Vizyoner Dergisi

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.

ISSN: 1308-9552
https://ror.org/04fjtte88

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

TOPIC MODELLING OF DOCTORAL THESES WRITTEN ON LUNG CANCER IN TÜRKİYE USING LDA*

TÜRKİYE'DE AKCİĞER KANSERİ ÜZERİNE YAZILMIŞ DOKTORA TEZLERİNİN LDA İLE KONU MODELLEMESİ

Lecturer Fatma ÜZÜMCÜ[1]
Prof. Dr. Nezihe TÜFEKCİ[2]

ABSTRACT

The aim of the study is to examine the research status, subject and content of doctoral theses on lung cancer in Türkiye. In December 2024, research documents are scanned using the text mining method in R software, employing topic-based text analysis. The search is conducted on the YOK National Thesis Centre page, selecting 'lung cancer', 'all', and 'doctorate'. The most frequently covered topics are found through the obtained thesis abstracts with the artificial intelligence-based 'Latent Dirichlet Allocation' algorithm. Content analysis is performed by examining the relationship between the subject headings and thesis abstracts. It is aimed to determine the most emphasized content in theses on lung cancer. As a result of the algorithm, the words are found to be compatible in the consistency test. The study shows that lung cancer research is mainly clinical and medical, but the data also has significant health management and health economics outputs. A detailed investigation of concepts like "quality of life, treatment process, cost, and value" identify areas for health policies and technology assessments. Latent Dirichlet Allocation (LDA) emerges as a tool to compare studies across databases, helping researchers choose topics and understand the subject density of theses conducted in Türkiye.

Keywords: Lung Cancer, Thesis Analysis, Bioinformatics, Text Mining, Latent Dirichlet Allocation (LDA).

JEL Classification Codes: I11, I21, M19, P46.

ÖZ

Bu çalışmanın amacı, Türkiye'de akciğer kanseri üzerine yapılan doktora tezlerinin araştırma durumunu, konusunu ve içeriğini incelemektir. Araştırma dokümanları Aralık 2024'te R yazılımında metin madenciliği yöntemi kullanılarak konu tabanlı metin analizi ile taranmıştır. Arama, YÖK Ulusal Tez Merkezi sayfasında "akciğer kanseri", 'tümü' ve "doktora" seçenekleri seçilerek gerçekleştirilmiştir. Elde edilen tez özetleri üzerinden yapay zekâ tabanlı 'Gizli Dirichlet Ayrımı' algoritması ile en sık ele alınan konular bulunmuştur. Konu başlıkları ile tez özetleri arasındaki ilişki incelenerek içerik analizi yapılmıştır. Akciğer kanseri konulu tezlerde en çok vurgulanan içeriklerin belirlenmesi amaçlanmıştır. Algoritma sonucunda, kelimelerin tutarlılık testinde uyumlu olduğu tespit edilmiştir. Çalışma, akciğer kanserine yönelik araştırmaların ağırlıklı olarak klinik ve tıbbi nitelikte olduğunu ancak elde edilen verilerin sağlık yönetimi ve sağlık ekonomisi açısından da önemli çıktılar sağladığını göstermektedir. "Yaşam kalitesi, tedavi süreci, maliyet ve değer" gibi önemli kavramların ayrıntılı bir şekilde incelenmesi, sağlık politikaları ve teknoloji değerlendirmeleri için alanlar belirlemiştir. Gizli Dirichlet Ayrımı (GDA), veritabanları arasında çalışmaları karşılaştırmak için bir araç olarak ortaya çıkmış ve araştırmacıların konuları seçmelerine ve Türkiye'de yazılan tezlerin konu yoğunluğunu anlamalarına yardımcı olmuştur.

Anahtar Kelimeler: Akciğer Kanseri, Tez Analizi, Biyoenformatik, Metin Madenciliği, Gizli Dirichlet Ayrımı (GDA).

JEL Sınıflandırma Kodları: I11, I21, M19, P46.

---

* The paper is prepared from the Ph. D. Dissertation titled "Investigation of Patients' Psychosocial Experiences in the Diagnosis and Treatment of Lung Cancer" prepared by "Fatma ÜZÜMCÜ" under the supervision of "Nezihe TÜFEKCİ".

[1] Akdeniz University, School of Health Services, Department of Medical Services and Techniques, fatmauzumcu@akdeniz.edu.tr

[2] Süleyman Demirel University, Faculty of Economics and Administrative Sciences, Department of Health Management, nezihetufekci@sdu.edu.tr

Makale Geliş Tarihi / Received : 15.03.2025
Makale Kabul Tarihi / Accepted : 31.10.2025
https://doi.org/10.21076/vizyoner.1658488

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

GENİŞLETİLMİŞ ÖZET

Amaç ve Kapsam:

Çağdaş teknolojik gelişmeler bağlamında, metin madenciliği, bilgi işleme ve değerlendirmesinin etkinliğini artırma amacıyla metinsel materyallerin tanımlanması, incelenmesi ve analiz edilmesi için öncü bir metodolojiyi ifade etmektedir (Karakuş, 2021; Güneş ve Yıldırım, 2022). Metin madenciliği, hacimli metinsel veri kümelerinden anlamlı ve eyleme dönüştürülebilir içgörüler çıkarmak için bir dizi istatistiksel metodoloji kullanır (Feldman ve Sanger, 2007). Metin madenciliği, veri madenciliğinin bir çeşididir; fark, veritabanları gibi yapılandırılmış verileri işleyen veri madenciliğinin aksine, metin madenciliği veri kaynaklarının yapılandırılmamış metinler olmasıdır (Hearst, 1999). Bu çalışmanın amacı Türkiye'de akciğer kanseri üzerine yapılan doktora tezlerinin mevcut araştırma, konu ve içerik durumunu incelemektir. Aralık 2024'te mevcut araştırma belgelerinin kapsamlı bir araştırması, R yazılım ortamında bir metin madenciliği yöntemi olan konu tabanlı metin analizi kullanılarak gerçekleştirilmiştir.

