ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2018-63264

# Classification of Statistics Books Based on Author's Education by Using Text Mining Methods

## İstatistik Kitaplarının Metin Madenciliği Yöntemleri Kullanılarak Yazarlarının Eğitimine Göre Sınıflandırılması

Betül KAN KILINÇ,[a]
Yonca YAZIRLI[b]

[a]Eskisehir Technical University, Faculty of Science, Department of Statistics, Eskisehir, TURKEY
[b]Eskisehir Technical University, Graduate School of Science, Department of Statistics, Eskisehir, TURKEY

*Received:* 11.10.2018
*Received in revised form:* 09.11.2018
*Accepted:* 09.11.2018
*Available Online:* 15.11.2018

*Correspondence:*
Betül KAN KILINÇ
Eskisehir Technical University, Faculty of Science, Department of Statistics, Eskisehir, TURKEY
bkan@eskisehir.edu.tr

**ABSTRACT Objective:** Nowadays, the machine learning techniques and understanding of the text-based data issues have gained a lot of interest. Publicly available information in web may be considered to utilize the text mining tools for several and especially for classification problems. The data mining methods aim to achieve the desired information through large data stacks. **Material and Methods:** In this paper, we focus on the information on statistics books and the bachelor's degree of the authors which are obtained from web sources. This paper presents a comparison of accuracy of different learning models for classifying the author's bachelor's degree by using a created data containing the information of 146 Turkish books of statistics whose written by statisticians, mathematicians, econometricians, engineers, biostatisticians, etc. The dependent variable is classified as binary such that the authors having a bachelor's degree in statistics and the others. To reduce the dimensionality in classification, irrelevant and ineffective words, special characters are filtered out at pre-processing phase. **Results:** The 70% of the data set is determined for training and 30% is used for testing for classification. Three machine learning algorithms including k-nearest neighbor (k-NN), support vector machine (SVM) and random forest (RF) are trained using the created data and accuracy performance is obtained. **Conclusion:** Comparing the results, it can be said that the best performance in classifying the authors' bachelor's degree is obtained from RF.

**Keywords:** Text mining; classification; random forest; machine learning.

**ÖZET Amaç:** Günümüzde makine öğrenme teknikleri ve metne dayalı verilerinin incelenmesi büyük ilgi görmektedir. Pek çok ve özellikle sınıflandırma problemleri için internetteki kamuya açık bilgiler kullanılarak metin madenciliği araçlarından faydalanmak mümkündür. Metin madenciliğinin amacı büyük veri yığınlarından istenilen bilgiye ulaşabilmektir. **Gereç ve Yöntemler:** Bu çalışmada, internet üzerinden yapılan bir araştırma ile elde edilen istatistik alanında yazılmış kitapların künye bilgileri ve bu kitapların yazarlarının lisans mezuniyetleri üzerinde durulmuştur. Farklı öğrenme yöntemlerinin sınıflandırma başarılarının karşılaştırılmasında istatistik, matematik, ekonometri, mühendislik, biyoistatistik, vb. lisans derecelerine sahip yazarlar tarafından yazılmış 146 istatistik kitabının künye bilgilerini içeren bir veri seti oluşturulmuştur. Bağımlı değişken, istatistikçiler ve diğerleri olmak üzere ikili sınıfta değerlendirilmiştir. Sınıflandırmada boyut problemini azaltmak için, ön işleme aşamasında, verideki ilgisiz, etkisiz ve özel karakterler çıkarılmıştır. **Bulgular:** Sınıflandırma yapılmadan önce, veri setinin %70' i eğitim için, geri kalan %30' luk veri seti ise test için ayrılmıştır. Üç makine öğrenme algoritması, k-en yakın komşuluk (k-NN), destek vektör makinesi (SVM) ve rasgele orman (RF) kullanılarak, oluşturulan veriler eğitilmiş ve sınıflandırma başarıları ölçülmüştür. **Sonuç:** Öğrenme algoritmalarının performansları incelendiğinde, lisans mezuniyetlerine göre yazarları sınıflandırmada rasgele orman algoritmasının en iyi performansı gösterdiği ortaya çıkmıştır.

**Anahtar Kelimeler:** Metin madenciliği; sınıflandırma; rastgele orman; makine öğrenmesi

One of the main advantages of technology is to provide us an easy access to the digital storage of data and hence saving time and space. Consequently, this situation has led to the formation of large data stacks and the analysis has become increasingly difficult. As a result, data mining has emerged as an important field of study.

