Journal of Information Systems and Management Research
Bilişim Sistemleri ve Yönetim Araştırmaları Dergisi

http://dergipark.gov.tr/jismar

Araştırma Makalesi / Research Article

Türkiye'de Yönetim Bilişim Sistemleri Alanında Yapılan Lisansüstü Tezlerin LDA Algoritması ile Konu Modellemesi*

Göktuğ İLİSU[a], Nursal ARICI[*,b]

[a] *Gazi Üniversitesi, Bilişim Enstitüsü, Bilişim Sistemleri, ANKARA, 06500, TÜRKİYE*
[b*] *Gazi Üniversitesi, Uygulamalı Bilimler Fakültesi, Yönetim Bilişim Sistemleri Bölümü, ANKARA, 06500, TÜRKİYE*

* *This article was produced from the Master's thesis prepared by Göktuğ İlısu under the supervision of Prof. Dr. Nursal Arıcı.*

MAKALE BİLGİSİ

Alınma: 29.09.2024
Kabul: 19.06.2025

Anahtar Kelimeler:
Yönetim bilişim sistemleri, konu modelleme, gizli dirichlet tahsisi algoritması

Sorumlu Yazar
e-posta:
goktugilisu@gmail.com

ÖZET

Yönetim Bilişim Sistemleri, işletmelerin ve kurumların stratejik, yönetsel ve operasyonel düzeylerdeki bilgi ihtiyaçlarını karşılamak için bilgi teknolojisi çözümlerini ve iş süreçlerini entegre etmeye odaklanan bir bilim alanıdır. Bu yönüyle bilgisayar bilimi, yönetim bilimi, istatistik, organizasyon teorisi, karar teorisi gibi çeşitli referans disiplinlerden beslenen çok disiplinli bir araştırma alanıdır. Bu çalışmanın temel amacı, Yönetim Bilişim Sistemleri bilim dalının Türkiye'deki lisansüstü tez konularına yansımalarını ve gelişimini incelemektir. Bu amaçla, 2002-2023 yılları arasında yönetim bilişim sistemleri alanında hazırlanan ve YÖK Ulusal Tez Merkezi web sitesi üzerinden erişilebilen 1070 lisansüstü tez (Yüksek Lisans-f: 951 ve Doktora-f: 119) inceleme kapsamına alınarak Gizli Dirichlet Tahsisi algoritmasıyla konu modellemesi gerçekleştirilmiştir. Konu modellemesinde kullanılan veri seti, lisansüstü tezlerin İngilizce özetleridir. Tez özetlerine öncelikle metin ön işleme ve kök çözümlemesi uygulanmıştır. Ortaya çıkan tüm kelimeler iç içe listelere dönüştürülüp LDA algoritması uygulanarak konu modelleri elde edilmiştir. Veri görselleştirme ile kelime bulutları, konu kümeleri, kelime sıklık histogramları ve belge- konu dağılımları oluşturulmuştur. Konu modellerinde çoğunlukla "data", "model", "research", "technology", "system" kelimelerinin yer aldığı tespit edilmiştir. Bu kelimelerin sıklıkla kullanıldığının tespit edilmesi yönetim bilişim sistemleri bilim dalında çalışılan konular için beklenen bir sonuç olarak değerlendirilmektedir.

DOI: 10.59940/jismar.1557818

Topic Modelling of Postgraduate Theses in the Field of Management Information Systems in Turkey with LDA Algorithm

ARTICLE INFO

Received: 29.09.2024
Accepted: 19.06.2025

Keywords:
Management information systems, topic modelling, latent dirichlet allocation algorithm

Corresponding Authors
e-mail:
goktugilisu@gmail.com

ABSTRACT

Management Information Systems is a branch of science that deals with integrating information technology solutions and business processes to meet the information requirements of businesses and organizations at strategic, managerial, and operational levels. In this respect, it is a multidisciplinary research field that draws upon several different disciplines, including computer science, management science, statistics, organization theory and decision theory. The primary objective of this study is to investigate the reflections and improvement of the field of Management Information Systems as evidenced by graduate thesis topics in Turkey. To this end, 1,070 graduate theses (951 MSc-f and 119 PhD-f) prepared in the field of Management Information Systems between 2002 and 2023 and accessible via the CoHE National Thesis Centre website were included in the scope of the study. Topic modeling was performed with the Latent Dirichlet Allocation algorithm. The data set employed in the topic modelling process comprised the English abstracts of graduate theses. The thesis abstracts were subjected to text preprocessing and stemming. The resulting words were converted into nested lists and topic models were obtained by applying the LDA algorithm. Data visualization was employed to create word clouds, topic clusters, word frequency histograms and document-topic distributions. It was established that the terms "data," "model," "research," "technology," and "system" were predominantly incorporated

into the topic models. The fact that these words are frequently used is considered as an expected result for the subjects studied in the field of management information systems.

DOI: 10.59940/jismar.1557818

1. *INTRODUCTION* *(GİRİŞ)*

Management information systems is a scientific discipline that centers upon providing solutions to the knowledge requirements of states, societies, organizations, groups, and individuals at strategic, managerial, and operational levels through information systems and technologies. It is a multidisciplinary field that develops methodologies to find a way out to real-life problems in an application-oriented framework and to rule information technology resources most appropriately by drawing on various reference study fields such as computer science, management science, statistics, organization theory, and operations research. The field is also concerned with issues in sociology, economics, and psychology, such as the utilization and influence of information technology [1].

Management Information Systems (MIS) is a study field that focuses on the strategic and practical use of technology to improve organizational performance. It lies at the intersection of business and technology, aiming to facilitate the flow of information within an institution to promote decision-making, coordination, control, analytics, and visualization of data [2].

Text mining significantly enhances the capabilities of management information systems by providing the required tools to analyze unstructured text data, leading to better decision-making, increased efficiency and a deeper understanding of both internal operations and external environments. MIS platforms facilitate the collection and storage of large scales of text data from several sources. Integrating text mining tools and techniques into MIS permits the processing and analysis of text data to generate actionable insights. Combining text mining with real-time data processing talents in MIS can provide up-to-date insights for timely decision-making [3].

1.1. Text Mining *(Metin Madenciliği)*

Scientific literature and documents from marketing and economic sectors are frequently gathered as extensive text data. Additionally, large datasets can be collected in semi-structured formats, such as log files from servers and networks. In this scenario, text mining analysis is highly useful for both unstructured and semi-structured textual data. Although text mining is akin to data mining, it specifically concentrates on text analysis rather than structured data [4].

