+Boosting	85.52%
(Aci and Çirak 2019)	Word2Vec+CNN	93.30%
(Kuyumcu *et al*. 2019)	FastText	93.52%
(Dogru *et al*. 2021)	CNN	94.17%
(Köksal and Akgül 2022)	FastText+CNN	**95.97%**
(Yürekli 2023)	PV-DBOW	93.51%
The proposed method	PSO+MNB	94.55%

Secondly, the classification results of the cutting-edge Turkish TC methods are presented in Table 3. According to the results in Table 3, the highest accuracy (95.97%) on TTC-3600 dataset is achieved by FastText-CNN developed by Köksal and Akgül (2022). They have utilized FastText for feature extraction and CNN classifier for TC. After FastText-CNN, the second highest Turkish TC classification accuracy (94.55%) is accomplished by the proposed method (i.e., PSO-MNB).

Table 1. Classification performances of the methods on TTC-3600 dataset.

Data	Accuracy								
	NB	**SVM**	**K-NN**	**ET**	**C4.5**	**RF**	**BN**	**MNB**	**PSO-MNB**
Non-stemming	76.11%	85.83%	52.83 %	62.4%	78.05%	88.3%	83.88%	92.05%	**93.44%**
Stemming	76.72%	86.94%	54%	65.75%	79%	89%	87.88%	91.91%	**94.55%**
	F1-score								
	NB	**SVM**	**K-NN**	**ET**	**C4.5**	**RF**	**BN**	**MNB**	**PSO-MNB**
Non-stemming	76.1%	85.8%	52.4%	62.5%	78%	88.3%	83.9%	92%	**93.4%**
Stemming	76.7%	87%	54.4%	65.8%	78.9%	88.9%	87.9%	91.9%	**94%**
	Precision								
	NB	**SVM**	**K-NN**	**ET**	**C4.5**	**RF**	**BN**	**MNB**	**PSO-MNB**
Non-stemming	76.6%	85.9%	62%	62.6%	77.9%	88.5%	84.6%	92.1%	**93.4%**
Stemming	77.3%	87.2%	61.9%	65.9%	78.9%	89%	88.1%	92%	**94%**
	Recall								
	NB	**SVM**	**K-NN**	**ET**	**C4.5**	**RF**	**BN**	**MNB**	**PSO-MNB**
Non-stemming	76.1%	85.8%	52.8%	62.5%	78.1%	88.3%	83.9%	92.1%	**93.4%**
Stemming	76.7%	86.9%	54%	65.8%	79%	89%	87.9%	91.9%	**94%**

On the other hand, the difference between the classification accuracies of PSO-MNB and FastText-CNN is only 1.42% which is not a great difference. Following FastText-CNN and PSO-MNB, Turkish TC methods based on CNN show excellent classification performances (93.30%-94.17%) on TTC-3600 dataset. Dogru *et al*. (2021) have implemented convolution to extract the features and CNN to classify the news texts. They have used BBC 20Newsgroups and TTC-3600 datasets for English and Turkish TC, respectively. They have achieved a classification result of 94.17% on TTC-3600 dataset for Turkish TC. Kuyumcu *et al*. (2019) have proposed FastText that is word embedding based classifier for Turkish TC. They have used TTC-3600 dataset to evaluate FastText classifier. They have attained a classification result of 93.52%. Yürekli (2023) has proposed an ensemble Turkish TC approach based on document processing, paragraph vector learning, and document similarity estimation. The experiments have been conducted on TTC-3600 dataset and a classification result of 93.5% is obtained by this approach.

Aci and Çirak (2019) have developed a Turkish TC method based on CNN and Word2Vec. They have tested their

method by using TTC-3600 dataset and achieved a classification result of 93.3% on this dataset. When an overall assessment is made, it can be inferred from the results that the proposed wrapper feature selection based Turkish TC method (i.e., PSO-MNB) eliminates the irrelevant Tf-Idf features from TTC-3600 dataset and increases the classification performance of MNB classifier. Moreover, it produces competitive Turkish TC classification results to SOTA.

5. Conclusions

In this study, a Turkish TC method based on wrapper feature selection using PSO and MNB classifier is proposed. In the study, to create features for the classification of the news texts from TTC-3600 dataset, Tf-Idf vectorization is employed. The proposed TC method includes two steps, the wrapper feature selection and the classification of the texts using Tf-Id features, respectively. In the study, MNB classifier is utilized both for computing the quality of the feature subsets in PSO algorithm and categorizing the news texts. According to the experimental results, the proposed wrapper feature selection based Turkish TC method (i.e., PSO-MNB) generates competitive classification results to both the existing feature selection based Turkish TC methods and the cutting-edge Turkish TC methods.

In future work, it is planned to use the proposed wrapper feature selection with CNN classifier to improve the classification accuracy for Turkish TC.

Declaration of Ethical Standards
The authors declare that they comply with all ethical standards.

Credit Authorship Contribution Statement
Author: Research, Methodology, Experiment, Writing

Declaration of Competing Interest
The authors have no conflict of interest to declare regarding the content of this article.

Data Availability
All data generated or analyzed during this study are included in this published article.

