SCIENTIFIC OASIS

Spectrum of Operational Research

Journal homepage: www.sor-journal.org
ISSN: 3042-1470

GA-LDA Approach for Topic Modelling in Turkish Accounting and Finance Articles: Performance Optimization in Text Classification

Mehmet Ozcalici[1,*], Meltem Kilic[2]

[1] Department of International Trade and Logisctics, Faculty of Economics and Administrative Sciences, Kilis 7 Aralik University, Kilis, Turkey
[2] Department of International Trade and Logisctics, Faculty of Economics and Administrative Sciences, Kahramanmaras Sutcu Imam University, Kahramanmaras, Turkey

ARTICLE INFO

Article history:
Received 10 September 2024
Received in revised form 19 December 2024
Accepted 7 February 2025
Available online 21 February 2025

Keywords:
Topic modeling; Text Mining; Latent Dirichlet Allocation; Genetic Algorithm.

ABSTRACT

The volume of research in the social sciences is expanding rapidly, creating significant challenges in extracting meaningful insights from unstructured text, particularly from articles lacking a classification system. Analysing these high-volume texts offers numerous advantages, including the ability to automatically identify topic relevance and track thematic trends over time. Such insights are valuable for journal management and enable researchers to access detailed information about evolving areas of study. Latent Dirichlet Allocation (LDA) is a widely used method for topic modelling, effectively extracting topics from textual data. However, its performance can be further enhanced through optimization techniques such as Genetic Algorithms (GA). This study introduces an intelligent GA-LDA framework designed to optimize word subsets for LDA, thereby improving its predictive capabilities. The proposed system is applied to a dataset of 928 abstracts from a Turkish-language academic journal specializing in accounting and finance, covering publications from 2005 to 2020. The genetic algorithm selects optimal word subsets for LDA analysis, with perplexity scores serving as the fitness function to guide the optimization process. Experimental results demonstrate that the GA-enhanced LDA significantly improves classification accuracy and topic modelling performance. This study not only underscores the potential of GA-LDA in handling unstructured text but also provides a robust tool for advancing automated content analysis in Turkish academic literature.

1. Introduction

The number of studies working for social sciences is increasing each day. Extracting meaningful data from these unstructured articles (i.e. text files without a classification system) is a great challenge. The necessity to automatically classify and label documents collected in digital platforms emerges. Topic modeling is one of the most powerful techniques in text mining for data mining, latent data discovery, and finding relationships among data and text documents [1]. Topic modeling identifies and classifies latent topics of each document [2]. There are various methods for topic

* *Corresponding author.*
E-mail address: mozcalici@gmail.com

https://doi.org/10.31181/sor21202521

modelling; Latent Dirichlet Allocation (LDA) is one of the most popular one in this field. LDA can be used to find related topics within unstructured text [3]. LDA is used for various subjects such as diabetic complication prediction [4], cardiology record label classification [5], modeling healthcare data [6], analyzing road safety inspections [7], language identification [8]. Improving the performance of LDA is also studied by various researchers as explained in the next section. The performance of the analysis may vary depending on the words used. Some of the words may be redundant when using with LDA analysis. It is possible to improve the performance of LDA by eliminating unnecessary words. In this study genetic algorithm is employed to select the word subset for LDA in a way that will increase the prediction power of LDA. Turkish is one of the languages that is used in natural language processing studies. Turkish language was subject of document classification [9–11] and sentiment analysis [12–14]. Moreover, topic extraction form Turkish language documents using LDA has been also studied [15–17].

In this study an intelligent system is developed to automatically classify the academic papers written in Turkish language. Contribution of the study can be summarized as below:

i. All of the articles published in the Journal of Accounting and Finance which publishes academic articles in Turkish language are investigated. Journal is one of the leading academic journals in the fields of accounting and finance in Turkey.
ii. A system based on GA and LDA to automatically label the articles are developed. In proposed system GA selects the best word subset to improve the prediction power of LDA.
iii. Classification results of proposed method are compared with the opinions of an expert.
iv. Proposed system is evaluated with whole dataset and results are visualized.

The remaining of this paper is organized as follow. Section 2 reviews the related literature. Section 3 describes Genetic Algorithm and LDA techniques. Section 4 presents the proposed methodology and describes the dataset as well as results of the analysis. Finally, Section 5 concludes this paper.

2. Literature Review

There were various studies about enhancing/improving the performance of LDA. For example, Shams and Dastjerdi [18] developed enhanced LDA for aspect extraction. They integrated LDA and co-occurance analysis to increase the performance of LDA. They applied their proposed method to texts written in Persian and English. They conclude that their method can be applied to all languages with reasonable accuracy. Yeh *et al.,* [19] developed Conceptual Dynamic Latent Dirichlet Allocation (CDLDA) for topic detection. Their model considers temporal features by introducing dynamic concepts. Guo *et al.,* [20] addresses the problem that traditional LDA ignores some of the semantic features hidden inside the medium and long texts. They proposed to refine the document into different semantic topic units before applying LDA. They report that their proposed method outperforms traditional LDA.

In literature, many studies have been conducted in different disciplines related to Latent Dirichlet Allocation (LDA) modeling, which is one of the subject modeling methods. Some of these studies are summarized as follows in chronological order.

In their study, Lin *et al.,* [21] extracted and processed abstract data from the Society of Neuroscience (SFN) annual meeting abstracts during the period 2001-2006. They employed Latent Dirichlet Allocation (LDA) method to extract topics from this data. Data cleaning and disambiguation methods are constructed a unified database. Using natural language processing, text mining, and other data analysis techniques, they examined the demographics of scientific collaboration network, the Dynamics of the field over time and major research trends. They report that the results of their

work can be useful for scientists, policy makers, and funding agencies seeking to gain a complete and unbiased picture of the community structure.

Lienou and Maitre [22] interested in the annotation of large satellite images, using semantic concepts defined by the user. They first annotated the images and then LDA model is used. After training process, the model is used to annotate the unseen images. Çelikyılmaz *et al.,* [23] investigated the application of the LDA model to the question-answer system. They calculated the similarity between the questions asked by the users and the answers given by the respondents, and ranked with LDA. In their studies, Ekinci and Omurca [24] tested the Latent Dirichlet Allocation model, which is one of the subject modeling methods, to extract product features through hotel-related user reviews. There were 1000 reviews written in Turkish language. From the experimental results, it was determined that a successful feature extraction was made with the modeling method of LDA. Pavlinek and Podgorelec [25] developed a self-training LDA (ST-LDA) method to overcome the problems encountered when only a small set of labeled documents is available. They used 11 small initial labeled sets to assess the performance of proposed model. They report that ST-LDA method performed significantly better in terms of classification accuracy.