Text mining, in other words, knowledge discovery, involves the extraction of valuable information from textual data. This field, called text analytics or natural language processing, integrates computer science, linguistics, statistics, and machine learning techniques to derive meaningful insights from unstructured text. Unstructured text data encompasses any text that lacks a predefined format or structure, such as emails, social media content, articles, and customer feedback [5].

1.1.1. Text preprocessing *(Metin önişleme)*

Before analysis, text data is typically subjected to preprocessing procedures to enhance its quality and ensure uniformity across different datasets. Text preprocessing represents a fundamental aspect of numerous text mining algorithms. A conventional text classification framework typically involves four main stages: preprocessing, feature extraction, feature selection, and classification. The research indicates that the effectiveness of the classification process is heavily influenced by the methods used in feature extraction, feature selection, and the choice of classification algorithm. However, the preprocessing stage has also been found to have a noticeable impact on the success of this process. Uysal et al. investigated the impact of preprocessing tasks, with a particular focus on their influence in the field of text classification. The preprocessing step typically comprises a series of tasks, including tokenization, filtering, lemmatization and stemming [6].

Tokenization is dividing a sequence of characters into discrete units, called tokens, which may include words or sentences. Punctuation marks may also be discarded. The resulting list of tokens is then used for further processing. The main aim here is to pinpoint individual words within a sentence. Effective text classification and mining rely heavily on a robust parser capable of accurately tokenizing documents.

The process of filtering typically involves the removal of specific words or phrases from documents. A common practice in text filtering is the removal of so-called "stop words." Stop words are lexical items that occur with high frequency in a text but lack significant content-bearing information. Examples of such words include prepositions and conjunctions. Similarly, words that appear with considerable frequency in the text are identified as having minimal information content and thus being unable to distinguish between different documents. Furthermore, words that appear infrequently are likely to be irrelevant and can be excluded from the documents [7].

Different capitalization patterns are used in the creation of text and document data points, forming sentences. Since documents contain many sentences, inconsistent capitalization can significantly complicate the classification of extensive documents. One standard method to tackle this issue is to convert all letters to lowercase, which effectively brings all words in the text and document into a consistent property space. Punctuation marks and private symbols are also excluded from the sentences because they can challenge classification algorithms.

Lemmatization examines words based on their morphological makeup, consolidating several inflexive forms of a word into a unique entity for analysis. In essence, stemming techniques strive to normalize verbs to their root forms and standardize nouns into a consistent format [5].

The objective of stemming methods is to identify the stem of derived words. The specific stemming algorithms employed vary depending on the language in question. In the case of English, the stemmer algorithm is a commonly utilized approach. Text stemming involves modifying words to generate different word forms by applying diverse linguistic operations like affixation (the attachment of prefixes and suffixes) [7].

1.1.2. Feature selection *(Özellik seçimi)*

After preprocessing, the text must be transformed into a numerical pattern suitable for machine learning analysis. A promising approach proposes that incorporating both syntactic and semantic features into text representations can be very effective for sentence selection, particularly in technical genomic texts. Another method to tackle syntactic challenges is to use the n-gram technique for feature extraction [8].

The n-gram technique involves identifying sequences of n-letters appearing in a specific order within a given text corpus. This method does not only serve as a direct representation of the text, but also rather functions as a feature for text representation. The Bag of Words (BOW) model represents text by using individual words non-sequentially. This model is simple to implement, and the text is represented by a vector, typically with a manageable dimensionality. An n-gram, in this context, is a BOW feature used to represent text through sequences of words. The use of two-letter and three-letter combinations is common. This approach allows the extracted text feature to detect more information than a single word [9].

In natural language processing, term frequency (TF) is a statistical metric used to determine how frequently a specific term or word appears in a corpus. The simplest form of weighted feature extraction involves TF, where each term is assigned a value based on its count throughout the corpus. More advanced approaches that build on TF often apply binary or logarithmic scaling to word frequencies for weighting. In these methods, documents are converted into vectors that represent word frequencies. While this technique is easy to understand, it can be limited by the overrepresentation of common words in the feature vectors [10].

The bag-of-words (BoW) model provides a streamlined and simplistic depiction of a text document by extracting key features such as word frequency. This approach finds applications in many areas, including document classification, information retrieval, computer vision, natural language processing (NLP), Bayesian spam filtering, and machine learning. In the BoW framework, a text—whether a document or a sentence—is represented as a collection of individual words. During the BoW process, word lists are generated. These words are not the elements that make up sentences and grammar; they are simply listed in a matrix without considering their semantic relationship. While the order of appearance and grammatical structure are ignored, the focal points of documents are still identified [11].

K. Sparck Jones proposed the concept of Inverse Document Frequency (IDF) to mitigate the influence of words that are inherently prevalent in a given sentence [11]. IDF dedicates higher weights to words that are either very frequent or infrequent across documents. The integration of Term Frequency (TF) and IDF is known as Term Frequency-Inverse Document Frequency (TF-IDF). The mathematical formula for calculating the weight of a term in a document using TF-IDF is expressed in equation (1).

$$W(d, t) = TF(d, t) * \log\left(\frac{N}{df(t)}\right) \quad (1)$$

In this framework, N denotes the overall documents, while df(t) indicates the number of documents that feature the term t. The initial component of the equation improves recall, whereas the latter component enhances the precision of the term's representation. Although TF-IDF helps reduce the impact of frequently occurring terms, it has its drawbacks. It fails to account for the semantic relationships between words within a document, treating each word in isolation. However, recent advancements in modeling, such as word embeddings,

offer new methods for capturing word similarities and integrating part-of-speech information [12].

1.2. Topic Modelling *(Konu Modelleme)*

A significant proportion of the literature is digitized and stored electronically in databases, either through digital libraries or social network databases. It is therefore necessary to have access to powerful automated tools to read this data and to realize the underlying themes. A significant pivotal objective of data evaluation is to discern the attributes that data entries exhibit in common. In the field of text analytics, this frequently entails the identification of the situations or constructs that a given document addresses. Although this information is intuitively understood by human readers, computer programs interpret text in its literal form. To address this challenge in programming, data scientists utilize topic modeling. Topic modeling is a widely adopted technique in text mining that reveals underlying patterns within large datasets. Though it is particularly effective for analyzing textual data, it is also valuable in fields such as bioinformatics, social science, and environmental studies. This approach helps structure extensive datasets, facilitating easier navigation and analysis [13].