Nowadays, many institutions and organizations have begun to attach importance to data mining by understanding the importance of data collection and the fact that only inquiry-based information will be obtained with past database inquiries.

The data mining methods aim to achieve the desired information through large data stacks. While classical statistical applications are organized and worked on summary data, the data mining methods deal with data stacks containing millions or even more variables.

While data mining is concerned with numerical data, text mining method has been developed for the analysis of data in text format. Various data sources have been discussed in the literature and the applicability of the newly developed text mining in different areas has been revealed by many scientific studies.

Gao and Eldin[1] used text mining methods to analyze employment data sent over the internet in their work. The aim of their work is to identify areas of knowledge, competencies and expertise related to the work in the construction sector. Between 14 November 2012 and 15 March 2013, over 20,000 job advertisement were downloaded from various websites. Once the qualifications were identified, the Latent Dirichlet Allocation model was used to identify groups the skills set that were required by employers. As a result, the qualifications required for the accountant, planner, supervisor, project engineer and project manager were determined. Bastin and Bouchet-Valat.[2] introduced R. TEMIS, a free software solution aimed at discovering new dimensions in text mining with special focus on media framing analysis in their work. The proposed module was obtained through R coding (one and two-way tables, time series, hierarchical clustering, response analysis, geographical mapping etc.) for social scientists investigating automation and texts of management procedures. In addition, it is designed to provide assistance in the context of expanding the scope of statistical tools. Luther et al.[3] aimed at revealing whether the statistical text mining could determine the fall-related injuries in electronic health record (EHR) documents in their work. In addition, they were intended to demonstrate the effectiveness of the STM training models obtained from one or more certified documents. Kehagias et al.[4] compared the performance of classifiers with the algorithms of different classifiers based on the words of classifiers based on sentences. Compared document collections were assumed as a subset of the explicit Brown Corpus semantic alignment. The study revealed that the use of senses did not result in any significant categorization development. Akar and Güngör [5] aim to examine the performance of RF algorithm using multispectral satellite images having different spatial resolutions and scene characteristics. To evaluate the performance of RF, the classification results were compared with the results obtained from Gentle AdaBoost (GAB), SVM and Maximum Likelihood Classification (MLC) algorithms. Ishikiriyama et al.[6] have addressed the breadth of business intelligence and their importance for researchers in their work. Their study aims to present a small sample of what is possible to achieve by analyzing text data from academic papers. The methodology occurred analyzing a sample of the top 35 most relevant papers regarding Business Intelligence obtained through an academic search engine and offered the results of this text mining study. Kılınç et al.[7] summarized certain academic publications on Research Gate, and using text mining methods, the articles were divided into classes and the success of classification was measured by k-NN. Bai[8] and Li et al.[9] have performed classification operations in their work using the sentiment analysis approach of text mining. The classification achievements of the algorithms used are 92.7% and 90.40%, respectively. Rasjid and Setiawan1[0] used k-NN and Naive Bayes approaches to study unstructured data and summary information for classification

and clustering in their study in 2017. The performance of the methods was compared in the study and the number of clusters in k-NN method reached the optimum result at k = 13. Chen et al.[11] mentioned that in text classification, the term weighting was a basic problem and directly affected the success of the classification. In Akşehirli et al. (12)' s study a supervised SVM used for classification or regression was performed in medical data set. Sun et al. (13) presented a comparative study on the strategies addressing imbalanced text classification by using SVM classifiers. Experimental results revealed the standard SVM often learn the best decision surface in most test cases.

In this paper, text mining algorithms and word cloud functions are used to explore the unstructured forms of books and authors' information obtained from the web sources. Using the structured form of the data set, the performance of three machine learning algorithms including k-NN, SVM, and RF are compared.

## MATERIAL AND METHODS

### TEXT MINING

The general definition of text mining is defined as the extraction of qualified information in raw data obtained from a text that is not in a proper format. To analyze and process semi-structured and unstructured (raw) text data, it is necessary to convert the text into a vector space. Hence, analytical algorithms can be used to large document databases.[14]

Text mining can be divided into seven areas of application based on their specific characteristics of each field.[14] These subfields are given as follows: search and information retrieval-IR, document clustering, document classification, web mining, information extraction-IE, natural language processing-NLP, concept extraction.