6. References

Aci, Ç. And Çirak , A., 2019. Turkish news articles categorization using convolutional neural networks and Word2Vec. *Bilişim Teknolojileri Dergisi, **12(3)**, 219-228*. https://doi.org/10.17671/gazibtd.457917

Alqaraleh, S., 2021. Efficient Turkish text classification approach for crisis management systems. *Gazi University Journal of Science, **34(3)**, 718-731*. https://doi.org/10.35378/gujs.715296

Borandağ, E., Özçift, A. and Kaygusuz, Y., 2021. 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*, **29(2)**, 514-530. https://doi.org/10.3906/elk-1911-116

Dogru, H. B., Tilki, S., Jamil, A. and Hameed, A. A., 2021. *Deep learning-based classification of news texts using doc2vec model*. 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 91-96.

Ghareb, A.S., Bakar, A.A. and Hamdan, A.R., 2016. Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications*, **49**, 31-47. https://doi.org/10.1016/j.eswa.2015.12.004

Heyong, W. and Ming, H., 2019. Supervised Hebb rule based feature selection for text classification. *Information Processing and Management*, **56**, 167-191. https://doi.org/10.1016/j.ipm.2018.09.004

Kayakuş, M. and Açıkgöz, F. Y., 2022. Classification of news texts by categories using machine learning methods. *Alphanumeric Journal*, **10(2)**, 155-166. https://doi.org/10.17093/alphanumeric.1149753

Kennedy, J. and Eberhart, R., 1995. Particle swarm optimization. *In Proceedings of ICNN'95-international conference on neural networks*, **4**, 1942-1948.

Kılınç, D., 2016. The effect of ensemble learning models on Turkish text classification. *Celal Bayar University Journal of Science*, **12(2)**, 215-220. http://dx.doi.org/10.18466/cbujos.04526

Kılınç, D., Özçift, A., Bozyigit, F., Yıldırım, P., Yücalar, F. and Borandag, E., 2017. TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science*, **43(2)**, 174-185. https://doi.org/10.1177/0165551515620551

Kim, K. and Zzang, S. Y., 2019. Trigonometric comparison measure: A feature selection method for text categorization. *Data & Knowledge Engineering*, **119**, 1-21. https://doi.org/10.1016/j.datak.2018.10.003

Köksal, Ö., 2020. *Tuning the Turkish text classification process using supervised machine learning-based algorithms*. International Conference on Innovations in Intelligent Systems and Applications (INISTA), Novi Sad, Serbia, 1-7.

Köksal, Ö. and Yılmaz, E.H., 2022. Improving automated Turkish text classification with learning-based algorithms. *Concurrency and Computation: Practice and Experience*, **34(11)**, e6874. https://doi.org/10.1016/j.datak.2018.10.003

Köksal, Ö. and Akgül, Ö., 2022. *A comparative text classification study with deep learning-based algorithms*. 9th International Conference on Electrical and Electronics Engineering (ICEEE), Alanya, Turkey, 387-391.

Kuyumcu, B., Aksakalli, C. and Delil, S., 2019. *An automated new approach in fast text classification (fastText): A case study for Turkish text classification without pre-processing*. 3rd International Conference on Natural Language Processing and Information Retrieval, Tokushima, Japan, 1-4.

McCallum, A. and Nigam, K., 1998. A comparison of event models for naive bayes text classification. *In AAAI-98 workshop on learning for text categorization*, **752**, 41-48.

Meetei, L. S., Singh, T. D., Borgohain, S. K. and Bandyopadhyay, S., 2021. Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation*, **55(4)**, 947-969. https://doi.org/10.1007/s10579-021-09541-9

Parlak, B., 2023. The effects of preprocessing on Turkish and English news data. *Sakarya University Journal of Computer and Information Sciences*, **6(1)**, 59-66. https://doi.org/ 10.35377/saucis...1207742

Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., and Mehmood, A., 2023. Impact of convolutional neural network and FastText embedding on text classification. *Multimedia Tools and Applications*, **82(4)**, 5569-5585. https://doi.org/10.1007/s11042-022-13459-x

Wang, D., Tan, D. and Liu, L., 2018. Particle swarm optimization algorithm: an overview. *Soft Computing*, **22**, 387-408. https://doi.org/10.1007/s00500-016-2474-6

Xie, L., Liu, G. and Lian, H., 2019. Deep variational auto-encoder for text classification. *In 2019 IEEE International conference on industrial cyber physical systems (ICPS)*, 737-742.

Yürekli, A., 2023. On the effectiveness of paragraph vector models in document similarity estimatıon for Turkish news categorization. *Eskişehir Technical University Journal of Science and Technology A-Applied Sciences and Engineering*, **24(1)**, 23-34. https://doi.org/10.18038/estubtda.1175001

Zhang, J., Wu, H., Jiang, M., Liu, J., Li, S., Tang, Y. and Long, J., 2023. Group-preserving label-specific feature selection for multi-label learning. *Expert Systems with Applications*, **213**, 118861. https://doi.org/10.1016/j.eswa.2022.118861

Zorarpaci, E., 2023. *A Turkish text classification based feature selection and density peaks clustering*. 31st Signal Processing and Communications Applications Conference (SIU), Istanbul, Turkey, 1-4.

Internet References

1-https://archive.ics.uci.edu/dataset/407/ttc+3600+benchmark+dataset+for+turkish+text+categorization. (15.01.2024)