The ability to derive valuable statistics and features from a dataset depends heavily on selecting the right methods. While contemporary topic modeling techniques greatly surpass earlier algorithms, they still need to be fine-tuned and optimized to ensure accurate results. Various topic modeling approaches are tailored to handle specific types of data relationships and structures, including short texts, long sequences, highly correlated information, and data with intricate structural patterns. To develop a topic modeling process that effectively meets the needs of a data analysis project, it is essential to grasp the distinctions between different models and the foundational algorithms of them [14].

1.2.1. Classification of Topic Modelling *(Konu Modelleme Sınıflandırması)*

Topic modelling is a statistical technique used to uncover the latent "topics" within a collection of documents. As a subset of unsupervised machine learning and natural language processing (NLP), it seeks to classify and structure extensive text corpora by identifying recurring patterns, themes, and structures. By applying a topic modelling algorithm to preprocessed data, one can discover these underlying patterns and topics. Notable algorithms in this field include Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

Latent Dirichlet Allocation (LDA) is a highly regarded method in topic modeling. It operates on the premise that documents are composed of a combination of topics, each of which is a collection of words. LDA functions by iteratively dedicating words to topics based on their co-occurrence patterns within the documents, gradually refining the association between words and topics [15].

Non-Negative Matrix Factorization (NMF) presents an alternative approach for uncovering topics within a corpus. It achieves this by decomposing the document-term matrix into two distinct lower-dimensional matrices: one matrix that captures the underlying topics and another that reflects how documents relate to these topics [16].

1.2.2. Topic modelling with LDA algorithm *(LDA algoritmasıyla konu modelleme)*

In the realm of natural language processing (NLP) and machine learning, topic modelling has emerged as a powerful tool for uncovering hidden themes and patterns within large text corpora. Among the various algorithms used for topic modelling, Latent Dirichlet Allocation (LDA) stands out as one of the most influential and largely used methods.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that postulates that each document in a corpus is a mixture of distinct topics, with each topic itself a mixture of words. The fundamental idea behind LDA is to reverse-engineer this generative process to uncover the hidden topic structure within the documents [17].

The Dirichlet distribution is a key component of LDA, serving as a prior distribution over the topic distributions in documents and the word distributions in topics. It is parameterized by a vector of positive reals and ensures that the resulting distributions are proper probability distributions. For document-topic distributions, the Dirichlet prior is denoted by α, and for topic-word distributions, it is denoted by β. These hyperparameters influence the sparsity of the distributions, with smaller values leading to sparser distributions.

The generation process essentially models how a set of words in a document can be generated given a set of topics. By applying Bayesian inference, LDA aims to reverse this process to discover the hidden topic structure. LDA creation process is shown in Figure 1 [18].

Text preprocessing is applied as tokenization, stop words removal, and lemmatization to the text dataset. Following preprocessing, textual data is typically transformed into a numerical format using a document-term matrix. Here, each document corresponds to a row, and each word corresponds to a column in the matrix.

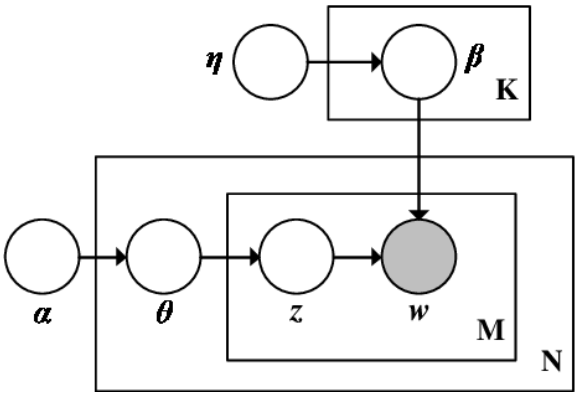


Figure 1. LDA creation process [18]
(LDA üretim süreci)

The number of topics is determined and LDA is applied to the preprocessed data by using the Gensim library of Python which provides an efficient and easy-to-use interface for running LDA.

The quality of the topics is evaluated using metrics such as perplexity (a metric for the model's sample prediction) and coherence score (an evaluation of the semantic similarity among high-probability words within a topic). Based on the evaluation results, the hyperparameters and the number of topics are also tuned to achieve the best performance.

Once the model converges, the topics are interpreted by examining the most probable words in each topic and the topic distribution is also interpreted for each document. Finally, the topics are visualized using tools like word clouds, topic distribution graphs, or interactive plots (e.g., LDAvis) to gain insights into the underlying themes in the corpus [19].

Latent Dirichlet Allocation (LDA) is a robust method for uncovering hidden topics in extensive text collections, offering insights into the underlying structure and themes of the data.

By leveraging the principles of Bayesian inference and Dirichlet distributions, LDA provides a robust framework for topic modelling, with wide-ranging applications in content analysis, recommendation systems, information retrieval, and beyond. Through careful implementation, evaluation, and interpretation, LDA can transform unstructured text data into meaningful and actionable knowledge [20].

2. LITERATURE REVIEW *(LİTERATÜR TARAMASI)*

A review of the literature revealed the existence of several studies on topic modelling with the LDA algorithm. However, there is a paucity of studies that have been conducted with the objective of conducting a quantitative and qualitative evaluation of postgraduate theses in terms of subject matter. The following section will provide an overview of these studies.

The research conducted by Çallı et al. [21] analyzed 574 graduate thesis abstracts completed between 2002 and 2020 in the MIS department in Turkey. The abstracts were examined using a text mining technique called the Latent Dirichlet Allocation algorithm. The analysis yielded 11 clusters, which were identified as follows: e-commerce and Marketing, System Development and Effects, Effects of Information Systems on Organizations, Data Mining, Human Resources Management, Organizational Change, Field Specific Studies I, Field Specific Studies II, Security, Education and Training, Prediction and Decision Support. In the context of this research, the similarities and differences between the estimation results and those presented in the national and international literature were discussed. This study aims to offer researchers in the field of management information systems insights and direction.

Parlina and Kusumarani [22] sought to employ bibliometric analysis to examine the intellectual structure and thematic development of the field of management information systems (MIS). A comprehensive analysis of the characteristics of publications in the three most prominent MIS journals in the SCOPUS database (IJIM, JSIS, and MIS Quarterly: Management Information Systems) was conducted, spanning the period from 1980 to 2021. In this study, the latent Dirichlet distribution (LDA) is incorporated into the approach to extend and improve the scientific research on MIS, resulting in a more comprehensive and up-to-date analysis. The indications of the study demonstrate the trend of publishing articles, the scientific structure, and the prominent issues in the top three journals.