Atıcı *et al.,* [26] determined product aspects in customer complaints by using LDA. There were 9378 complaints belongs to 6 firms. The aim of their study was to determine complaints and dissatisfactions about products, service or companies from the complaints in a website dedicated to complaints. Li *et al.,* [27] used the LDA model to digitize and visualize the Financial Stability Report. Unlike digital data, text data expresses more information and intuitive senses. In their study, the Chinese Financial Stability Report was analyzed by LDA modeling to measure financial stability. As a result of LDA modeling, the basic terms and specific terms of each subject can be drawn with each field in finance. They analyzed 5-year or annual issues, thus revealing a design matrix and analyzing financial stability trends. At the end of the study, iy was easily depicted using teh macro environment word cloud in finance.

Onan [15] measured the predictive success of machine learning classifiers in sentiment analysis by testing them with sentiment analysis based on Latent Dirichlet Allocation in Turkish twitter messages. Author used five different learning algorithms in the study. As a result of the tests, it is determined that the LDA method is a suitable, effective and concise method for classifying Turkish text documents. Drosatos *et al.,* [28] presented an analysis of eHealth topics and trends in published literature indexed in PubMed, based on unsupervised topics modeling and trend analysis. Their findings indicate a slightly declining publication trend when compared to the overall PubMed corpus growth. Guven *et al.,* [16] used LDA to classify the emotions in tweets. They employed normal LDA and n-stage LDA models. There are 4000 tweets written in Turkish language in the dataset. There are five emotions. These emotions are; happiness, fear, surprise, anger and sadness. Classification results are presented.

Hagen [29], evaluated e-petitions, a popular tool used for political activities, with content analysis LDA models. Author stated that since e-petitions are based on unsupervised machine learning, a reliable training and verification process is required. In this study, the subject modeling algorithm LDA and e-petition data were used and a framework for verification is explained. With rigorous training and evaluation 87% of LDA-generated topics made sense to human judges. Bastani *et al.,* [30] investigated the complaints received by The Customer Financial Protection Bureau (CFPB). To authors, since the number of complaints is increasing over time, it is not practical to review the documents manually. They developed an intelligent system to analyze narratives automatically and provide insightful knowledge to the experts.

Bailon-Elvira *et al.,* [31], used LDA for topic modeling in the official Bulletin of the Spanish government (BOE). The analyzed the data in the suggestion system in the Official Bulletin of the State of Spain with LDA. All meta-data of the documents published in the proposal system were analyzed to know the scope ıof the system. As a result of the analysis, they found that more than 89% of the documents cannot be recommended, because they are not well described at the documentary level, some of their key meta-data are empty. Therefore, it proposes a method that automatically labels document based on LDA. As a result of this test, they found that using this approach, more than twice the documents currently made by the system can be proposed. Gangadhara and Gupta [32] define the names in agricultural documents using LDA-based subject modeling technique. Plant names, soil types, pathogen names, plant diseases and fertilizers have been identified in the field of agriculture. They tested the model using 3000 sentences. The result of the analysis was evaluated by the authorized people and 80% accuracy was provided. Chang *et al.,* [33] applied NLP to analyze environmental education research topics from 2011–2020, using text mining, cluster analysis, LDA, and co-word analysis on Web of Science abstracts. Expert reviews and TF-IDF-based cluster techniques identified seven categories, and comparisons between K-means and LDA indicated largely consistent topic groupings. These results underscore the usefulness of AI-driven methods for refining and interpreting environmental education research trends.

Sharma *et al.,* [34] presented a comprehensive review of 8,320 Scopus-indexed publications on smart cities from 2010 to 2022, analyzed using Latent Dirichlet Allocation. It identifies key research trends, highlights notable international collaborations among institutions and authors, and concludes with the need for deeper investigation into rapidly growing areas of IoT-driven smart city research. Modzik *et al.,* [35] presents a large-scale bibliometric literature review of 116,759 supply chain research documents (1969–2021) to identify prevalent themes and trends, including the impact of COVID-19. Thirteen distinct research domains emerge—ranging from ecology and IT support to inventory management and disruptive risk—offering a comprehensive overview of how supply chain scholarship has evolved over the past four decades. Park *et al.,* [36] identifies research trends on the metaverse by analyzing 451 publications through Latent Dirichlet Allocation, revealing six main topics (e.g., virtual and real-world engagement, crypto marketplaces, game-mediated communities). The findings indicate a notable increase in metaverse-related studies since 2007 and suggest future directions for research in virtual education, commerce, and identity representation. Shashank and Behera [37] examines the factors that influence product recommendations in women's e-commerce clothing by analyzing a diverse set of online reviews using Latent Dirichlet Allocation and natural language processing techniques. Findings highlight that product quality, consumer satisfaction, and the overall shopping experience play crucial roles in shaping positive recommendations, while clothing categories and review lengths further impact recommendation likelihood.

The reviewed studies demonstrate that Latent Dirichlet Allocation (LDA) and its various extensions or adaptations (e.g., enhanced LDA, Conceptual Dynamic LDA, self-training LDA) have been successfully employed in multiple domains—including aspect extraction from reviews, classification of tweets, topic detection in research abstracts, e-petitions, and large-scale bibliometric analyses of supply chain and smart city research. Researchers have integrated LDA with methods such as co-occurrence analysis, dynamic concepts, and document segmentation to improve its accuracy and applicability. LDA's utility extends to identifying hidden topics, extracting salient features, automating annotation tasks, and providing actionable insights on trends in datasets (e.g., customer complaints, financial stability reports, eHealth publications). Findings consistently indicate that LDA-based approaches can outperform traditional text classification techniques, particularly when dealing with large or diverse text corpora, even when only a small labeled set is available.

Collectively, these studies underscore LDA's versatility for unearthing latent themes in large-scale textual data and improving classification, annotation, and trend analysis tasks. By automating the detection of hidden structures in text, LDA enables organizations, policymakers, and researchers to make data-driven decisions, optimize resource allocation, and gain a deeper understanding of user-generated content and emerging trends. Methodological enhancements to LDA (e.g., incorporating domain-specific features, adopting semi-supervised training, and combining LDA with other natural language processing techniques) further broaden its scope, making it more robust and adaptable across different languages, domains, and data types.

3. Methodology

3.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is the most popular subject modeling methods that consider the text as the source of the data. Before using LDA subject modeling method, Deerwested *et al.,* [38] introduced Latent Semantic Indexing (LSI) method, which is an automatic indexing model. LSI analysis uses singular-value decomposition to determine semantic relations between terms. A large matrix is taken with the term and associated data, and a semantic field is created where closely related terms and documents are placed close together [38]. In 1999, the Probabilistic Latent Semantic Analysis (LSA) model, which is an alternative to the LSI model, was developed by Hofmann. According to Hofmann [39] the probabilistic LSA model is based on a mixture decomposition derived from a latent class model. In other words, latent classes constitute a single class [39]. In this test, each word consists of a single topic, and different words in a document can be produced from different topics [40]. Although Hofmann's work is a useful method for probabilistic modeling of the text, it is an incomplete method since it does not create a probabilistic model at the document level. In this model each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers [40].