Yöntem:

Araştırma, "akciğer kanseri" üzerine yazılmış doktora tezlerinin genelliği hakkında yargılarda bulunma açısından betimsel nitel bir yaklaşım kullanmaktadır. Nitel araştırmada, incelenen durumlara ve olaylara özgü daha ayrıntılı bir değer yargısına ulaşma çabası vardır (Morgan, 1996). Araştırma materyali akciğer kanseri üzerine yapılmış doktora çalışmalarından oluşmaktadır. Belgelerin analizi, R yazılım ortamında bir metin madenciliği yöntemi olan konu tabanlı metin analizi yaklaşımı kullanılarak gerçekleştirilmiştir. "Akciğer kanseri" arama terimi, aranacak alan olarak belirlenmiş, izin durumu "tümü" olarak ayarlanmış ve tez türü "doktora" olarak belirlenmiştir. Arama, Yükseköğretim Kurulu (YÖK) Ulusal Tez Merkezi web sitesi kullanılarak Aralık 2024'te gerçekleştirilmiştir. Arama teriminde "akciğer kanseri" teriminin kullanılması, İngilizce tezlerin bulunmasını kolaylaştırmak amacıyla tasarlanmıştır. Kapsamlı bir tarama sürecinin ardından YÖK Ulusal Tez Merkezi'nde kayıtlı toplam 1.007 tez olduğu, bunlardan 514'ünün tıpta uzmanlık tezi, 357'sinin yüksek lisans tezi, 130'unun doktora tezi ve 6'sının tıpta yan dal tezi olduğu tespit edildi. Bu araştırma kapsamında Türkiye'de akciğer kanseri üzerine yapılmış 130 doktora tezinin otomatik ve anlamsal analizini yapan bir konu modelleme yöntemi olan Latent Dirichlet Allocation (LDA) uygulanmıştır.

Bulgular:

Araştırmada kullanılan dokümanların metin madenciliği çalışması, Konu-5'in dokümanların %40'ından fazlasını kaplayarak en baskın konu olduğunu ortaya koymuştur. Bu bulgu, Konu-5'in dokümanlarda birincil tema olduğunu göstermektedir. Benzer şekilde, Konu-2, dokümanların yaklaşık %20'sini kapsayarak ikinci en yaygın konu olarak ortaya çıkmıştır. Bu bulgu, Konu-2'ye ait dokümanların konuya özgü bir alt temayı veya araştırma alanını kapsadığını göstermektedir. Konu 1, 3 ve 4'e ait dokümanlar daha düşük bir temsil düzeyi sergilemiştir. Bu konular, dokümanların metinlerin yaklaşık %10-15'ini temsil ettiğini ve daha az yoğun konuların göstergesi olduğunu göstermektedir. Analizden elde edilen bulgular, tek bir konu baskın olsa da metinde farklı temaların temsil edildiğini ortaya koymaktadır. Bu bulgu, analiz edilen metinlerin kapsadığı araştırma temalarının kapsamlı ve çeşitli doğasını vurgulamaktadır. Her konu için en yüksek kelime ağırlığına sahip beş kelime, konuların temalarını incelemek için analiz edilmiştir (Tablo 1). Konu-1 için en yüksek ağırlığa sahip kelimeler "hücre" (95,98), "EGFR" (63,19), "tedavi" (54,43), "tümör" (52,63) ve "mutasyon" (49,94) idi. Konu-2 için, "hücre" (55,32), "sclc" (25,74), "kök" (15,19), "ekzosomlar" (14,20) ve "emt" (12,97) terimleri en yüksek sıklığa sahiptir. Konu-3 için, 'genotip' (60,59), 'polimorfizm' (55,78), 'kontrol' (55,20), 'seviye' (48,18) ve 'önemli' (44,19) terimleri en yüksek sıklığa sahiptir. Konu-4 için, 'hücre' (285,39), 'ifade' (134,01), 'gen' (113,63), 'seviye' (85,67) ve 'etki' (76,58) terimleri en yüksek sıklığı göstermektedir. Beşinci konu için, "kontrol" (37,28), "kalite" (37,19), "yaşam" (27,94), "bulunan" (26,99) ve "tedavi" (24,60) terimleri en yüksek sıklığı göstermiştir.

Sonuç ve Tartışma:

İlk dört konunun kapsamlı bir analizi, akciğer kanseri araştırmalarının ağırlıklı olarak klinik ve tıbbi nitelikte olduğunu ve bulguların tıbbi bağlamda yorumlandığını ortaya koymaktadır. Genel bulgular Sağlık Yönetimi ve Sağlık Ekonomisi alanında gelecekte kullanılabilecek kavramları öngörmektedir. Ortaya çıkan konularla ilişkili anahtar sözcüklerin kullanımı hem ulusal hem de uluslararası düzeyde yayınlanan en çok atıf alan araştırmaların da kanıtladığı gibi, araştırmanın "Akciğer Kanserinin Tedavi Süreci ve Yaşam Kalitesinin Etkisi"ne doğru yönlendirilmesini kolaylaştırır. Konu 5 altında akciğer kanserinde "kontrol, kalite, yaşam, maliyet, bulunan, tedavi, değer" anahtar sözcüklerinin daha kapsamlı bir şekilde incelenmesi, Sağlık Ekonomisi alanındaki odak noktalarını aydınlatır. Bu anahtar sözcüklerin yakından incelenmesi, Konu 5'in yaşam kalitesi ölçümleri, tanı ve tedavi süreçleri, maliyet etkinliği ve genel sağlık dahil olmak üzere bir dizi temayı kapsadığını ortaya koymaktadır. Sonuç olarak, ülkeler bu anahtar sözcükleri kullanarak sağlık ekonomisi ve sağlık teknolojisi değerlendirmesi için çözümler ve sağlık politikaları belirleyebilir. Gizli Dirichlet Ayrımı analizi, farklı veri tabanlarından (örneğin PubMed, Scopus, Web of Science, vb.) alınan makale ve çalışmaların karşılaştırılmasını kolaylaştırmaktadır. Araştırmada yapılan analizler doğrultusunda, Türkiye'de yapılan alan tezlerinin konu yoğunluğu tespit edilebilir. Bu analiz, araştırmacılara uygun konuları seçmede rehberlik edebilir. Bu bağlamda, Gizli Dirichlet Ayrımı'nda ortaya çıkan etiketlerin önem düzeylerinin belirlenmesiyle etkili çıkarımlar yapılabilir.

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

1. INTRODUCTION

Lung cancer represents a significant and pressing health concern, with both global and national ramifications. As such, a comprehensive investigation into the treatment processes and their effects on patients' quality of life is essential. Cancer is characterized by genetic alterations that disrupt normal cell cycle checkpoints, allowing cancer cells to modify their metabolism in order to support unchecked proliferation and invasion (Yeşilmen, 2024). These metabolic reprogramming events, such as increased glucose uptake, enhanced glutaminolysis, and augmented fatty acid synthesis, play a pivotal role in promoting tumor growth and progression (Fadaka et al., 2017).

Lung cancer, which remains the most common form of cancer worldwide, represents 12.3% of all cancer cases. Its prevalence is significantly influenced by various factors including geographic location, gender, race, and, notably, tobacco use. Smoking, particularly long-term smoking, can increase the risk of developing lung cancer by as much as 30-fold (Minna et al., 2002). Lung cancer is categorized into two primary types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Among these, approximately 85% of lung cancer cases are classified as non-small cell lung cancer (Gridelli et al., 2015).