Özköse and Gencer [23] conducted a comprehensive analysis of the field of Management Information Systems (MIS) through the use of bibliometric mapping. To achieve this objective, 222 journals that are indexed in the Science Citation Index Expanded (SCI-E) and the Social Science Citation Index (SSCI) were selected from the Web of Science and Scopus

databases. To determine the corpus of journals, 24 journals were selected for analysis, with the input of experts who could provide a more nuanced interpretation of the field. Initially, 20,497 English-language articles from these journals were gathered from the Web of Science (WoS) Core Collection between the years 1980 and 2015. Following text mining, the most influential organizations, authors and countries are displayed on graphs with statistical analysis using BibExcel. Furthermore, the development per annum of published articles is illustrated, and a trend analysis of these articles is presented. Additionally, the most cited articles are provided. Subsequently, using VosViewer, the most pertinent terms in this field were extracted through co-occurrence analysis from abstracts and keywords. The terms and their clusters are displayed on a graph. Density maps were also employed. The graphs and density maps are interpreted in detail, respectively.

The objective of this study is to analyze the master's and doctoral theses published in the field of MIS in universities in Turkey and the TRNC by text mining. Topic modelling with the LDA algorithm is employed as a method, although a multitude of data visualization techniques are utilised.

3. *METHODOLOGY (YÖNTEM)*

The aim of this study is to perform topic modelling of postgraduate theses prepared in the field of management information systems in universities in Turkey by text mining. Latent Dirichlet Allocation (LDA) algorithm is used in topic modelling for this purpose.

3.1. Dataset Creation Process (*Veri Seti Oluşturma Süreci*)

The data set consists of 1070 postgraduate theses (Master' s-f: 951 and Doctorate-f: 119) prepared in the field of Management Information Systems between 2002 and 2023 and accessible through the YÖK National Thesis Centre website [24].

The YÖK National Thesis Centre website was accessed and the detailed search section was selected via the link https://tez.yok.gov.tr/UlusalTezMerkezi/ . Subsequently, the term "Management Information Systems" was entered into the primary discipline, branch of science and subject sections. This process was repeated for each of these occasions. The dataset was created by this way as an Excel table. To make the dataset suitable for use in Python, the relevant dataset was organized so that Turkish and English topics, Turkish and English thesis names, Turkish and English keywords, and thesis abstracts were included in separate columns of the Excel table.

3.2. Text Preprocessing (*Metin Önişleme*)

For the post graduate theses published in the field of MIS, there are thesis abstracts in both Turkish and English on the YÖK National Thesis Centre website. Within the scope of the study, text preprocessing was performed based on the English abstracts of these theses and the Subject (English) column of the data structure related to the organized data set. The text preprocessing process was applied with Python before the subject models, which are intended to be created by using only thesis abstracts as a data structure. Through the Python Re library, all punctuation marks and numbers in the English abstracts of the post graduate theses were removed. The content of all remaining texts was converted into lower case letters.

Stopwords of the English language are accessible through NLTK library of Python. In addition to the aforementioned stop words, the NLTK library also includes a list of words that have been evaluated as having lost their meaning in English thesis abstracts. These include:

[‘In’, 'using', 'used', 'also', 'however', 'since', 'via', 'within', 'although', 'among', 'besides','whereas', 'dont', 'u', 'can', 'non', 'thus', 'may', 'towards', 'according', 'study', 'thesis', 'one', 'result', 'obtained', 'different', 'many', 'first', 'second', 'third', 'important', 'use', 'along', 'therefore', 'around', 'moreover', 'furthermore', 'nevertheless', 'whether', 'with', 'without', 'could', 'would', 'should', 'often', 'fourth', 'fifth', 'sixth', 'always', 'generally', 'sometimes', 'never', 'whenever', 'hence', ‘across’, 'thereby', 'thesis', 'before', 'after', 'meanwhile']

The Python Gensim library offers a simple preprocess function that can be used to break down text into words. In this case, the function was used to tokenize the abstracts. The stop words were then extracted from the English thesis abstracts using the simple preprocess function. A nested list was created for all remaining words. Each sub-list in the nested list consists of the words in a thesis abstract that have been removed from all stop words.

The English natural language processing model, provided by the Spacy library, was initially loaded using the command nlp = spacy.load(‘en_core_web_sm’). Subsequently, the words within each sub-list, which constituted the nested list, underwent lemmatization.

During the lemmatization process, only the inflectional suffixes were removed from the end of each word. As the removal of these suffixes does not alter the meaning of the word, the resulting stems were deemed to be appropriate for inclusion in the nested list. However, no changes were made to words with construction suffixes, as the addition of these suffixes alters the meaning of the root word. Several words were identified as being repeated in the sub-lists of the nested list of lemmatized words. These were removed, and each word was permitted to appear only once in each sub-list.

3.3. Evaluation of TF, IDF and TF-IDF Values *(TF, IDF ve TF-IDF Değerlerinin Hesaplanması)*

To achieve this process, the nested list containing the unique words in the thesis groups (repeated words are removed) is converted into merged texts using the Python programming language. Subsequently, a term frequency-inverse document frequency (TF-IDF) matrix is generated for the concatenated texts through the application of the TF-IDF vector generation function within the scikit-learn library. Subsequently, the matrix is transformed into a word sequence. The term frequency of the words in the sequence is calculated using Python code. Ultimately, graphs are generated to illustrate the term frequency (TF), inverse document frequency (IDF) and term frequency-inverse document frequency (TF-IDF) frequency of the most frequently used words.

3.4. Conversion of Text Data Into Numerical Format *(Metin Verilerinin Sayısal Formata Dönüştürülmesi)*

This is the final stage to be applied before topic modelling with LDA. A word dictionary is created with the corpora function in the Python Gensim library, utilizing the words in the nested list and removing any instances of repetition. This dictionary assigns an identification number (ID) to each unique word. The id2word.doc2bow function generates a list comprising the ID2WORD ID number of each word in the dictionary and the number of times this word occurs in the text. Here, doc2bow denotes 'document to bag-of-words'.

This process is repeated for each word in the lexicon, with the resulting data stored in a list called "corpus." This corpus contains numerical representations of the frequencies of each word. An example of a visual representation of the corpus is presented in Figure 2.