Blei *et al.,* [40] have developed the Latent Dirichlet Allocation (LDA) topic modeling method by revealing the deficiencies in the probabilistic LSA (pLSA) method. LSA is not suitable for mixture modeling while both LSA and pLSA suffers from lack of generalization ability [30]. With LDA these deficiencies are eliminated. LDA, is an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well [1]. Moreover, LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. It learns the various distributions such as the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document [3].

In the LDA algorithm, all words in each document are randomly assigned a subject. After the subject assignment is made to the documents, various statistics are extracted with this information. Local statistics show how many words are assigned to the topics in each document, while global statistics show how many times each word is assigned to each topic for the entire document. After obtaining statistical information, each word is reassigned to each word for each document [16]. Moreover, the LDA model is not necessarily tied to text, and has applications to other problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics [40].

Terms and notations can be extended as follows [41]:

- ✓ D is the number of documents in the entire corpus
- ✓ T is the number of topics and it is assumed to be known and fixed.
- ✓ Each topic ϕ_t, where $1 \leq t \leq T$, is a distribution over a fixed vocabulary of terms and $\phi_{tw}$is the term proportion of term w in topic t.
- ✓ θ_d is the topic mixture of the dth document and $\theta_{dt}$is the topic proportion of topic t in document d.
- ✓ z_d are the topic assignments for document d, where $z_{d,n}$ is the topic assignment fort he nth term in document d.
- ✓ w_d are the terms occuring in document d, where $w_{d,n}$ is the nth term in document d. All terms are elements of a fixed vocabulary.
- ✓ β is the Dirichlet prior on the topic-terms distributions
- ✓ α is the Dirichlet prior on the document-topics distributions.

LDA model is represented in Figure 1. Figure 1 indicates that there are three levels to the LDA representation. The parameters α and β are corpus level parameters and follows Dirichlet distribution. The variable θ is document-level variable. z and w are Word-level variables [40].

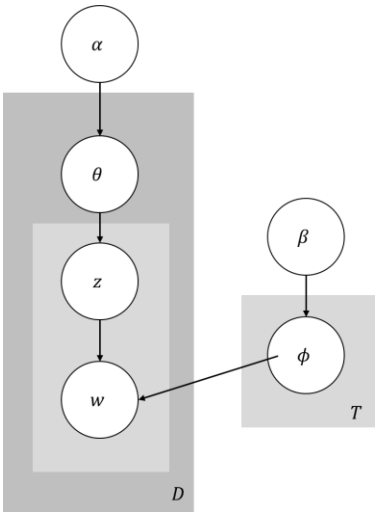


Fig. 1. Graphical representation of LDA. Adapted from [40]

LDA generative process works as follows [41]:

i. For each topic, choose a multinomial distribution ϕ_t from a Dirichlet distribution with parameter β. In other words, choose $\phi_t \sim Dir(\beta)$, where $1 \leq t \leq T$.
ii. For each document d, choose a multinominal distribution θ_d form a Dirichlet distribution with parameter α; i.e., $\theta_d \sim Dir(\alpha)$.
iii. For each term in document d, pick a topic assignment $z_{d,n}$ from the distribution θ_d fort he nth term in document d.
iv. Pick a term $w_{d,n}$ from the distribution $\phi_{z,d,n}$.

Advantages of the LDA model that is one of the most popular models about the topic modeling can be stated as follow [42]:

i. LDA model is the total probability generation model and can use efficient probability inference algorithms to calculate.

ii. The size of LDA model's parameter space and the number of training documents are independent making model more suitable for handling large scale corpus.
iii. LDA is a hierarchical model that is suitable and less likely to overfit than the non-hierarchical model.

3.2. Genetic Algorithms

Genetic algorithms which were introduced by John Holland [43] are biologically inspired search approaches that are suitable to a wide range of optimization problems [44].

A genetic algorithm for a particular problem must have the following five components [45]: (1) a genetic representation for potential solutions to the problem, (2) a way to create an initial population of potential solutions, (3) an evaluation function that plays the role of the environment, rating solutions in terms of their fitness, (4) genetic operators that alter the composition of children and (5) values for various parameters that the genetic algorithm uses (population size, probabilities of applying genetic operators, etc.).

Genetic algorithm procedure can be summarized as follow [46]. An initial population $P(t)$ is created in generation t. Each individual represents a potential solution to the problem considered. Each individual is evaluated to give some measure of its fitness with the help of fitness function. Some individuals undergo stochastic transformations by means of genetic operations to create new solution candidates (individuals). There are two types of transformation: crossover, which creates new individuals by combining parts from two individuals and mutation which creates new individuals my making changes in a single individual. Created individuals, called offspring $C(t)$, is then evaluated. A new population is formed by selecting the individuals having the highest fitness values from the parent population and the offspring population. After several iterations (generations), the algorithm converges to the best individual (solution), which hopefully represents an optimal or suboptimal solution to the problem. Pseudocode of the genetic algorithm is presented in Figure 2.

```
begin
    t ←0
    initialize P(t)
    evaluate P(t)
    while (not termination condition) do
    begin
        recombine P(t) to yield C(t);
        evaluate C(t);
        select P(t + 1) from P(t) and C(t);
        t ← t + 1;
     end
end
```

Fig. 2. Pseudocode of the Genetic Algorithm [49]

The fitness function plays an important role in any successful GA implementation since the main task of GA is to minimize the fitness function. The fitness function is a function which returns a numerical value measuring the goodness of an individual [47]. The fitness function accepts a candidate solution and produces an objective value as a measure of the performance of the candidate solution.

Crossover operation in GA implements a mechanism that mixes the genetic material of the parents. In crossover operation two solutions at n positions are split up and alternately assembled to

obtain new individuals. The motivation of such an operator is that both strings might represent successful parts of solutions that when combined even outperform their parents [47-50]. Another genetic operation in GA is mutation. It changes a solution by disturbing with random changes.

Users must determine when the GA will stop iterations. It is possible to allow GA to run for a predetermined number of generations and that is the most popular termination condition. It is also possible to allow run GA for a specific time or until no significant improvement in fitness value is observed.

There are some advantages and disadvantages of Genetic Algorithms [51]. Some of the advantages are: parallelism, liability, using only function evaluations, can be easily modified for different problems, can handle large or poorly understood search spaces easily. Limitations are: the problem of identifying fitness function, definition of representation for the problem, cannot use gradients, no effective terminator and require large number of response (fitness) function evaluations.

4. Analysis Section

In this section, the proposed model is explained (Figure 3). It is proposed to develop an auto-labeling process for "The Journal of Accounting and Finance". Auto-labeling process uses genetic algorithm and Latent Dirichlet Allocation simultaneously. Genetic algorithm is employed to reduce the number of words, while LDA is used to produce word probabilities that are used to classify the abstracts.