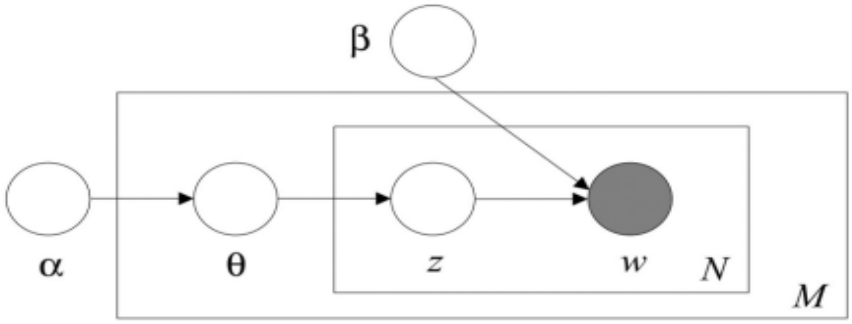
Lung cancer is the leading cause of cancer-related deaths, accounting for 25% of all cancer fatalities (Hoy et al., 2019). The survival rates for lung cancer patients are strongly correlated with the stage at diagnosis, with rates varying from 60% for localized tumors to a mere 6% for those with metastatic disease (Hoy et al., 2019). Although early detection of lung cancer is critical for improving patient outcomes, it remains a significant challenge due to the absence of specific early symptoms and the overlap with other chronic conditions. This highlights the urgent need for advanced molecular imaging technologies, which can aid in the early diagnosis and treatment of lung cancer (Kennedy et al., 2022).

In the context of modern technological advancements, text mining represents an innovative approach for identifying, examining, and analyzing textual data, with the ultimate goal of improving the efficiency of information processing and evaluation (Karakuş, 2021; Güneş & Yıldırım, 2022). This methodology utilizes a variety of statistical techniques to extract meaningful, actionable insights from large-scale textual datasets (Feldman & Sanger, 2007). While text mining is often considered a subset of data mining, it distinguishes itself by focusing on unstructured textual data, unlike traditional data mining, which primarily deals with structured data such as databases (Hearst, 1999).

As a branch of data mining, text mining incorporates a diverse range of techniques, including natural language processing, information retrieval, text summarization, text classification, and clustering. These techniques are increasingly utilized in modern online platforms and applications (Baker & Yacef, 2009; Karakuş, 2021). One of the key methods within text mining is topic modeling, which involves a set of algorithms designed to uncover the latent structures of topics within documents (Blei et al., 2003).

Latent Dirichlet Discriminant (LDD) is a probabilistic model often employed in machine learning and natural language processing. The Latent Dirichlet Allocation (LDA) model, in particular, was developed to address text mining and topic modeling challenges. LDA is widely used to identify latent topics across a collection of documents, enabling the representation of each document in terms of these underlying topics.

Figure 1. Latent Dirichlet Allocation (LDA) Graphical Model



Source: (Blei, Ng and Jordan, 2003, 997).

The LDA model's equation is expressed as follows:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (1)$$

In the formula, θ represents the document-specific topic distribution, z denotes the topic assignments, w indicates the words, α is the dirichlet prior parameter (document topic), β is the dirichlet parameter (topic word), and N signifies the number of words. In the context of Latent Dirichlet Allocation (LDA), topics and topic distributions per document are considered latent variables, whereas documents are observed entities. The allocation of each word to a topic is a latent construct called latent (Boyd-Graber et al., 2017). The algorithm is thus designated as Latent Dirichlet Allocation (LDA). The fundamental premise underlying the LDA approach is that a document is represented as a random composite of hidden topics, each of which possesses a unique characteristic determined by the frequency of word occurrences within the document, as illustrated in Figure 1 (Blei et al., 2003). The LDA modeling framework encompasses distinct levels, as illustrated in Figure 1. The parameters α and β represent the distribution of topics at the corpus level, which is defined as a collection of documents (M). The parameter α determines the distribution of topics in the document; that is, an elevated α value indicates a more substantial mix of topics in the document, and vice versa. The parameter β, on the other hand, governs the distribution of words within a topic. A higher value of β indicates a greater number of words in a topic, and vice versa. The variable θ, representing the topic distribution within a document, is a document-level variable (M). A higher value of θ indicates a greater diversity of topics in the document, while a lower value indicates a more specific topic.z_n and w_n are word-level variables (N). The variable z represents the topic of a particular word in a document, while ww represents the word related to a particular topic in the document (Blei et al. 2003).Latent Dirichlet Allocation (LDA) is a modelling approach in which the universe of documents is called 'structure' and the terms in these documents represent words (Maier et al., 2018). By using the words in the documents and their weights, hidden information is revealed and thus, the topics in the documents can be identified.

Sugiantoro et al.'s research employed abstracts as the primary source to identify commonly used topics in undergraduate thesis research through text mining using Latent Dirichlet Allocation (LDA). The research collected a dataset of 666 abstracts and employed LDA to identify dominant themes and topics within it, demonstrating that the LDA model achieved a consistency score of up to 0.448, indicating a reasonable level of consistency in the topics identified. The analysis yielded six primary topics: systems analysis and design, data mining, computer networks, decision support systems, software testing, and computer security (Sugiantoro et al., 2023).

Van Nieuwenhove (2017) conducted an investigation into the Latent Dirichlet Allocation (LDA) algorithm for topic extraction from documents, with a particular emphasis on the selection of parameters that influence the determination of the number of topics and the algorithm's instability. To achieve this, he performed a comparative analysis of LDA with Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA). The evaluation of these algorithms was based on several metrics, including model complexity (perplexity), model similarity, topic uniqueness, computation time, the cumulative distribution of annotated documents, and topic consistency. A novel metric introduced in the study was "uniqueness," which assesses the distinctiveness of topics and aids in determining the optimal number of topics (Van Nieuwenhove, 2017). The study concluded that LDA performed particularly well in terms of uniqueness and complexity, establishing it as the most suitable algorithm for further analysis. However, NMF outperformed LDA in certain metrics and proved to be a faster alternative. Despite this, issues regarding the instability of LDA remain unresolved. This research paper provides a balanced overview of the strengths and weaknesses of LDA, offering insightful perspectives into its applications and potential areas for improvement, particularly in comparison to other algorithms (Van Nieuwenhove, 2017).

Ekinci et al. (2020) conducted an automatic and semantic analysis of medical articles published by Turkish researchers using the LDA method. The study aimed to identify topics within the medical literature by analyzing articles from the PubMed database published over the past 11 years. The results of this study indicated that the LDA method effectively identified emerging trends in medical research and contributed to the discovery of significant themes within the literature (Ekinci et al., 2020).