[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 1), (27, 1), (28, 1), (29, 1 9, 1), (40, 1), (41, 1), (42, 1), (43, 1), (44, 1), (45, 1), (46, 1), (47, 1), (48, 1), 1), (59, 1), (60, 1), (61, 1), (62, 1)], [(2, 1), (7, 1), (12, 1), (20, 1), (22, 1), (2 1), (58, 1), (63, 1), (64, 1), (65, 1), (66, 1), (67, 1), (68, 1), (69, 1), (70, 1), (7 1), (81, 1), (82, 1), (83, 1), (84, 1), (85, 1), (86, 1), (87, 1), (88, 1), (89, 1), (9 1), (100, 1), (101, 1), (102, 1), (103, 1), (104, 1), (105, 1), (106, 1), (107, 1), (10 (117, 1), (118, 1), (119, 1), (120, 1), (121, 1), (122, 1), (123, 1), (124, 1), (125, 1 1), (42, 1), (51, 1), (54, 1), (67, 1), (99, 1), (104, 1), (118, 1), (123, 1), (130, 1) 9, 1), (140, 1), (141, 1), (142, 1), (143, 1), (144, 1), (145, 1), (146, 1), (147, 1),

Figure 2. Creation of the corpus
(Külliyatın oluşturulması)

3.5. Determination of LDA Topic Modelling Performance Metrics *(LDA Konu Modellemesi Performans Ölçülerinin Belirlenmesi)*

Performance measures for LDA topic modeling are perplexity, coherence, exclusivity and corpus distance.

Perplexity is a measure of the uncertainty of a language model. It is often used to evaluate how well a language model performs. A lower perplexity value indicates that the model performs better and better represents the text data. Perplexity, which is the negative exponent of the logarithmic probability of the model, is calculated as given in equation (2).

$$\text{Perplexity} = \exp\left[-\frac{\sum_{d=1}^{M} \log P(w_d)}{\sum_{d=1}^{M} N_d}\right] \qquad (2)$$

Here $P(w_d)$ represents the probabilities of the words in the document. N_d represents the total number of words in the document.

Coherence is a measure of the extent to which the topics in a topic model are meaningful and coherent. Topic models represent topics, which are usually made up of words. Coherence measures how well these words relate to each other. A higher cohesion score indicates more meaningful and coherent topics. Cohesion uses the frequency of co-occurrence of words and similarities between word vectors.

The exclusivity of a topic model is a measure of the degree to which the topics are distinct from one another. A higher exclusivity value indicates that the topics are more unique and separable from one another. This measure is particularly crucial in models with a substantial number of topics. The exclusivity of a topic model is typically calculated by examining the number of overlaps between the top-ranked words of each topic. If there are minimal overlaps between the high-frequency words in the topics generated by the model, the exclusivity will be high.

The corpus distance is a metric used to assess the similarities and differences between the topics of

various LDA models. It provides insight into the extent to which the topics produced by the models are analogous or disparate. The corpus distance is typically calculated using measures such as the Kullback-Leibler (KL) divergence or the Jensen-Shannon divergence. These measures quantify the discrepancy between two probability distributions. For instance, the KL divergence between the topics of two LDA models indicates the extent of their divergence.

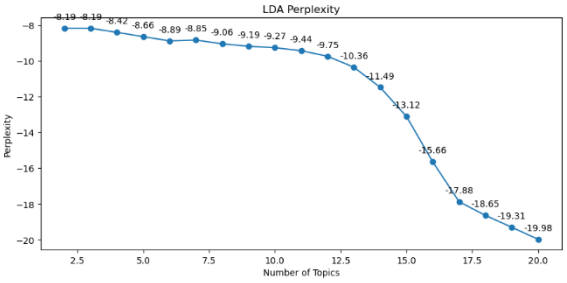
The formula for KL divergence, where P and Q are the probability distributions obtained from two different LDA models, is as given in equation (3).

$$D_{KL}(P \parallel Q) = \sum_i \log \frac{P(i)}{Q(i)} \qquad (3)$$

The combined use of perplexity, compatibility, exclusivity and corpus distance measures helps to select the best model in subject modelling processes and to comprehensively evaluate the performance of the model [17].

4. TOPIC MODELLING WITH LDA ALGORITHM *(LDA ALGORİTMASIYLA KONU MODELLEME)*

In this section, the perplexity, corpus distance, coherence and exclusivity values were calculated using the Python programming language, and graphs were created for each measure. To calculate these values, the CoherenceModel, LdaModel, similarities and TfidfModel functions of the Python gensim library were installed. Subsequently, the corpus was transformed with TF-IDF values. The minimum number of topics to be formed was determined to be two, while the maximum number of topics was determined to be twenty. The LDA model was trained for different numbers of topics with the LDA model function. Finally, graphs for perplexity, corpus distance, compatibility and exclusivity values were created. Related graph is presented in Figure 3.



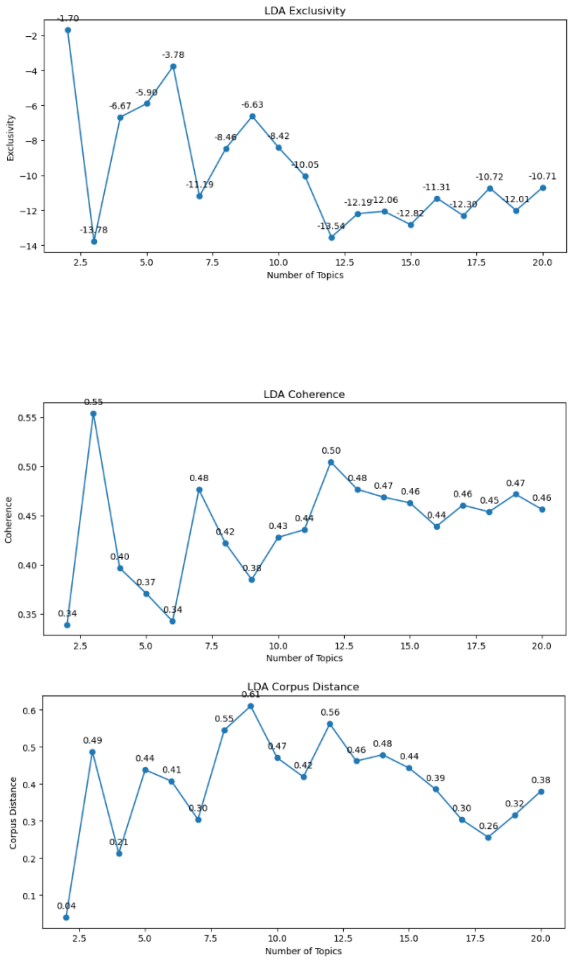


Figure 3. LDA topic modelling measures
(LDA konu modelleme ölçüleri)

Once the measures of perplexity, corpus distance, compatibility and exclusivity have been determined, the coherence values of the LDA models for a given set of topics are analyzed in order to ascertain the number of topics with the lowest compatibility value. This process is employed to ascertain the optimal number of topics that most accurately represent the text data. The coherence value is typically a measure of the consistency and meaningfulness of the identified topics. Consequently, determining the optimal number of topics represents a crucial step in the text analysis and model evaluation process. At this stage, the optimal number of topics was automatically determined by Python code based on the coherence value. The number of topics is assigned as "2" automatically by Python. Thus, there exist two topics by using English abstracts of postgraduate thesis as a dataset.