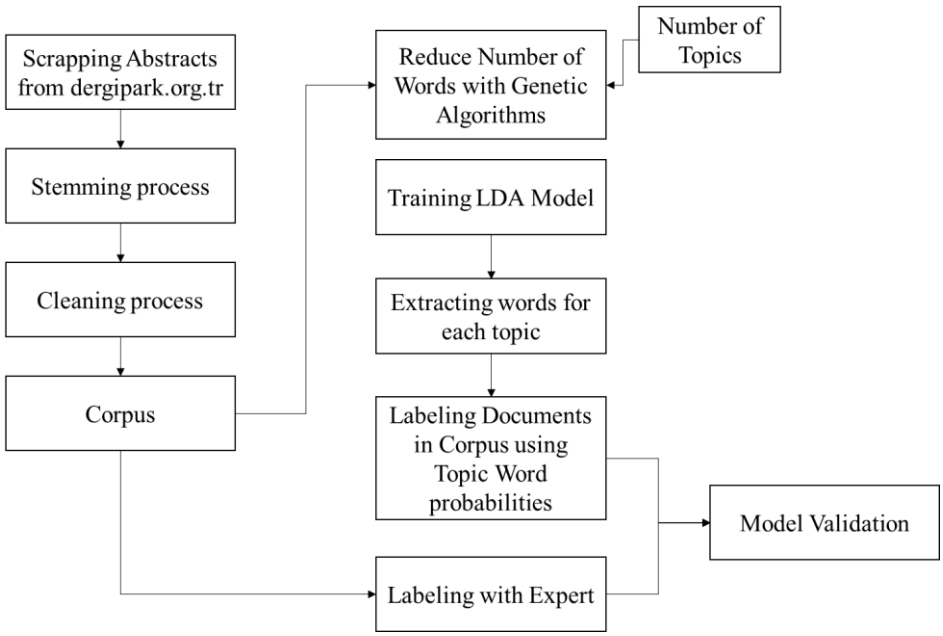


Fig. 3. Proposed Model

4.1. Dataset Description

The Journal of Accounting and Finance (The Journal of Accounting and Finance, 2020) only publishes academic articles in two fields namely accounting and finance. In other words, the number of topics will be fixed at two.

Abstracts are scrapped from dergipark.org.tr which is an internet site provides online hosting services for academic journals published in Turkey. Abstracts can be accessed from the website directly as raw text. However, full texts are presented in portable document format (PDF). Therefore,

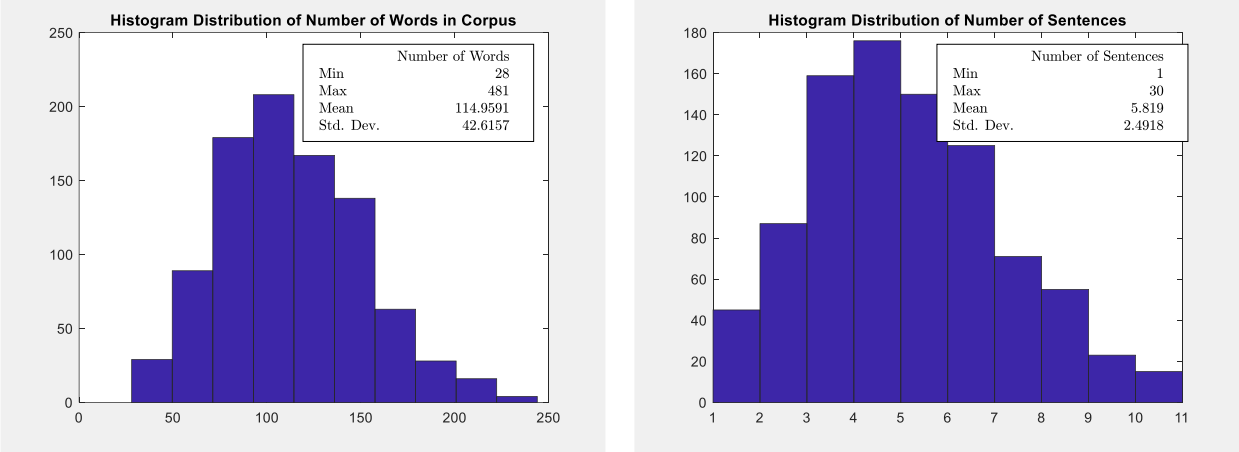
only abstract texts were used in the study. Number of articles in each year between 2005-2020 is presented in Table 1.

Initially 936 abstracts between 2005 and 2020 are scrapped. After the first manual inspection, it was determined that the abstracts of four articles were not available and the other six were in English. Since it was aimed to analyze Turkish texts in the study, it was decided to exclude six abstracts without Turkish abstracts. As a result, 926 Turkish documents are ready to be used in analysis. In this step there are 19021 unique terms in the corpus. All of the words in texts are converted to lowercase and special characters and punctuation marks (!\%$#&/:=()*?,.;) and numbers are removed since they don't contribute to the analysis.

Table 1
Number of Articles for Each year

Year	Number of Articles	Year	Number of Articles
2005	79	2013	41
2006	75	2014	39
2007	74	2015	44
2008	70	2016	43
2009	71	2017	60
2010	75	2018	54
2011	44	2019	100
2012	42	2020	17

Figure 4 contains the histogram distribution of the number of sentences and the number of words in 928 documents in corpus.



(a) (b)

Fig. 4. Histograms of dataset (a) Histogram of number of sentences (b) Histogram Distribution of Number of Words

Preprocessing steps applied in this study are as follow:

i. Since the suffixes of the words will affect the labeling process, stemming process is applied. In this study, for stemming purpose, Zemberek library [48] which is a natural language processing tools (version 0.17.1) for Turkish language and developed by Ahmet Akın is used. It has been shown that stemming has little effect when doing classification on Turkish corpus [49]. However, in this study, in order to select the word subset from a corpus a stemming process is needed.

ii. After stemming, the verbs remained in infinitive form. Verbs are used in both accounting and finance fields. This common use will not be useful in the separation of documents. In addition, some verbs may be used in specific areas. However, in this study, it was decided to remove the verbs from the corpus. After the stemming process, all verbs in infinitive form were excluded from the analysis.

iii. Words shorter than 2 characters and longer than 15 characters are eliminated. After this step, 9233 unique terms remained in the corpus.

iv. Words that appear two times or fever are removed from the corpus. At this step there are 3000 terms remained in the corpus.