It is not easy for computer software to analyze the unstructured (raw) data. That is, the data should be exchanged to a form which is understandable by the computer software to be used in text mining applications. For this reason, the texts have to go through a pre-processing step. Afterwards, the edited data is transferred to the computer and the texts are converted into numbers and the processing steps are performed.[15]

The pre-processing step is to promote the data to the computer with several operations consisting of punctuation, extraction of characters such as spaces, special characters for some languages, conversion to lower case, extraction of ineffective words (with, and, or, all, some, etc.) from the main text.

A vector space (VS) model is used to transform each text into a numerical value in order to generate the complete document matrix.[16] If a term is included in the document, the value in the vector changes according to the frequency of its occurrences. In short, a value in a vector expresses the frequency of this term lying in the text.

It is important to weigh each word in the documents for constructing the VS model. Vector weighting affects the success of classification. Three most commonly used methods are Term Frequency (TF), Inverse Document Frequency (IDF) and TF-IDF method for weighting.

#### Term Frequency (TF) Method

In this method, the number of times a word appears in the text is measured and more words in the text are thought to be more important. The calculation of the TF weighting method is as follows:[16]

$$TF_{ij} = \frac{n_{ij}}{d_i} \quad (1)$$

where $d_i$ is ith document and $n_{ij}$ is the frequency of passage of the $w_j$ word in $d_i$.

### Inverse term frequency (IDF) method

In this method, fewer words in the text are thought to be distinctive.[16]

$$IDF_j = \log\frac{n}{n_j} \tag{2}$$

where $n_j$ is the number of documents contains the $w_j$ word and n is the number of documents in a set of documents.

### TF-IDF method

This method combined TF and IDF methods to weigh the terms.[16]

$$x_{ij} = TF_{ij} \times IDF_j \tag{3}$$

where $x_{ij}$ is the TF-IDF weight of the $w_j$word in $d_i$ document.

After the weighting process, the document term matrix is generated as shown in Figure 1.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_j \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & \cdots & t_j \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2j} \\ & & & & \\ & & & & \\ x_{j1} & x_{j2} & x_{j3} & \cdots & x_{jj} \end{pmatrix} \end{array}$$

**FIGURE 1:** Document term matrix.

### Evaluation Criteria

The confusion matrix is a $n \times n$ matrix of size which is used to represent the actual and predicted values of classes.

**TABLE 1:** Confusion Matrix

| | | Estimated Class | |
|---|---|---|---|
| | | A (+) | B (-) |
| Real Class | A (+) | TP | FN |
| | B (-) | FP | TN |

The confusion matrix shown in *Table 1*;

TP, "True Positive" that the element in class A is assigned to class A; FN, "False Negative" that the element in class A is assigned to class B; FP, "False Positive" the element that cannot be in class B is assigned to class A; TN, "True Negative" the element that cannot be in class B is not assigned to class B.

The success of the algorithm is calculated by the accuracy rate, which indicates the correct classification. Misclassification measure is calculated as 1-Accuracy.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

## k-NN ALGORITHMS

The k-NN algorithm is a classification algorithm proposed by Cover and Hart.[17] The nearest neighbor decision rule assigns an unclassified sample point to the nearest of a set of previously classified points.

The main purpose of the k-NN algorithm is to determine classes of unknown sample points based on its (sample point's) nearest neighbor whose class if already known. *k* is known as the number of neighbors being compared, so that it must be an integer. After the training set constructed, distance (Euclid, Manhattan, Murkowski, etc.) is evaluated from all training points. The Euclidean distance given in *Equation 5* is preferred because it is the most commonly used distance criterion in classification and clustering algorithms.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2} \quad (5)$$

where p refers to number of variables, $i = 1,2, \dots, n$ and $j = 1,2, \dots, n$.

The processing steps of the k-NN algorithm are as follows:

*Step 1:* First, k is determined.

*Step 2:* The distances between each sample query and the samples in the whole training set are calculated.

*Step 3:* Distances are sorted from small to large.

*Step 4:* k. neighbors up to the distance are selected after their classes are decided.

The k-NN algorithm is especially preferred for classification applications due to its lack of training, its easy implementation, its adaptability to any local knowledge, its resistance to noisy training data.[18]

## SUPPORT VECTOR MACHINES

SVM is a learning algorithm based on the statistical learning theory.[19] This method is mainly thought of as a linear classifier for solving two class problems. Later, it was started to be widely used in solving these problems by generalizing classifying problems which cannot be separated linearly or multi class.

The principle of SVM is based on defining a hyperplane which can distinguish the two classes from each other in an optimal way. While it is possible to distinguish linearly separable data from a plane of the size to which they belong, it is possible to separate the nonlinearly separable data by a hyperplane by moving it to a higher dimension than its size.