4.1. Creating LDA Topic Model *(LDA Konu Modelinin Oluşturulması)*

In the LDA topic model, which is suitable for the data structure in which postgraduate English thesis abstracts were used, the number of topics was automatically determined as 2. Subsequently, the corpus, number of topics and word identification number were used through the Python pprint library and the LdaMulticore function of the gensim library, resulting in the creation of the LDA topic model. A visual representation of the resulting topic model is presented in Figure 4.

```
[(0,
  '0.013*"datum" + 0.011*"information" + 0.011*"technology" + 0.010*"analysis" '
  '+ 0.010*"method" + 0.010*"system" + 0.008*"model" + 0.008*"process" + '
  '0.008*"research" + 0.007*"application"'),
 (1,
  '0.012*"datum" + 0.011*"research" + 0.009*"system" + 0.009*"technology" + '
  '0.009*"information" + 0.008*"model" + 0.008*"process" + 0.008*"method" + '
  '0.008*"analysis" + 0.007*"development"')]
```

Figure 4. LDA Topic Model
(LDA Konu Modeli)

As illustrated in Figure 4, the LDA topic model identifies two primary topics: Topic 0, which encompasses "Data and Information Technology Analysis" and Topic 1, which pertains to "Research and Development in Information Systems".

4.2. Visualization of LDA Topic Model *(LDA Konu Modelinin Görselleştirilmesi)*

Following the generation of the topic model, a visual representation of the LDA topic model was produced using the gensimvis function of the Python pyLDAvis library. This is presented in Figure 5.

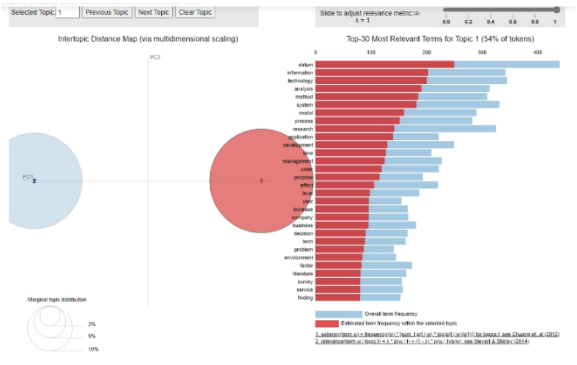


Figure 5. LDA topic model visualization when λ = 1
(λ = 1 iken LDA konu modeli görselleştirilmesi)

When the scroll bar is completely to the right (λ = 1), the words in Topic 0 (Data and Information Technology Analysis) are shown in Figure 5 as an example.

The lambda (λ) used in the PyLDAvis visualization is a tool for effectively exploring the results of the LDA topic model. In the PyLDAvis interface, the variable λ, which is used to determine how relevant a particular topic is to a particular word, can be set between 0 and 1. The properties of the variable λ are as follows [25]:

- λ = 1 means that the general frequency of words is more prominent.
- If λ = 0, the specific relevance of words to a particular topic is more prominent.

A shift to the left of the scroll bar (λ < 1) results in alterations to the position and frequency of words within the topics. This phenomenon is exemplified in Figure 6.

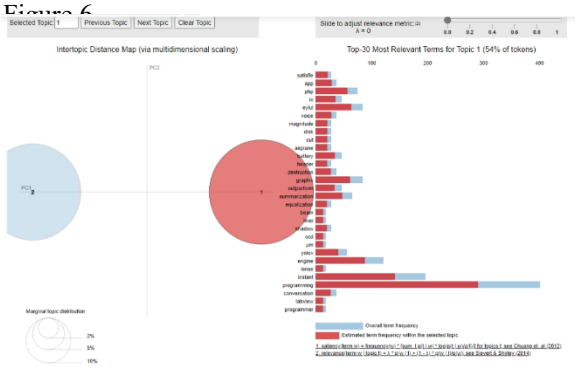
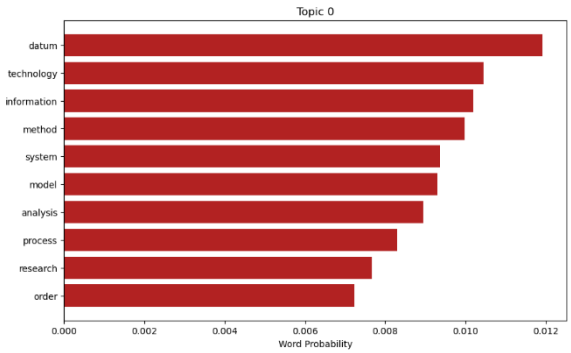


Figure 6. LDA topic model visualization when λ = 0
(λ = 0 iken LDA konu modeli görselleştirilmesi)

In the LDA topic model, which is represented as a cluster, Topic 0 is the predominant topic within the model, as it is the cluster with the largest area (54% of tokens are represented).

4.3. Displaying Topics with a Bar Graph *(Konuların sütun grafiğiyle görüntülenmesi)*

Matplotlib library of Python is used to create bar graphs to express the most used ten words in topics of the model. These bar graphs are illustrated in Figure 7 as a visualization of Figure 4.