Altıntaş et al. (2021) collected user posts related to cancer from the Reddit platform and performed topic modeling on the data using the LDA algorithm. A content analysis was conducted to identify the most frequently discussed topics in cancer-related posts. The topics identified were found to be consistent with each other based on the results of consistency tests, and the relationships between these topics were further analyzed using the t-SNE technique. This research provides a valuable contribution by analyzing the sharing of cancer-related information on social media platforms and identifying key themes within this content (Altıntaş et al., 2021).

Budak (2024) utilized the SCM framework, incorporating LDA as a crucial text mining method, to uncover hidden patterns and concerns regarding waste and recycling within reports from the European Environment Agency (EEA). By applying LDA to analyze word classification and co-occurrence within these reports, the study generated relevant tags that accurately reflected various aspects of waste management and recycling practices. This approach has been shown to enhance the accessibility of these reports, thereby supporting more informed decision-making and improved planning for waste management and recycling practices across Europe. The application of LDA in this context facilitates the identification of key issues and contributes to evidence-based studies focused on sustainable waste management (Budak, 2024).

Lung cancer has been observed to demonstrate high incidence and mortality rates in the population of Türkiye. This phenomenon poses a considerable burden on healthcare systems and represents a significant public health concern. The disease predominantly affects men, with an increasing prevalence among women. Treatment processes have been shown to exert pressure not only on medical resources but also on economic, social, and psychological support mechanisms. Thematic mapping of academic knowledge production is therefore crucial for accurately understanding the burden of the disease in healthcare management and developing evidence-based policies. Doctoral dissertations constitute a substantial repository of in-depth knowledge pertaining to healthcare delivery, treatment strategies, psychosocial support, and health economics. Current LDA applications primarily focus on social media posts or institutional reports. To the best of the author's knowledge, systematic topic modelling of doctoral dissertations on lung cancer in Türkiye has not been conducted. The present study is distinguished from other examples in the literature by virtue of the use of long-form, academic texts as a data source, the exclusive focus on lung cancer, and the identification of thematic trends and priority areas for health policy. In this respect, it is important to note that it fills an important gap in the literature in terms of data type and target area. The thematic evidence it produces at a national scale is significant for health management and policy planning.

2. MATERIALS AND METHODS

This study employs a descriptive qualitative approach to explore the general trends and patterns within doctoral theses focused on "lung cancer." In qualitative research, the objective is to formulate a more nuanced and context-specific value judgment that reflects the unique circumstances and phenomena under examination (Morgan, 1996). The research material comprises doctoral dissertations on lung cancer. The analysis of these documents was performed using a topic-based text analysis methodology, specifically utilizing a text mining technique, within the R software environment.

For the purposes of this study, the search term "lung cancer" was chosen as the primary keyword, and the search parameters were set to include all available documents, with the thesis type specified as "doctorate." The search was conducted in December 2024 via the National Thesis Center of the Council of Higher Education (YÖK) website. The selection of "lung cancer" as the search term was intended to streamline the retrieval of relevant theses written in English. Following a thorough screening process, it was determined that a total of 1,007 theses were registered in the YÖK National Thesis Center database. Of these, 514 were medical specialty theses, 357 were master's theses, 130 were doctoral theses, and 6 were sub-specialty medical theses.

Within the scope of this research, the Latent Dirichlet Allocation (LDA) model a topic modeling technique designed for the automatic and semantic analysis of text data was applied to analyze the 130 doctoral dissertations on lung cancer in Türkiye.

Figure 2. The Methodological Framework of the Study

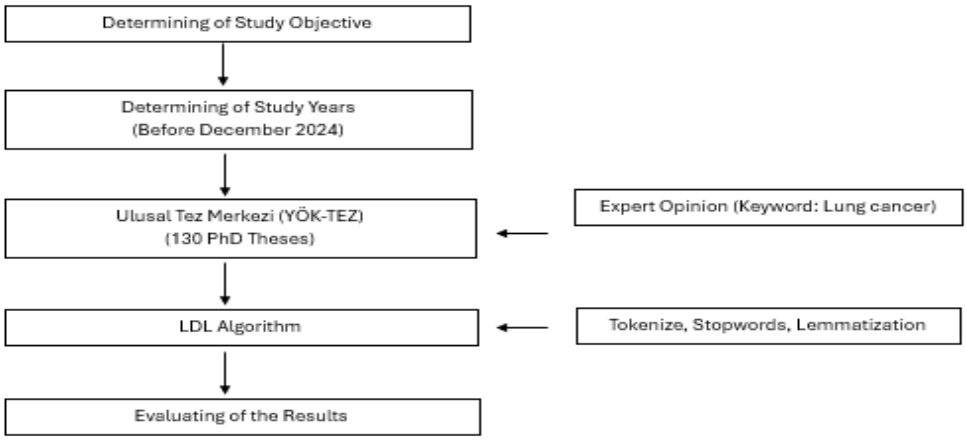


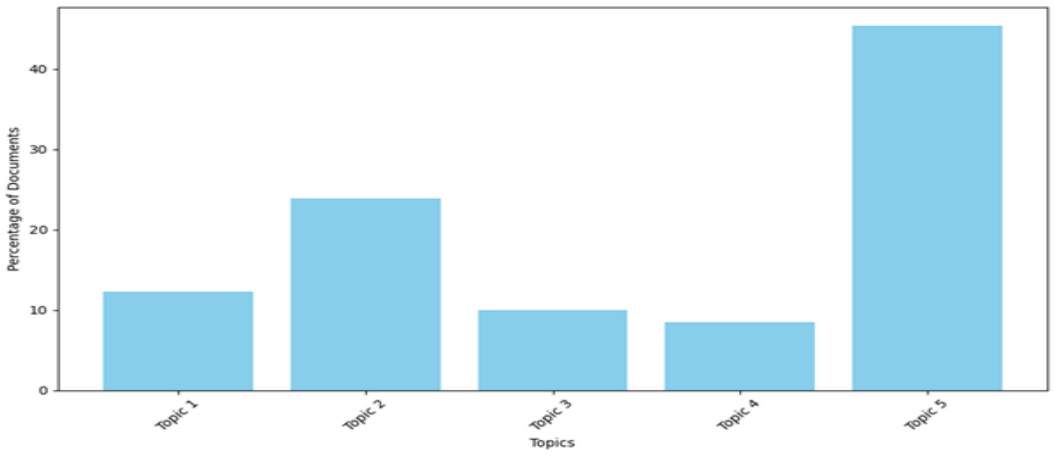
Figure 2 presents the flow diagram of the algorithm employed in this study concerning LDA. Only doctoral theses that included the term "lung cancer" and were uploaded to the National Thesis Center and published before December 2024 were included in the study, which constitutes a limitation of the research. While LDA is effective in identifying latent themes in text, model outputs are sensitive to initial values and hyperparameter settings. The outcomes of such studies may be subject to variation depending on the researcher's preferences, including the selection of the number of topics (K), the text preprocessing, and the stopword list (Van Nieuwenhove, 2017). In this study, the aforementioned limitations were mitigated by consistency scores and expert evaluations.