4.2. Choosing number of topics

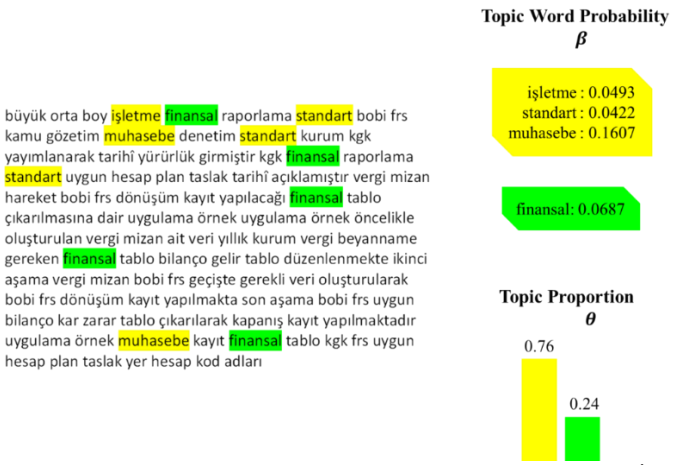
The journal we are reviewing is an academic journal that only publishes articles in the field of accounting or finance. Therefore, it is known that the number of topics is two. In addition, perplexity scores in different topic numbers were also examined. Perplexity score is used to determine the number of topics. Perplexity is a standard method to measure the prediction power of LDA [40].

$$perplexity = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (1)$$

Where N_d stands for the total number of terms occurred in the dth document. w_d is the dth word in the document. A lower perplexity score reveals a higher prediction power of the model [50].

Perplexity scores in different topic numbers are experimented and the lowest perplexity score is obtained when the number of clusters is two. It is observed that the perplexity score increases as the number of topics increases. Thus, it is confirmed that the number of topics should be two.

An illustrative example is presented in Figure 5. All of the preprocessing steps are applied to the first document. Moreover, some of the words are eliminated with GA-LDA method as will be explained later. Remained words are presented in Figure 5. Topic-word probabilities and topic proportions are also presented.



Abstract in English: Financial Reporting Standards for Large and Medium Sized Enterprises (BOBİ FRS) were published by the Public Oversight Accounting and Auditing Standards Authority (POA) and entered into force on 1.1.2018. The POA has announced the Draft Chart of Accounts in accordance with the Financial Reporting Standards. In this study, a complete application example is developed on how to make the transformation records to BOBİ FRS based on the tax balance and consequently the preparation of financial statements. In the application example, firstly, by using the data of the tax balance created, it is prepared in the financial statements (balance sheet and income statement) which should be annexed to the annual corporate tax return. In the second stage, the necessary data for the transition from tax balance to BOBİ FRS are created and the conversion records are made to BOBİ FRS and in the last stage, the balance sheet and profit / loss statement in accordance with BOBİ FRS are issued and closing records are made. balance sheet and profit / loss statement in accordance with BOBİ FRS are issued and closing records are made. In the complete application example, accounting codes and names in the FRS Draft Account Plan published by POA are used in the creation of accounting records and financial statements.

Fig. 5. Illustrative Example of document id: 01.

4.3. Optimizing Word Subset by Using Genetic Algorithm

After preprocessing, 3000 words remain in the data set. The number of words will be reduced by carrying out a selection process rather than using these words as they are. Removing the words that reduce the LDA score in the dataset from the dataset, will reduce the size of the dataset and can enhance the performance of LDA analysis. The parameters of the genetic algorithm used in the study are as follows: Crossover fraction 0.8, elite individual count 3, number of generations 500 and population size is 200. The fitness function of the genetic algorithm is perplexity score. Therefore, the genetic algorithm tries to optimize the subset of words that will minimize the perplexity score in the LDA model where the number of topics is fixed at 2. Pseudocode of the fitness function is presented in Figure 6.

input : ***cs*** and ***bag***
cs : candidate solution
bag : bag of words
1. ***new_bag*** = remove the words indexed by ***cs*** from ***bag***
2. create a ***new LDA model*** with two topics with ***new_bag***
3. calculate perplexity of ***new LDA model***.
4. Function output : perplexity of ***new LDA model***

Fig. 6. Fitness function design

The genetic algorithm reduced the number of words initially from 3000 to 2382 by reducing 618 words. In other words, the number of words (size of the data set) decreased by 21%.

4.4. Applying the Final LDA Model

After reducing the number of words with the genetic algorithm, a final LDA analysis was applied with the selected words. In the final LDA analysis, the number of clusters was determined to be 2, as in the fitness function in GA. Word cloud of topics is presented in Figure 7.

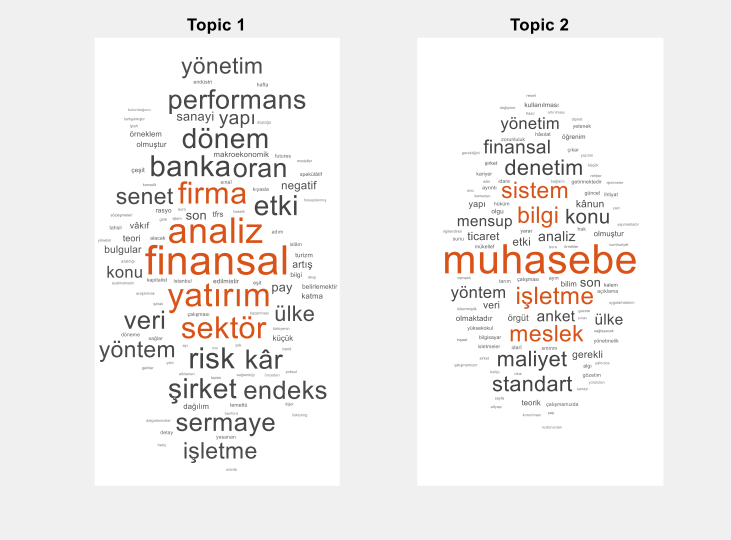


Fig. 7. Word Cloud of Topics (bilingual reference (English-Turkish) are presented in Table 2

4.5. Automatic Labeling Documents with LDA Scores

In this study, the documents are labeled according to their LDA scores. Final LDA Model produced topic word probabilities for both of the topics. These probability scores indicate the closeness of word to a topic. In other words, if a topic word probability is higher for a specific topic, than it is concluded

that this word is related with this topic. Topic word probabilities for some words are listed in Table 2. In the table, the first five words are related to the first topic, the financial area, and the last five words are related to the second topic, the accounting area. It should be noted that the probability values of the words in their own topics are higher than the probability values in the other topic.