When binary classification is considered for a data set that can be linearly separated, there are infinite numbers of hyperplanes that can distinguish it. While SVM creates the decision surface, it tries to maximize the distance between the two classes. There is only one hyperplane with a maximum boundary between these planes. The hyper-planar optimal hyperplane, which subtracts the maximum to the maximum, is called the optimal hyperplane. The points that limit the border width are called "support vectors". The support vector algorithm tries to minimize the training error while classifying with the separator hyperplane with the largest boundary width. This method aims to have a linear discriminant function with the largest marginal difference between the classes. After the decision function is determined, the class

to which the new instance belongs is determined according to the value it receives in the function. If the dataset cannot be separated linearly, the data are transported to another higher dimension where they can be linearly separated, and the classification is done in that space.

### RANDOM FOREST

One of the data mining tools is known as RF algorithm. It is similar to classification trees whereas a tree is constructed based on the splitting rules. Indeed, a large number of trees are constructed in RF algorithm using a randomized number of predictor variables. At each node of the tree, the predictors are selected to find the best split. After reaching the entire trees (default is maximum size of 500), predictions from all trees are combined. Advantage of using RF is that it produces a variable importance plot obtained from predictor variables.[20]

## APPLICATION

### OBTAINING DATA

A data set including book titles, name of the author(s), publishing year, publisher obtained from several online libraries have been created for 146 records. Besides the collection of the information of statistics books, the bachelor's degree of each author of the books is examined.

When a book has more than one edition, the latest edition is taken into consideration. The degree(s) of the authors who wrote the books have been obtained from web. The results show that authors who wrote statistics books are from different subjects such as mathematics, statistics, management, economy, engineering, biology, chemistry, etc. Hence the dependent variable is classified as binary such that the authors having a bachelor's degree in statistics and the others. In this sense, a value of 1 is assigned for a bachelor's degree in statistics, and 0 is assigned for other degrees.

### EXPERIMENTAL STUDY

In this study statistics books written in Turkish are examined so ineffective words such as "and, some," and "-sel, -ler, -li" are extracted manually. Also, some of the Turkish letters such as "ç, ş, ğ, ü, ö " have been converted into "c, s, g, u, o". Capital "i" is converted to small case "i" and the small "ı" is converted to small case "i". All computations are implemented by using R Studio.[21]

In the data pre-processing phase, first the tm_map function from library(tm) is used to convert to each term to lower case and unnecessary punctuations in the text are removed. After these steps, the document term matrix (DTM) in which each document is represented as a vector is constructed according to weighting process given in Section 2.1. The elements of the vector form the terms for each document. In DTM, there are 146 rows and 570 columns. The 70% of the data set is determined for training and 30% is used for testing for classification.

Additionally, the frequencies of the most 20 observed words are given by a bar chart in Figure 2 whereas the word cloud of the DTM are presented in Figure 3.

The created data set contains the title of statistics books in Turkish, the names of the authors, years, publishers, years of publication and the information on the bachelor's degree of the authors. In short, "listatistik" is used for a bachelor's degree in statistics, "yistatistik" is used for a master's degree in statistics, and "distatistik" for PhD in statistics. The term "999" represents the authors whose degree of PhD could not been accessed. The letter "y" is added to the end of Publisher's names. In Figure 2, a bar chart of the 20 most occurred words are plotted.

As seen in Figure 2, the most common word observed in book' s title is "istatistik" whereas "yontem" occurred 17 and "uygulamalı" occured 15 times. The word "distatistik" occurred 61 times meaning that the authors of the books have a PhD degree in statistics whereas the word "yistatistik" occurred 49 indicating the authors have a master's degree in statistics. The word "listatistik" indicates only 38 of the authors have a bachelor's degree in statistics whereas the number of the authors having a mathematics degree is only 30. On the other hand, the publisher "Nobel Akademi" published 21 of these books and 19 of the books published in 2016.