A probabilistic topic modelling approach was employed to categorise the text, create a topic model for doctoral dissertations, and analyse this model using LDA. The LDA algorithm was utilised to assess the efficacy of various variations in determining the optimal number of tags and keywords, employing diverse iterations. Initially, each word in the thesis was randomly assigned to a topic, and then iterative adjustments were made based on the likelihood of the words occurring more frequently. The calculations were arranged using the Gibbs sampling procedure, with five topics and 20 keywords under each topic over 2,000 iterations. The numerical value thus obtained was designated as "Weight." This established the proportion of the topics within the total and predicted the topic on which the subsequent study would be written.

3. FINDINGS

The results obtained in the study are explained through the five topics determined based on the objectives of the research. These five topics identified in the study are presented in Figure 3.

Figure 3. Subject Distribution Among Documents



The text mining analysis conducted on the documents used in this study revealed that Topic 5 emerged as the most dominant theme, accounting for more than 40% of the total documents. This substantial proportion suggests that Topic 5 plays a central role in the corpus, making it the primary focus across the analyzed materials. In a similar manner, Topic 2 was identified as the second most prominent theme, representing approximately 20% of the documents. This indicates that Topic 2 encompasses a specific sub-theme or research area that is significant but secondary to the overarching themes in the dataset.

On the other hand, Topics 1, 3, and 4 were represented to a lesser extent, with each of these topics contributing between 10% and 15% of the overall document content. This suggests that while these topics are present, they are less central and represent more niche or specialized areas within the broader context of the study. The analysis highlights that, although one topic dominates the corpus, multiple themes are still represented, reflecting the multifaceted nature of the research subjects covered in the documents. This diversity in thematic representation emphasizes the broad scope and varied research areas that are integrated within the texts under investigation.

Table 1. The Words with the Highest Weight for Each Topic

Topic-1		**Topic-3**		**Topic-5**	
Word	Weight	Word	Weight	Word	Weight
cell	95.97	genotype	60.59	control	37.27
egfr	63.19	polymorphism	55.78	quality	37.19
treatment	54.43	control	55.20	life	27.94
tumor	52.63	level	48.18	found	26.99
mutation	49.94	significant	44.18	treatment	24.60
Topic-2		**Topic-4**			
Word	Weight	Word	Weight		
cell	55.32	cell	285.39		
sclc	25.74	expression	134.01		
stem	15.19	gene	113.63		
exosomes	14.20	level	85.67		
emt	12.97	effect	76.58		

To examine the themes related to each topic, the five words with the highest word weights for each topic were analyzed (Table 1). For Topic 1, the words with the highest weight were "cell" (95.98), "EGFR" (63.19), "treatment" (54.43), "tumor" (52.63), and "mutation" (49.94). These results suggest that the subject matter is strongly related to cell biology and cancer research. Specifically, the terms "EGFR" (epidermal growth factor receptor) and "tumor" indicate a focus on cancer-related processes, while "treatment" and "mutation" point to the investigation of therapeutic strategies and their relationship with genetic variations in cancer.

In Topic 2, the terms with the highest frequency were "cell" (55.32), "SCLC" (small cell lung cancer) (25.74), "stem" (15.19), "exosomes" (14.20), and "EMT" (epithelial-mesenchymal transition) (12.97). These terms indicate that the documents in this topic predominantly focus on small cell lung cancer and stem cells as key research areas. The inclusion of "exosomes" and "EMT" suggests that the research also encompasses biological processes related to extracellular vesicles and tumor metastasis, which are critical to understanding cancer progression.

For Topic 3, the most frequent terms were "genotype" (60.9), "polymorphism" (55.78), "control" (55.20), "level" (48.18), and "significant" (44.19). These terms suggest that the research documented in this topic is centered around genetic studies, particularly focusing on genetic variations and polymorphisms. The terms "control" and "level" indicate that experimental controls and the measurement of genetic levels are integral to the research methodologies employed in these studies, highlighting the precision and rigor of genetic investigations.

Topic 4 is characterized by the terms "cell" (285.39), "expression" (134.01), "gene" (113.63), "level" (85.67), and "effect" (76.58). This combination of terms underscores the primary focus of the documents on gene expression and the regulation of biological processes at the cellular level. The prominence of "cell" and "expression" supports

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

this, while the terms "gene" and "effect" suggest that the impact of genetic factors and their expression levels are key areas of intensive research within this topic.

Finally, Topic 5 features the terms "control" (37.28), "quality" (37.19), "life" (27.94), "found" (26.99), and "treatment" (24.60). These terms suggest that the research in this topic is primarily concerned with the quality of life and the factors that influence it, particularly in the context of disease treatment and management. The repeated appearance of "quality" and "life" highlights the central role of health-related outcomes in the documents, while "treatment" emphasizes the importance of therapeutic interventions in improving life quality.

3.1. Findings Related to Topic-1

When the words addressed in Topic 1 are evaluated, the word cloud shown in Figure 4 emerges. Additionally, when the frequencies of the words are assessed, the frequency distribution presented in Figure 5 becomes apparent.