Table 2
Topic word probabilities for some words

Word	Topic 1 (finance)	Topic 2 (accounting)
Finansal (financial)	0.0687	0.0347
Analiz (analysis)	0.0526	0.0188
Yatırım (investment)	0.0465	1.8824e-07
Etki (effect)	0.0326	0.0141
Şirket (company)	0.0311	0.0048
Muhasebe (accounting)	1.6903e-07	0.1607
Işletme (business)	0.0270	0.0493
Denetim (auditing)	1.6903e-07	0.0466
Meslek (profession)	1.6903e-07	0.0513
Standart (standard)	1.6903e-07	0.0422

A metric called Labeling Score has been developed to determine which document belongs to which topic. In this metric, the topic-word probability values of each word are summed. Here, the number of words to be determined by the user is taken into consideration. For example, the top five most likely words can be considered. In this case, the highest five topic-word probabilities are used in auto-labeling process. Let ρ number of words that will be used in labeling the documents, $p(w_t)$ represents the topic word probability for word w in topic t. In this case the labeling score of document d for topic t is defined as follows:

$$LS_{dt} = \sum_{i=1}^{\rho} p(w_{it}) \quad (2)$$

For some documents LS_{dt} scores will be same for both of the topics. If the LS score is equal for each topic, then it is not possible to label the abstract on either topic. That is why higher LS values are preferred. Experiments have been made to determine how many words will be used in the LS score calculation. Word numbers from 1 to 30 were applied, and the abstract number in which the LS score was the same for both topics was recorded and presented in Figure 8. In cases where more than 20 words are used, there is no significant decrease between the LS scores. Therefore, it was decided to use twenty words in the auto-labeling process.

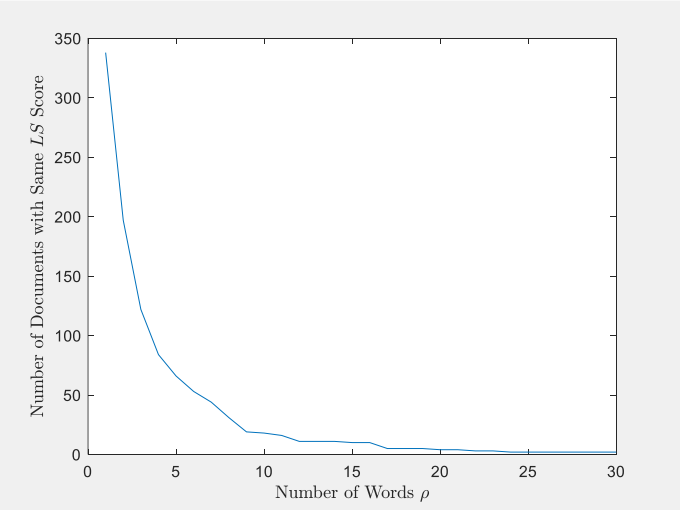


Fig. 8. LS scores for various number of words

4.6. Comparing the results with Expert Opinion

In this study, performance of two LDA models are compared. One is a single LDA model without a GA word selection process. The other LDA model (GA-LDA model) uses the genetic algorithm to reduce the number of words. Randomly selected 100 out of 928 abstracts were used for testing purposes to compare the performance of the models. In order to measure the classification performance of models, true labels are needed. True labels of these 100 abstracts are determined by an academician studying on Finance. Expert is evaluated the abstracts and provided the true document labels as finance or accounting. True labels and predicted labels can be presented in a diagnostic table (confusion matrix) as follow:

Confusion matrix can be defined as follow:

	True Finance	True Accounting	
Predicted Finance	True Positive (tp)	False Positive (fp)	(3)
Predicted Accounting	False Negative (fn)	True Negative (tn)	

Classification performance is compared with 5 different scores. These scores are defined as follow:

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (4)$$

$$Precision = \frac{tp}{tp+fp} \quad (5)$$

$$Recall = \frac{tp}{tp+fn} \quad (6)$$

$$Specificity = \frac{tn}{tn+fp} \quad (7)$$

$$F\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

Classification performance of LDA and GA-LDA models are compared with classification performance indicators. Results produced by the models with the results produced by the expert are presented in Table 3. The accuracy of GA-LDA model is %82. However, when a single LDA analysis is used, 79 out of 100 abstracts are the same as the expert's assessment.

Table 3
Classification Results

GA-LDA						
		Finance	Accounting	Total	Accuracy	0.8200
Expert	Finance	47	17	64	Precision	0.7344
	Accounting	1	35	36	Recall	0.9792
	Total	48	52	100	Specificity	0.8393
					F Score	0.9792

LDA						
		Finance	Accounting	Total	Accuracy	0.7900
Expert	Finance	44	17	61	Precision	0.7213
	Accounting	4	35	39	Recall	0.9167
	Total	48	52	100	Specificity	0.8073
					F Score	0.9167

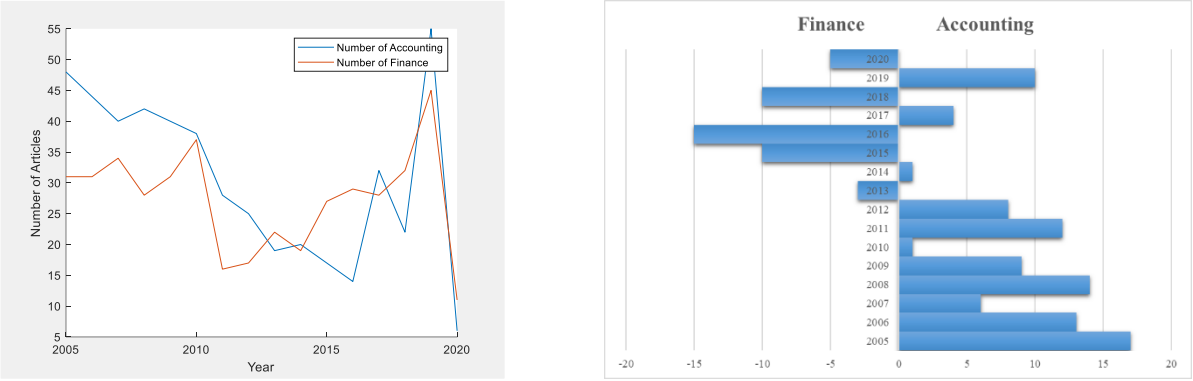
4.7. Number of Articles Published Over the Years

This section contains the results of the auto-labeling of the proposed GA-LDA analysis of 928 articles published over the years 2005-2020. All of the 928 articles are auto-labeled with the help of GA-LDA analysis.

Figure 9 represents the auto-labeling results. In part (a) of the figure, the number of auto-labeled articles in each topic is presented. It is noteworthy that there are more articles published in the field of accounting during the period. In general, there is a decrease in the number of articles published between 2010 and 2018.

In section (b) of Figure 9, there is a difference between the number of articles published in the field of accounting and finance. The fact that this difference is on the positive side shows that the publication of the publication for the year in question is high, while the negative side shows that the articles published in the field of finance for the year in question are predominant. It turns out that the years when the journal started to be published, the journal focused on accounting and nowadays it focuses on financial publications.

In section (b) Figure 9, the difference between the number of articles published in the field of accounting and finance is visualized with the help of bars. If the bar is to the right (left) side, the number of articles published in the field of accounting (finance) is high. It turns out that the years when the journal started to be published, the journal focused on accounting and nowadays it focuses on financial publications.