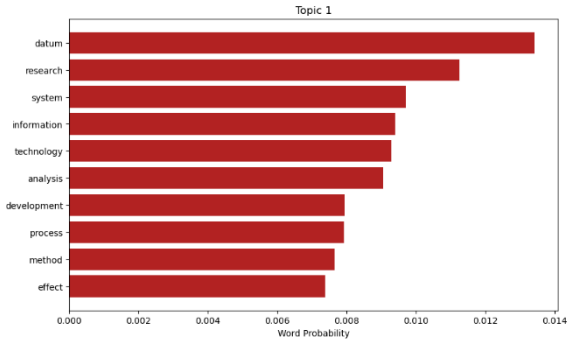


Figure 7. LDA topic model bar graphs
(LDA konu modeli bar grafikleri)

4.4. TF, IDF, TF-IDF values for postgraduate thesis abstracts *(Lisansüstü Tezlere ilişkin TF, IDF ve TF-IDF Değerleri)*

The TF, IDF and TF-IDF values are presented in Figure 8, with the graph displaying the 20 words with the highest term frequency, inverse document frequency and term frequency-inverse document frequency values, as observed in postgraduate thesis abstracts.

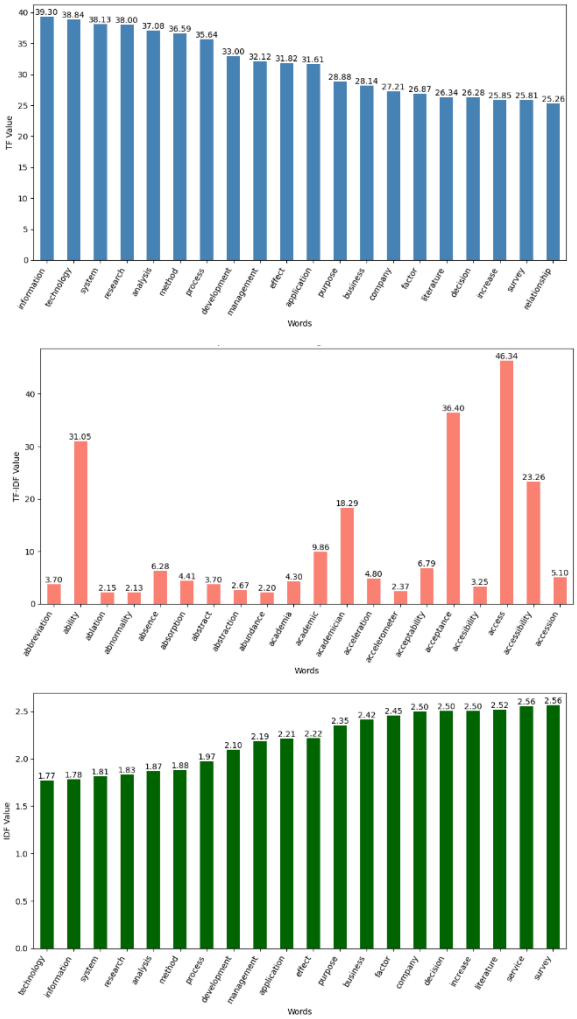


Figure 8. Words with the highest TF, IDF, TF-IDF values in postgraduate thesis abstracts
(Lisansüstü tez özetlerinde en yüksek TF, IDF, TF-IDF değerlerine sahip olan kelimeler)

Figure 8 illustrates that the word with the highest term frequency is 'information', with a TF value of 39,29. Conversely, the word with the highest inverse document frequency is 'service' and 'survey', with an IDF value of 2,56. The term 'access' had the highest term frequency-inverse document frequency (TF-IDF) score, recorded at 46,34.

4.5. Creating Word Clouds for the Topics in the LDA Topic Model *(LDA Konu Modelindeki Konular için Kelime Bulutlarının Oluşturulması)*

Figure 4 presents the word clouds containing the 10 most frequently mentioned words in the topics generated by the LDA topic model. These word clouds were created using Python libraries, specifically the wordcloud and matplotlib packages. The word clouds for the topic model are presented in Figure 9. It is seen that the words in the bar graphs in Figure 7 are the same as the words in the word clouds in Figure 9.

Figure 9. Word Clouds for LDA Topic Model
(LDA Konu Modeli Kelime Bulutları)

4.6. Creating a Heatmap of the LDA Topic Model
(LDA Konu Modeli için Isı Haritası Oluşturulması)

The output of the Latent Dirichlet Allocation (LDA) topic model can be used to create a visual representation of the distribution of documents (thesis abstracts) across topics, in the form of a heatmap. The LDA topic model takes the topic probabilities in each document and stores these probabilities in a matrix, which is then visualised as a heatmap. The resulting heatmap is presented in Figure 10.

The map provides a more accessible and analytically tractable representation of the output of the LDA topic model. The degree of colour intensity in each cell is indicative of the relevance of the document in question to the topic in question. In creating this heat map, the colour palette was set to 'YlGnBu'. The use of dark colours to represent high probability and light colours to represent low probability allows for a clear visualisation of the distribution of topics among documents and the degree of relatedness between documents and topics. The heat map demonstrates that both topics exhibit a high document probability, as indicated by the high density of dark colors. In this context, the term 'document' refers to the number of words in the original word list from which repeated words have been extracted for each thesis.

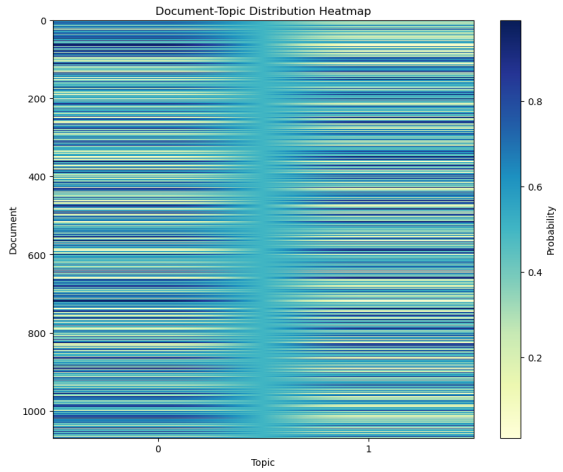


Figure 10. LDA Topic Model Document-Topic Distribution Heatmap
(LDA Konu Modeli Belge- Konu Dağılımı Isı Haritası)

5. CONCLUSION *(TARTIŞMA)*

In latent Dirichlet allocation (LDA) topic modelling, fit values serve as a means of evaluating the model's efficacy and the interpretability of the topics it identifies. A higher fit value indicates that the model performs better and generates more meaningful topics. The application of the LDA algorithm to the data set yielded topic models in which the words 'data', 'model', 'research', 'technology', and 'system' were identified as predominant.

In topic modelling, the size of the data set and its compatibility within itself are important factors. Despite the topic models being created from master's and doctoral theses in the field of management information systems, the different subjects and contents affect the topic model performance.