Figure 4. Word Cloud Related to Topic-1 as a Result of the Analysis of Documents

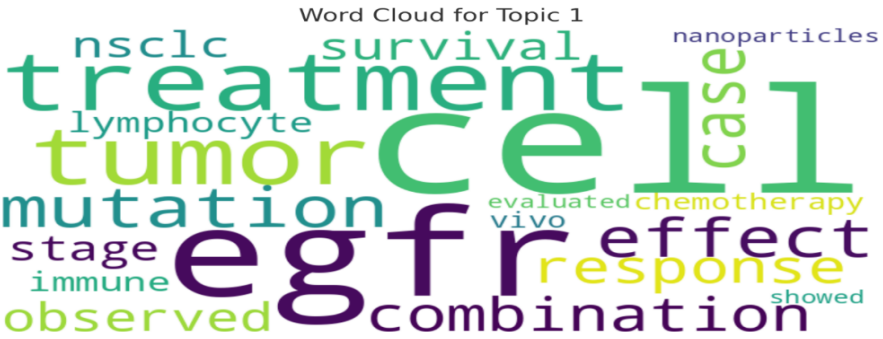
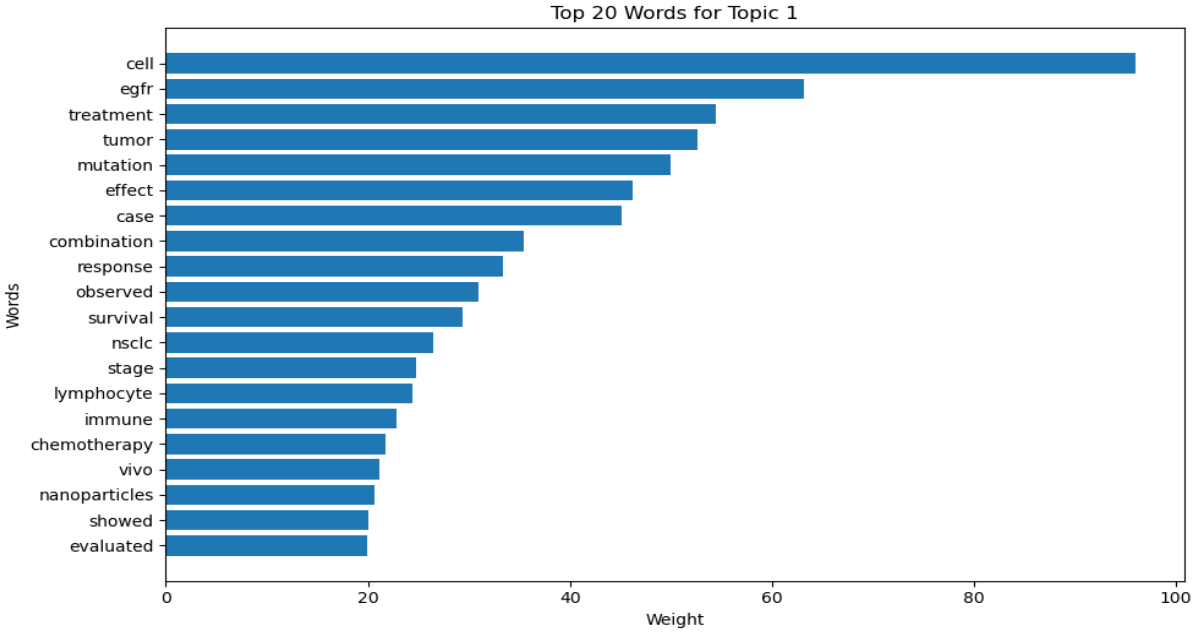


Figure 4 illustrates the word cloud generated from the lexicon of the sources related to Topic 1. In this visualization, the size of each word corresponds to its frequency within the dataset, providing a clear indication of the prominence of specific terms. Upon analyzing the word cloud, it becomes evident that the word "cell" appears most frequently, making it the most prominent term in Topic 1. Following "cell," the next most frequent words are "EGFR," "treatment," and "tumor," in that order. Additionally, Figure 5 displays a list of the top 20 words with the highest frequency in Topic 1, offering further insight into the key concepts and themes prevalent within the topic.

Figure 5. 20 Words with the Highest Frequency Related to Topic 1



Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

When Figure 5 is examined, the findings related to Topic 1 are listed as follows: cell, EGFR, treatment, tumor, mutation, effect, case, combination, response, observed, survival, NSCLC, stage, lymphocyte, immune, chemotherapy, vivo, nanoparticles, showed, and evaluated. Additionally, Figure 5 also illustrates the weight of each word.

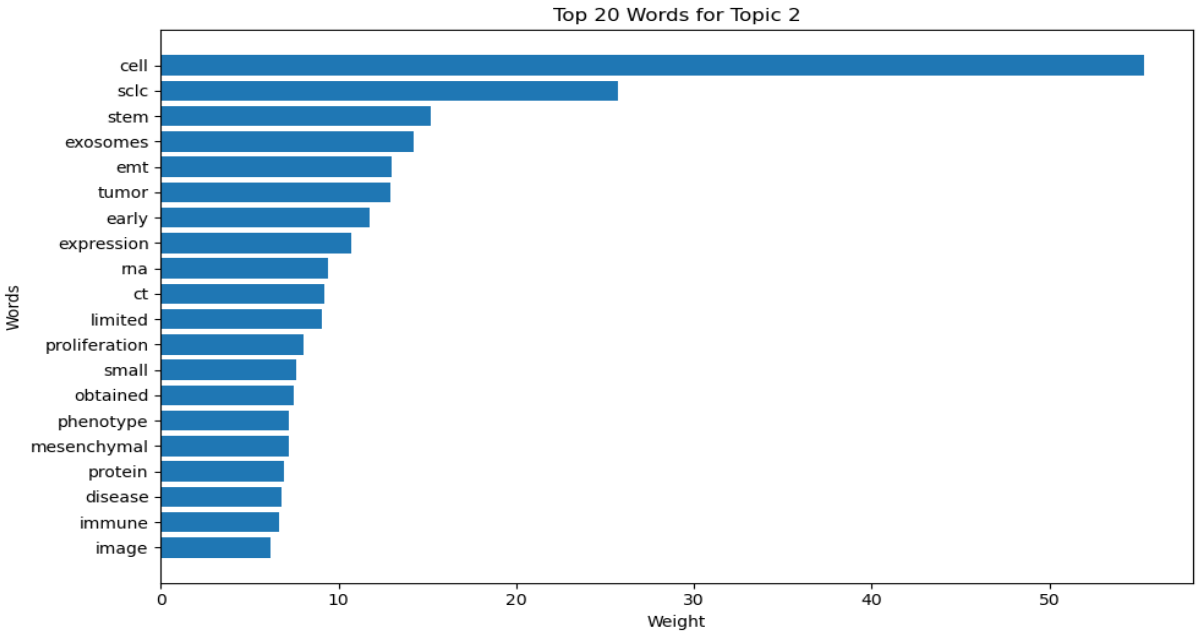
3.2. Findings Related to Topic-2

When the words addressed in Topic 2 are evaluated, the word cloud shown in Figure 6 emerges. Additionally, when the frequencies of the words are assessed, the frequency distribution presented in Figure 7 becomes apparent.

Figure 6. Word Cloud Related to Topic-2 as a Result of the Analysis of Documents

The word cloud, generated based on the lexicon of the sources related to Topic 2, is presented in Figure 6. It is evident that the size of each word in the cloud reflects its frequency of occurrence in the dataset. Upon analyzing the word cloud, the term "cell" is identified as the most frequent, followed by "SCLC", "stem," and "exosomes," in that order. Additionally, Figure 7 displays the 20 most frequent words for Topic 2, providing further insight into the key terms and themes related to this topic.