(a) (b)

Fig. 9. Time trends of two topics from 2005 and 2020

4.8. Comparison with other datasets

Proposed GA-word-selection model is also examined in various dataset. Benchmarking datasets are downloaded from [51]. In this website, various datasets are available in Turkish language.

i. News Dataset: This dataset, includes 1150 news in five categories (economics, magazine, health, politics, sports). In each category there are 230 different news. For each category randomly selected 20 news are assigned to testing dataset, while other 1050 news are used for training purpose.

ii. Movie Reviews Dataset: In this dataset, there are 105 movie reviews. There are three categories namely positive, negative and neutral reviews. In each category there are 35 reviews. For each category, randomly selected 10 writings are assigned to testing dataset, while other 75 reviews are used for training purpose.

iii. Mood Dataset: In this dataset, there are 157 diary writings. Four categories available (mixed, happy, nervous and sad). In each category there are 40 writings. For each category

randomly selected 10 writings are assigned to testing dataset while other 117 writings are used for training purpose.

There are two models used in benchmarking. The first model is the single LDA model where no improvements are made with GA. The second model is the GA-LDA model. In this model the Genetic Algorithm is used to reduce the number of words. Class numbers for each dataset is assigned as topic numbers in LDA models. Classification performance with benchmark datasets presented in Table 4. In each dataset, the GA-LDA models are outperformed LDA models.

Table 4
Classification performance with Benchmark Dataset

	News Dataset		Movie Reviews Dataset		Mood Dataset	
	LDA	GA-LDA	LDA	GA-LDA	LDA	GA-LDA
Accuracy	0.4700	0.1600	0.2667	0.3333	0.225	0.2750
Precision	0.1000	0.2800	0.3125	0.3333	0.1667	0.2941
Recall	0.1000	0.3500	0.5000	0.7000	0.3000	0.5000
Specificity	0.1000	0.3111	0.3846	0.4516	0.2143	0.3704
F Score	0.1000	0.3500	0.5000	0.7000	0.3000	0.5000

5. Conclusion

In this study, an intelligent GA-LDA system is developed to automatically label the articles published in the journal of "Accounting and Finance" which publishes academic papers mostly written in the Turkish language. This journal publishes papers on two topics which are accounting and finance. 928 abstracts published between 2005 and 2020 are scrapped from the journal's website. After scrapping the abstracts from the website, a cleaning process has been undergone. In the proposed system, the genetic algorithm selects the best word subset for LDA analysis. The fitness function is the perplexity score of LDA analysis by using the selected words.

Two LDA models are employed. One model uses GA to reduce the number of words, and the other is the raw LDA model. Since there is no readily available label set, we asked an expert to label the test set which contains twenty abstracts. In order to measure the performance of LDA models, expert opinions and LDA outputs are compared with 100 abstracts. Results indicate that the GA-LDA model performed better than LDA model. Purposed GA-LDA model is applied three different datasets and in each dataset GA-LDA model outperformed LDA model.

The advantage of the proposed system is that it can automatically classify documents without the need for expert knowledge. However, the question of what should be the number of topics still remains an area open to development.

This is a demonstration of a general approach that can be used for more complex journal systems which includes more topics. Moreover, if topic number is not known a priori, perplexity scores (or other techniques used to determine the number of topics) can be applied to determine the number of topics. The results obtained from this study provides insightful knowledge to the experts. For example, editors can use this system to automatically assign the referees.

Acknowledgement
This research was not funded by any grant.

Conflicts of Interest
The authors declare no conflicts of interest.

References

[1] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, *78*, 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[2] Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, *152*, 113401. https://doi.org/10.1016/j.eswa.2020.113401

[3] Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74-88. https://doi.org/10.1016/j.infsof.2018.02.005

[4] Ding, S., Li, Z., Liu, X., Huang, H., & Yang, S. (2019). Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model. *Information Sciences*, *499*, 12-24. https://doi.org/10.1016/j.ins.2019.05.037

[5] Pérez, J., Pérez, A., Casillas, A., & Gojenola, K. (2018). Cardiology record multi-label classification using latent Dirichlet allocation. *Computer methods and programs in biomedicine*, *164*, 111-119. https://doi.org/10.1016/j.cmpb.2018.07.002

[6] Lu, H. M., Wei, C. P., & Hsiao, F. Y. (2016). Modeling healthcare data using multiple-channel latent Dirichlet allocation. *Journal of biomedical informatics*, *60*, 210-223. https://doi.org/10.1016/j.jbi.2016.02.003

[7] Roque, C., Cardoso, J. L., Connell, T., Schermers, G., & Weber, R. (2019). Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads. *Accident Analysis & Prevention*, *131*, 336-349. https://doi.org/10.1016/j.aap.2019.07.021

[8] Zhang, W., Clark, R. A., Wang, Y., & Li, W. (2016). Unsupervised language identification based on Latent Dirichlet Allocation. *Computer Speech & Language*, *39*, 47-66. https://doi.org/10.1016/j.csl.2016.02.001

[9] Aydoğan, M., & Karci, A. (2020). Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification. *Physica A: Statistical Mechanics and its Applications*, *541*, 123288. https://doi.org/10.1016/j.physa.2019.123288

[10] Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing*, *50*, 135-141. https://doi.org/10.1016/j.asoc.2016.11.022

[11] Bay, Y., & Çelebi, E. (2016). Feature selection for enhanced author identification of Turkish text. In *Information Sciences and Systems 2015: 30th International Symposium on Computer and Information Sciences (ISCIS 2015)* (pp. 371-379). Springer International Publishing. https://doi.org/10.1007/978-3-319-22635-4_34

[12] Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, *35*(1), 136-147. https://doi.org/10.1016/j.tele.2017.10.006

[13] Parlar, T., Özel, S. A., & Song, F. (2016). Interactions between term weighting and feature selection methods on the sentiment analysis of Turkish reviews. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 335-346). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-75487-1_26

[14] Demirci, G. M., Keskin, Ş. R., & Doğan, G. (2019). Sentiment analysis in Turkish with deep learning. In *2019 IEEE international conference on big data (big data)* (pp. 2215-2221). IEEE. https://doi.org/10.1109/BigData47090.2019.9006066

[15] Onan, A., (2017). Türkçe Twitter mesajlarında Gizli Dirichlet Tahsisine dayalı duygu analizi. In Akademik Bilişim.