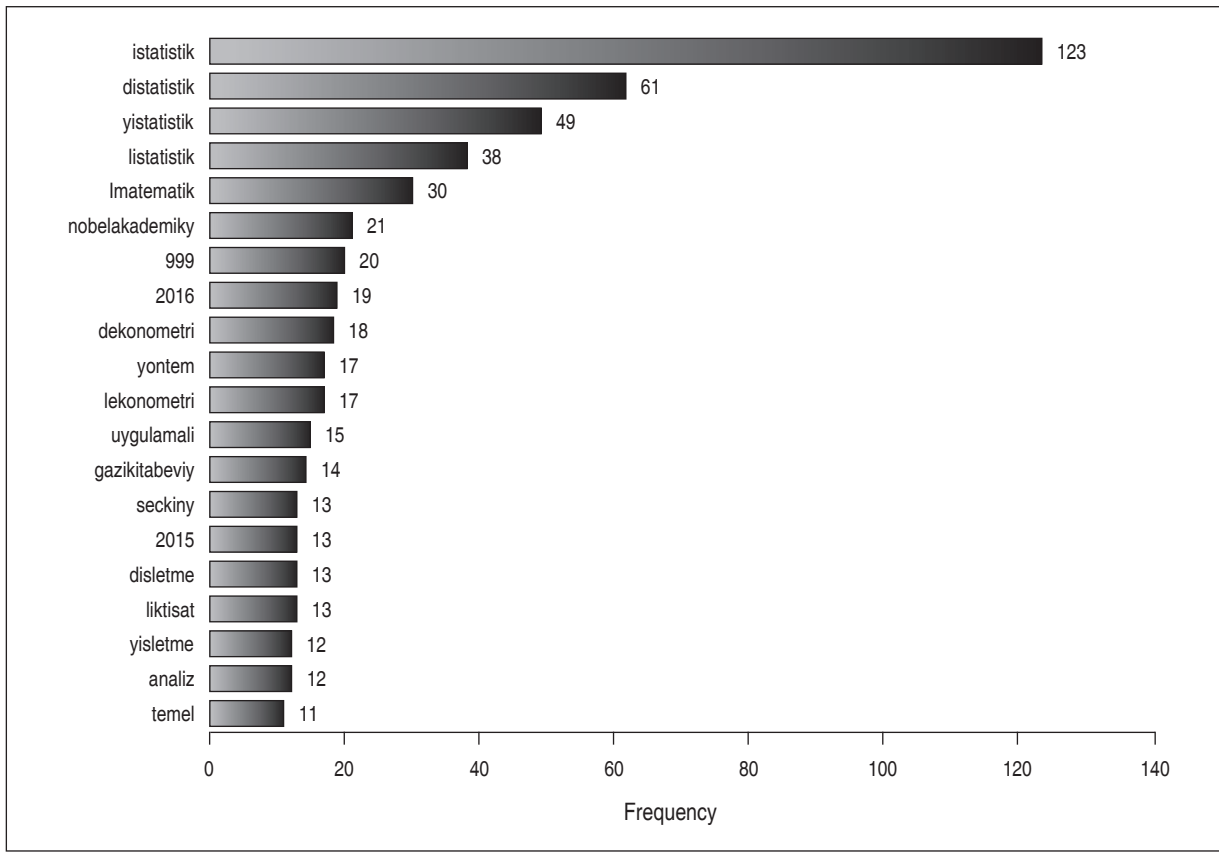


**FIGURE 2:** Word frequencies related to book kits.


**FIGURE 3:** Word cloud of the document term matrix.

Footnote: Words which have less than 5 frequencies are extracted.

In Figure 3, an interesting representation of the document term matrix is given by a word cloud which can be used to visualize the unstructured data. As it can be seen in figure, frequently occurred words stand out with larger characters. The figure indicates the word "istatistik" has the greatest frequency in the data as the larger the word in the figure the more it occurred the in the data. The least frequent words are not shown.

Moreover, it can be easily claimed that most of the authors preferred to use "istatistik" in the book titles with a ratio of 84% whereas some of the them preferred to use "yontem" with 12% and the others preferred "uygulamalı" with 10%. 42% of the authors achieved a doctorate degree, 34% achieved a master's degree and 26% had a bachelor's degree in the field of statistics (14% of authors had no available information on having a PhD degree). 21% of authors had a bachelor's degree in mathematics.

The results for author's education classification achieved an accuracy value 81.40%, 65.12% and 88.37% for k-NN, SVM and RF, respectively. According to these results, random forest is much better a classifier than k-NN and SVM for future selection.

This study shows not only a list of 146 records of statistics books which written by the authors having a degree in statistics but also indicates that approximately 74% of the authors have a bachelor's degree different from statistics.