A comparison of the results obtained in the studies presented in Section 2 with the results obtained in this thesis can be expressed as follows:

In the study [21], the Latent Dirichlet Allocation algorithm, a text mining method, was employed to analyze 574 graduate thesis abstracts completed between 2002 and 2020 in the MIS department. In this thesis, the same method was used to analyze 1170 graduate thesis abstracts completed between 2002 and 2023. This indicates that the number of theses published during the three years between 2020 and 2023 is higher than the number of theses published during the 18 years between 2002 and 2020. Furthermore, subject differentiation was observed in the results obtained by the subject models. The analysis conducted with the LDA algorithm revealed the prevalence of topics such as data analysis, decision-making, system analysis, system development, and information management in postgraduate theses.

In the study [22], the application of topic modelling with the LDA algorithm led to the conclusion that the most frequently obtained topics in the first three MIS journals in the SCOPUS database were business performance, value management, data analysis, training, knowledge management and model use. Similarly, the topic modelling with the LDA algorithm applied in this thesis study yielded results that highlighted data analysis, knowledge management and modelling as prominent topics.

In the study [23], words such as "study", "research", "analysis", "use", "method", "algorithm" were identified as prominent in the density maps within the scope of bibliometric analysis studies conducted on leading journal articles in the field of management information systems in WoS. In light of the aforementioned findings, a comparison of the subject model presented in Figure 5, which is among the article's key findings, reveals a similar trend. This suggests that graduate theses and international journal articles share a significant overlap in terms of content.

In the future studies, topic models and model performances can be evaluated by including

postgraduate theses prepared in 2024. The processes can be improved by using different algorithms such as Top2Vec, LSA, Bertopic instead of the LDA algorithm. Apart from this study, an examination of the articles published in the field of MIS can be determined as another study topic.

REFERENCES *(KAYNAKLAR)*

[1] Baskerville, R. L., & Myers, M. D. (2002), "Information Systems as a Reference Discipline". *MIS Quarterly,* 26(1), 1–14. https://doi.org/10.2307/4132338

[2] Laudon, K. & Laudon, J. (2006), "Management Information Systems: Managing the Digital Firm", *9th ed. Prentice Hall.*

[3] Berry, M. W., & Castellanos, M. (2007). "Survey of Text Mining: Clustering", *Classification, and Retrieval.*

[4] O'Mara-Eves, A., Thomas, J., McNaught, J. *et al.* "Using text mining for study identification in systematic reviews: a systematic review of current approaches", *Syst Rev* 4, 5 (2015). https://doi.org/10.1186/2046-4053-4-5

[5] Peersman, C., Edwards, M., Williams, E. & Rashid, A. (2022), "A Survey of Relevant Text Mining Technology", *10.48550/arXiv.2211.15784.*

[6] Parlak, B., & Uysal, A. K. (2015, May). Classification of medical documents according to diseases. In *2015 23nd signal processing and communications applications conference (siu)* (pp. 1635-1638). IEEE.

[7] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D. (2019), "Text Classification Algorithms: A Survey", *Information 2019*(10), 150; doi:10.3390/info10040150.

[8] Liang, H., Sun, X., Sun, Y., Gao, Y. (2017), "Text feature extraction based on deep learning: a review", *Liang et al. EURASIP Journal on Wireless Communications and Networking*, (2017)211. doi: 10.1186/s13638-017-0993-1.

[9] Cavnar, W. B., & Trenkle, J. M. (1994, April), "N-gram-based text categorization", In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval (Vol. 161175*, p. 14).

[10] Biricik, G., Diri, B., Sönmez, A.C. (2012), "Abstract feature extraction for text classification", *Turk J Elec Eng & Comp Sci*, Vol.20, No. Sup.1, 2012, TÜBİTAK doi:10.3906/elk-1102-1015.

[11] Sakai, T., & Sparck-Jones, K. (2001, September), "Generic summaries for indexing in information retrieval", In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 190-198.

[12] Vayansky, I & Kumar, S. (2020), "A review of topic modeling methods", *Information Systems*, 94. 101582. 10.1016/j.is.2020.101582.

[13] Jurafsky, D., & Martin, J. H. (2000), "Speech and Language Processing: An Introduction to Natural Language Processing", *Computational Linguistics, and Speech Recognition*.

[14] Garoufallou, E., & Gaitanou, P. (2021), "Big data: opportunities and challenges in libraries, a systematic literature review", *College & Research Libraries, 82*(3), 410.

[15] Alghamdi, R., & Alfalqi, K. (2015), "A survey of topic modeling in text mining", *Int. J. Adv. Comput. Sci. Appl. (IJACSA), 6*(1).

[16] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), "Latent dirichlet allocation", *Journal of machine Learning research, 3*(Jan), 993-1022.

[17] Onan, A., Korukoglu, S., & Bulut, H. (2016), "LDA-based topic modelling in text sentiment classification: An empirical analysis", *Int. J. Comput. Linguistics Appl., 7*(1), 101-119.

[18] Kuang, D., Choo, J., & Park, H. (2015), "Nonnegative matrix factorization for interactive topic modeling and document clustering", *Partitional clustering algorithms*, 215-243.

[19] Sievert, C., & Shirley, K. (2014). "LDAvis: A method for visualizing and interpreting topics", Proceedings of *the Workshop on Interactive Language Learning, Visualization, and Interfaces*.

[20] Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021), "Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)", In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*, 341-354. Springer Singapore.

[21] Çallı, L., Çallı, F., & Alma Çallı, B. (2021), "Yönetim Bilişim Sistemleri Disiplininde Hazırlanan Lisansüstü Tezlerin Gizli Dirichlet Ayrımı Algoritmasıyla Konu Modellemesi", *MANAS Sosyal*

Araştırmalar Dergisi, 10(4), 2355-2372. https://doi.org/10.33206/mjss.894809

[22] Parlina, A., & Kusumarani, R. (2023), "A Latent Dirichlet Allocation-Based Bibliometric Exploration of Top-3 Journals in Management Information Systems", *Jurnal Studi Komunikasi dan Media*, 27(1), 77-92.

[23] Özköse, H., & Gencer, C. T. (2017). "Bibliometric analysis and mapping of management information systems field", *Gazi University Journal of Science, 30*(4), 356-371.

[24] Internet: YÖK National Thesis Center Web Site. [Accessed: 02/05/2024.] https://tez.yok.gov.tr/UlusalTezMerkezi/

[25] Sievert, C., & Shirley, K. (2014), "LDAvis: A method for visualizing and interpreting topics." *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*.