[16] Güven, Z. A., Diri, B., & Çakaloğlu, T. (2018, April). Classification of TurkishTweet emotions by n-stage Latent Dirichlet Allocation. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)* (pp. 1-4). IEEE. https://doi.org/10.1109/EBBT.2018.8391454

[17] Balcıoğlu, Y. S. (2024). Analyzing Customer Sentiments and Trends in Turkish Mobile Banking Apps: A Text Mining Study. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, (80), 49-69. https://doi.org/10.51290/dpusbe.1391631

[18] Shams, M., & Baraani-Dastjerdi, A. (2017). Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. *Expert Systems with Applications*, *80*, 136-146. https://doi.org/10.1016/j.eswa.2017.02.038

[19] Yeh, J. F., Tan, Y. S., & Lee, C. H. (2016). Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation. *Neurocomputing*, *216*, 310-318. https://doi.org/10.1016/j.neucom.2016.08.017

[20] Guo, C., Lu, M., & Wei, W. (2021). An improved LDA topic modeling method based on partition for medium and long texts. *Annals of Data Science*, *8*(2), 331-344. https://doi.org/10.1007/s40745-019-00218-3

[21] Lin, J. M., Bohland, J. W., Andrews, P., Burns, G. A., Allen, C. B., & Mitra, P. P. (2008). An analysis of the abstracts presented at the annual meetings of the Society for Neuroscience from 2001 to 2006. *PLoS One*, *3*(4), e2052. https://doi.org/10.1371/journal.pone.0002052

[22] Lienou, M., Maitre, H., & Datcu, M. (2009). Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters*, *7*(1), 28-32. https://doi.org/10.1109/lgrs.2009.2023536

[23] Celikyilmaz, A., Hakkani-Tur, D., & Tür, G. (2010). LDA based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search* (pp. 1-9).
[24] Ekinci Ekin, O. & S. İlhan, (2016). Ürün özelliklerinin konu modelleme yöntemi ile çıkarılması. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 9(1), 51–58.
[25] Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, *80*, 83-93. https://doi.org/10.1016/j.eswa.2017.03.020
[26] Atici, B., S.I. Omurca & E. Ekinci, (2017). Kullanici Şikayetlerindeki Ürün Özelliklerinin Gizli Dirichlet Ayirimi ile Saptanmasi, In 2nd International Conference on Computer Science and Engineering, UBMK 2017.
[27] Li, G., Zhu, X., Wang, J., Wu, D., & Li, J. (2017). Using lda model to quantify and visualize textual financial stability report. *Procedia computer science*, *122*, 370-376. https://doi.org/10.1016/j.procs.2017.11.382
[28] Drosatos, G., Kavvadias, S.E., Kaldoudi, E. (2018). Topics and Trends Analysis in eHealth Literature. In: Eskola, H., Väisänen, O., Viik, J., Hyttinen, J. (eds) EMBEC & NBC 2017. EMBEC NBC 2017 2017. IFMBE Proceedings, vol 65. Springer, Singapore. https://doi.org/10.1007/978-981-10-5122-7_141
[29] Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?. *Information processing & management*, *54*(6), 1292-1307. https://doi.org/10.1016/j.ipm.2018.05.006
[30] Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*, *127*, 256-271. https://doi.org/10.1016/j.eswa.2019.03.001
[31] Bailón-Elvira, J. C., Cobo, M. J., Herrera-Viedma, E., & López-Herrera, A. G. (2019). Latent Dirichlet Allocation (LDA) for improving the topic modeling of the official bulletin of the spanish state (BOE). *Procedia Computer Science*, *162*, 207-214. https://doi.org/10.1016/j.procs.2019.11.277
[32] Gangadharan, V., & Gupta, D. (2020). Recognizing named entities in agriculture documents using LDA based topic modelling techniques. *Procedia Computer Science*, *171*, 1337-1345. https://doi.org/10.1016/j.procs.2020.04.143
[33] Chang, I. C., Yu, T. K., Chang, Y. J., & Yu, T. Y. (2021). Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals. *Sustainability*, *13*(19), 10856. https://doi.org/10.3390/su131910856
[34] Sharma, C., Batra, I., Sharma, S., Malik, A., Hosen, A. S., & Ra, I. H. (2022). Predicting trends and research patterns of smart cities: A semi-automatic review using latent dirichlet allocation (LDA). *IEEE Access*, *10*, 121080-121095. https://doi.org/10.1109/access.2022.3214310
[35] Madzík, P., Falát, L., & Zimon, D. (2023). Supply chain research overview from the early eighties to Covid era–Big data approach based on Latent Dirichlet Allocation. *Computers & Industrial Engineering*, *183*, 109520. https://doi.org/10.1016/j.cie.2023.109520
[36] Park, H., Ahn, B., & Kim, T. (2024). An exploration of research trends on metaverse: topic modeling with latent dirichlet allocation. *Quality & Quantity*, 1-20. https://doi.org/10.1007/s11135-024-01931-9
[37] Shashank, S., & Behera, R. K. (2024). Factors influencing recommendations for women's clothing satisfaction: A latent dirichlet allocation approach using online reviews. *Journal of Retailing and Consumer Services*, *81*, 104011. https://doi.org/10.1016/j.jretconser.2024.104011
[38] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9
[39] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). https://doi.org/10.1145/312624.312649
[40] Campbell, J. C., Hindle, A., & Stroulia, E. (2015). Latent Dirichlet allocation: extracting topics from software engineering data. In *The art and science of analyzing software data* (pp. 139-159). Morgan Kaufmann. https://doi.org/10.1016/b978-0-12-411519-4.00006-9
[41] Momtazi, S. (2018). Unsupervised Latent Dirichlet Allocation for supervised question classification. *Information Processing & Management*, *54*(3), 380-393. https://doi.org/10.1016/j.ipm.2018.01.001
[42] Liu, Z., Li, M., Liu, Y., & Ponraj, M. (2011, July). Performance evaluation of Latent Dirichlet Allocation in text mining. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (Vol. 4, pp. 2695-2698). IEEE. https://doi.org/10.1109/FSKD.2011.6020066
[43] Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
[44] Kramer, O., & Kramer, O. (2017). *Genetic algorithms* (pp. 11-19). Springer International Publishing.
[45] Rocke, D. M., & Michalewicz, Z. (2000). Genetic algorithms+ data structures= evolution programs. *Journal of the American Statistical Association*, *95*(449), 347. https://doi.org/10.2307/2669583
[46] Gen, M., & Cheng, R. (1999). *Genetic algorithms and engineering optimization*. John Wiley & Sons. https://doi.org/10.1002/9780470172261

[47] Kaya, M., & Alhajj, R. (2005). Genetic algorithm based framework for mining fuzzy association rules. *Fuzzy sets and systems*, *152*(3), 587-601. https://doi.org/10.1016/j.fss.2004.09.014
[48] Akın, A, available online [https://github.com/ahmetaa/zemberek-nlp]
[49] Çağataylı, M., & Çelebi, E. (2015). The effect of stemming and stop-word-removal on automatic text classification in Turkish language. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part I 22* (pp. 168-176). Springer International Publishing. https://doi.org/10.1007/978-3-319-26532-2_19
[50] Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, 142-156. https://doi.org/10.1016/j.elerap.2018.04.003
[51] http://www.kemik.yildiz.edu.tr/?id=28