## DISCUSSION AND CONCLUSION

This study presents an application of the performance comparison in classification problems by using text mining methods and learning algorithms. A total of 146 books of statistics and the regarding information based on the author's bachelor's degree are used for comparisons of three machine learning algorithms. Specifically, the bachelor's degree of the authors is examined for each book and categorized in two classes such as "statistics-based authors" and the "others". The created dataset is first examined by text mining tools due to its lack of formatting. Then, k-NN, SVM and RF algorithms are used to classify the authors' bachelor's degree. Experimental results show that text mining is very useful to gain effective information from an unstructured data and shows that statistics books are written by statisticians as well as other scientists from different field of areas. Regarding to the performance of the comparison, random forest classified effectively the class of authors' education more than the others.

For future studies, the learning methods can be applied to larger data sets.

***Source of Finance***

*During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.*

***Conflict of Interest***

*No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.*

***Authorship Contributions***

***Idea/Concept:*** *Betül KAN KILINÇ;* ***Design:*** *Betül KAN KILINÇ;* ***Control/Supervision:*** *Betül Kan Kılınç;* ***Data Collection and/or Processing:*** *Betül KAN KILINÇ, Yonca YAZIRLI;* ***Analysis and/or Interpretation:*** *Betül KAN KILINÇ;* ***Literature Review:*** *Yonca YAZIRLI;* ***Writing The Article:*** *Betül KAN KILINÇ, Yonca YAZIRLI;* ***Critical Review:*** *Betül KAN KILINÇ;* ***References and Fundings:*** *Betül KAN KILINÇ*

## REFERENCES


1. Gao L, Eldin N. Employers' expectations: a probabilistic text mining model. Procedia Eng 2014;85:175-82.
2. Bastin G, Bouchet-Valat M. Media corpora, text mining, and the sociological ımagination- a free software text mining approach to the framing of julian assange by three news agencies using R.TeMiS. Bulletin of Sociological Methodology 2014;122:5-25.
3. Luther SL, McCart JA, Berndt DJ, Hahm B, Finch D, Jarman J, et al. Improving identification of fall-related injuries in ambulatory care using statistical text mining. Am J Public Health 2015;105(6):1168-73.
4. Kehagias A, Petridis V, Kaburlasos VG, Fragkou P. A comparison of word-and sense-based text categorization using several classification algorithms. J Intell Inf Syst 2003;21:227-47.
5. Akar Ö, Güngör O. Classification of multispectral images using random forest algorithm. Journal of Geodesy and Geoinformation 2012;1(2):105-12.
6. Ishikiriyama CS, Miro D, Gomes CFS. Text mining business intelligence: a small sample of what words can say. Procedia Comput Sci 2015;55:261-7.
7. Kılınç D, Borandağ E, Yücalar F, Tunalı V, Şimşek M, Özçift A. [Classification of scientific articles using text mining with KNN algorithm and R language]. Marmara Fen Bilimleri Dergisi 2016;3:89-94.
8. Bai X. Predicting consumer sentiments from online text. Decision Support Systems 2011;50(4):732-42.
9. Li YM, Li TY. Deriving market intelligence from microblogs. Decision Support Systems 2013;55:206-17.
10. Rasjid ZE, Setiawan R. Performance comparison and optimization of text document classification using k-NN and naïve bayes classification techniques. Procedia Computer Science 2017;116(C):107-12.
11. Chen K, Zhang Z, Long J, Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst Appl 2016;66(C):1339-51.
12. Yılmaz Akşehirli Ö, Ankaralı H, Aydın D, Saraçlı Ö. [An alternative approach in medical diagnosis: support vector machines]. Turkiye Klinikleri J Biostat 2013;5(1):19-28.
13. Sun A, Lim EP, Liu Y. On strategies for imbalanced text classification using SVM: a comparative study. Decision Support Systems 2009;48:191-201.
14. Miner GD, Delen D, Elder J, Fast A, Hill T, Nisbet RA. The seven practice areas of text analytics. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 1st ed. Massachusetts, USA: Academic Press; 2012. p.29-41.
15. Turban E, Sharda R, Delen D. Decision Support and Bussiness Intelligence Systems. 9th ed. New Jersey, USA: Prenctice Hall; 2011. p.696.
16. Vishnu MG, Vardhan DBV, Sarangam K, Reddy P, Pal V. A comparative study on term weighting methods for automated telugu text categorization with effective classifiers. IJDKP 2013;3(6):95-105.
17. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 1967;13(1):21-7.
18. Bhatia N, Vandana. Survey of nearest neighbor techniques. IJCSIS 2010;8(2):302-5.
19. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121-67.
20. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Standford, California:Springer; 2008. p.745.
21. R Development Core Team. R: a language and environment for statistical computing, Vienna, Austria. R Foundation for Statistical Computing; 2013. http://www.R-project.org.