Figure 7. 20 Words with the Highest Frequency Related to Topic 2



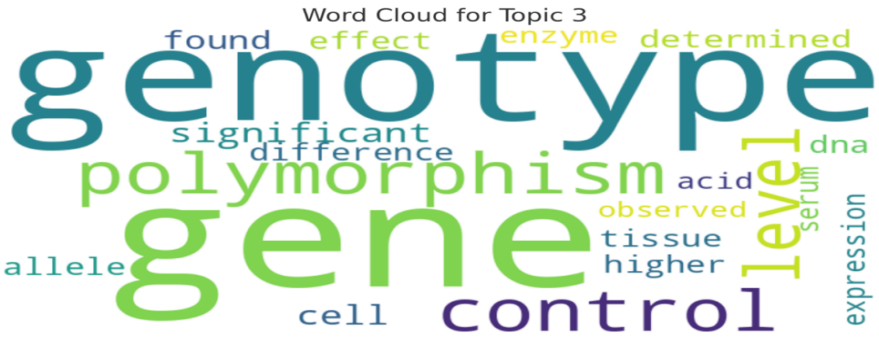
Upon examining Figure 7, the words are listed as follows: "cell," "SCLC", "stem," "exosomes," "EMT", "tumor," "early," "expression," "MA," "CT," "limited," "proliferation," "small," "contained," "phenotype," "mesenchymal," "protein," "disease," "immune," and "image." Additionally, the table presents the corresponding weights for each

word, which are determined based on their frequency values, offering further insight into the relative significance of each term within the topic.

3.3. Findings Related to Topic-3

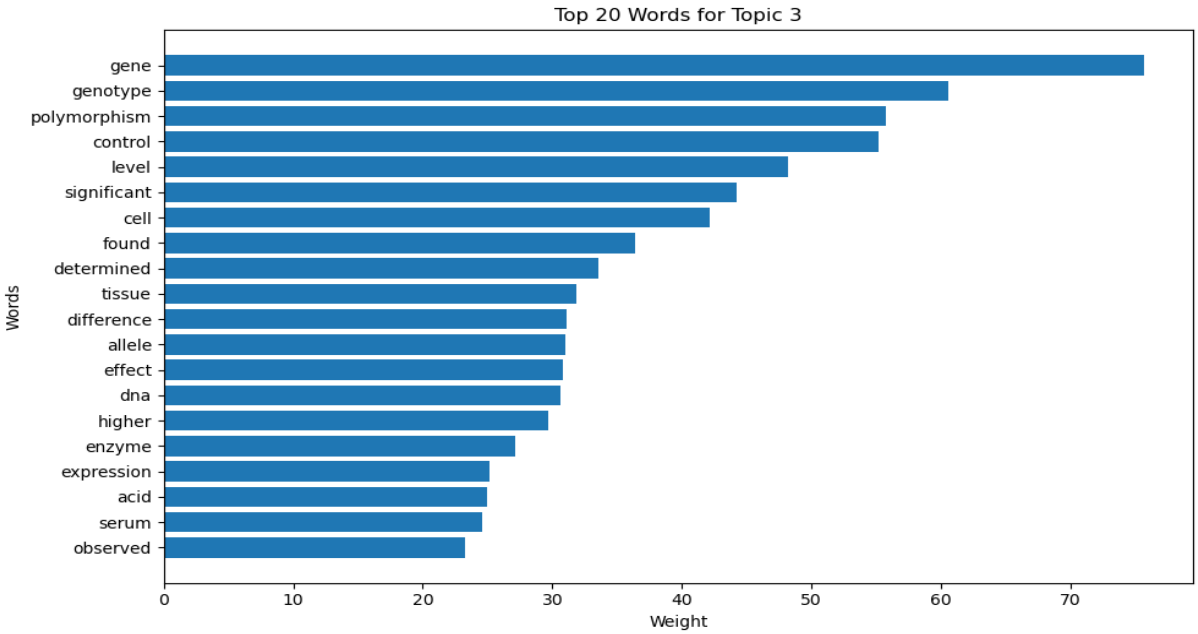
When the words addressed in Topic 3 are evaluated, the word cloud shown in Figure 8 emerges. Additionally, when the frequencies of the words are assessed, the frequency distribution presented in Figure 9 becomes apparent.

Figure 8. Word Cloud Related to Topic-3 as a Result of the Analysis of Documents



The word cloud, as determined by the lexicon of the sources defined for Topic-3, is illustrated in Figure 8. It is evident that the dimensions of the text are indicative of the frequency with which a word is used. When the word cloud in the figure is subjected to analysis, the word with the highest frequency is identified as "gene," i.e., the word "gene." Following the word "gene" are the words "genotype," "polymorphism," and "control," respectively. Furthermore, Figure 9 presents the 20 most frequently occurring words for Topic 3.

Figure 9. The 20 words with the highest frequency for Topic 3



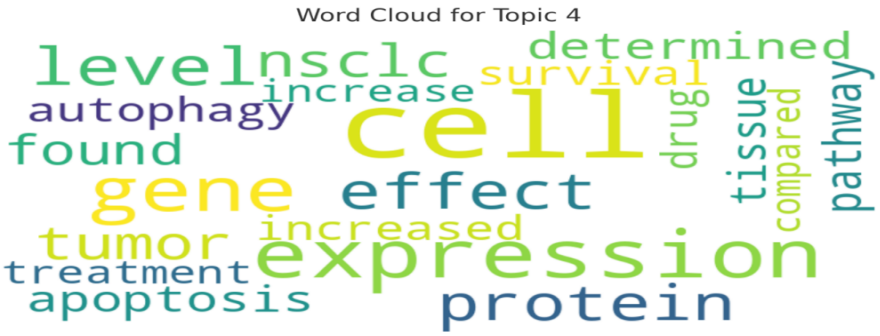
Upon examining the word frequencies in Figure 9, the words are ranked in the following order: "gene," "genotype," "polymorphism," "control," "level," "significant," "cell," "found," "determined," "tissue," "difference," "allele," "effect," "DNA," "higher," "enzyme," "expression," "acid," "serum," and "observed." The table also displays the weights of the words.

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

3.4. Findings Related to Topic-4

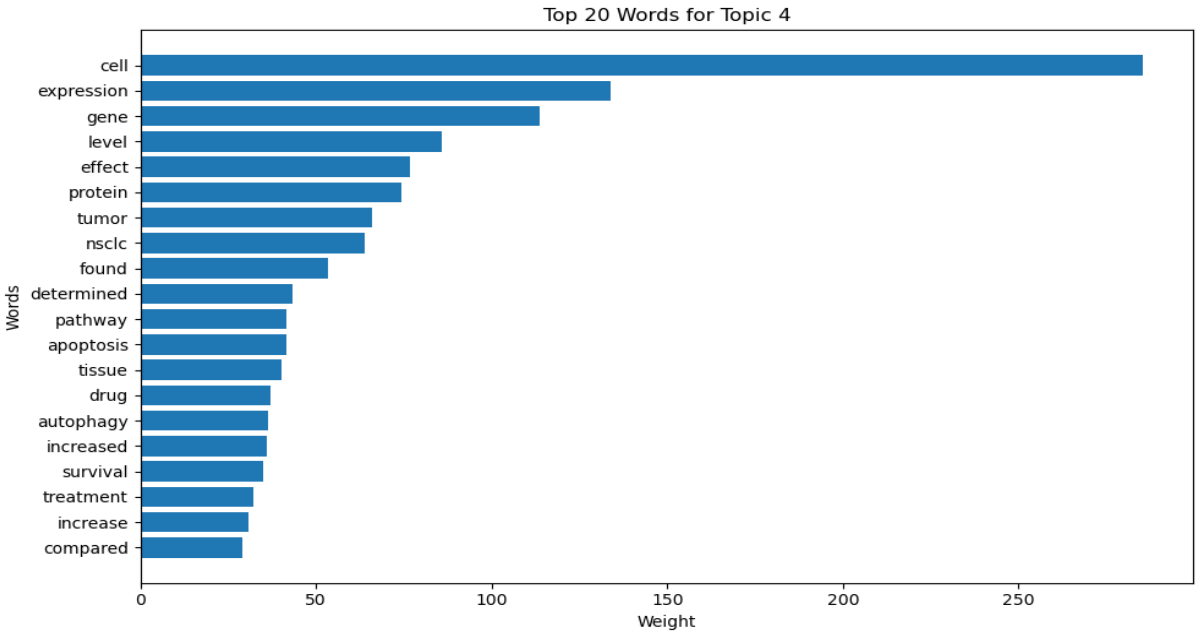
When the words addressed in Topic 4 are evaluated, the word cloud shown in Figure 10 emerges. Additionally, when the frequencies of the words are assessed, the frequency distribution presented in Figure 11 becomes apparent.

Figure 10. Word Cloud Related to Topic-4 as a Result of the Analysis of Documents



The word cloud, as determined by the lexicon of the sources defined for Topic 4, is illustrated in Figure 10. It is evident that the text size serves as an indicator of word frequency. A subsequent analysis of the word cloud in the figure reveals that the word with the highest frequency is "cell." Following the word "cell" are the words "expression," "gene," and "level," respectively. Furthermore, Figure 11 presents the 20 most frequently occurring words for Topic 4.

Figure 11. 20 Words with the Highest Frequency Related to Topic 4



Upon examining Figure 11 and analyzing the words related to Topic 4, the following ranking emerges: "cell," "expression," "gene," "level," "effect," "protein," "tumor," "NSCLC" (non-small cell lung cancer), "found," "determined," "pathway," "apoptosis," "tissue," "drug," "autophagy," "increased," "survival," "treatment," "increase," and "compared." The table also displays the weights of the words.

Süleyman Demirel Üniversitesi Vizyoner Dergisi, Yıl: 2025, Cilt: 16, Sayı: 48, 1401-1418.
Süleyman Demirel University Visionary Journal, Year: 2025, Volume: 16, No: 48, 1401-1418.
ISSN: 1308-9552

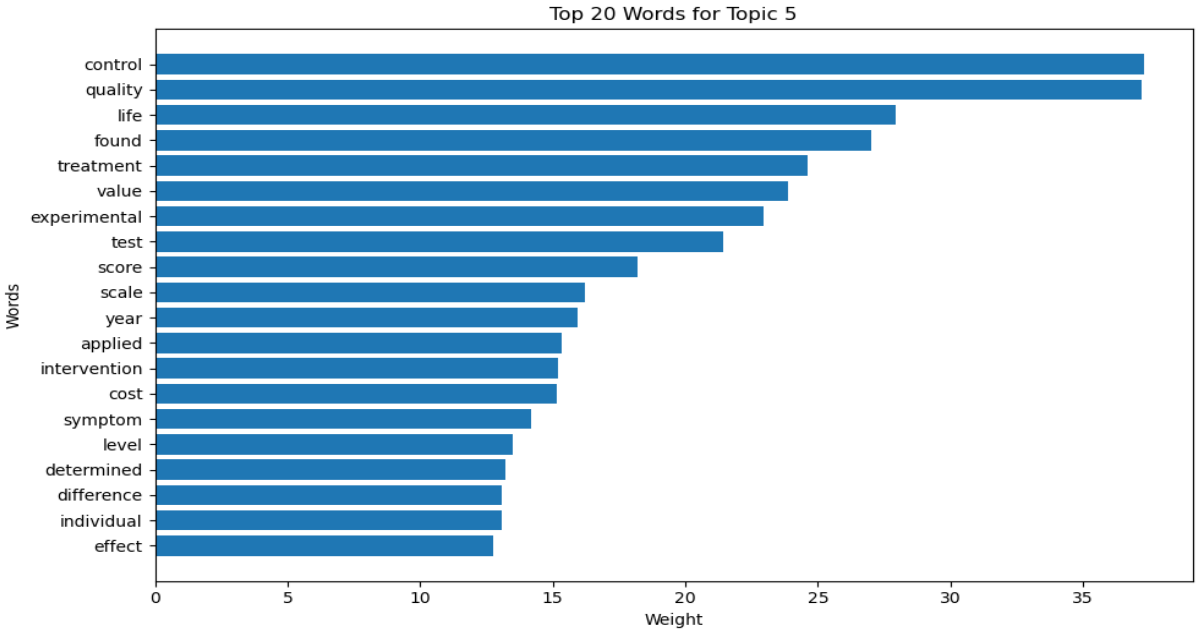
3.5. Findings Related to Topic-5

When the words addressed in Topic 5 are evaluated, the word cloud shown in Figure 12 emerges. Additionally, when the frequencies of the words are assessed, the frequency distribution presented in Figure 13 becomes apparent.

Figure 12. Word Cloud Related to Topic-5 as a Result of the Analysis of Documents

The word cloud, as determined by the lexicon of the sources defined for Topic 5, is illustrated in Figure 12. It is evident that the text size serves as an indicator of word frequency. A subsequent analysis of the word cloud reveals that the word with the highest frequency is "control." The words 'quality,' 'life,' and 'found' follow 'control' in terms of frequency. Furthermore, Figure 13 presents the results of the 20 words with the highest frequency for Topic 5.

Figure 13. 20 Words with the Highest Frequency Related to Topic 5



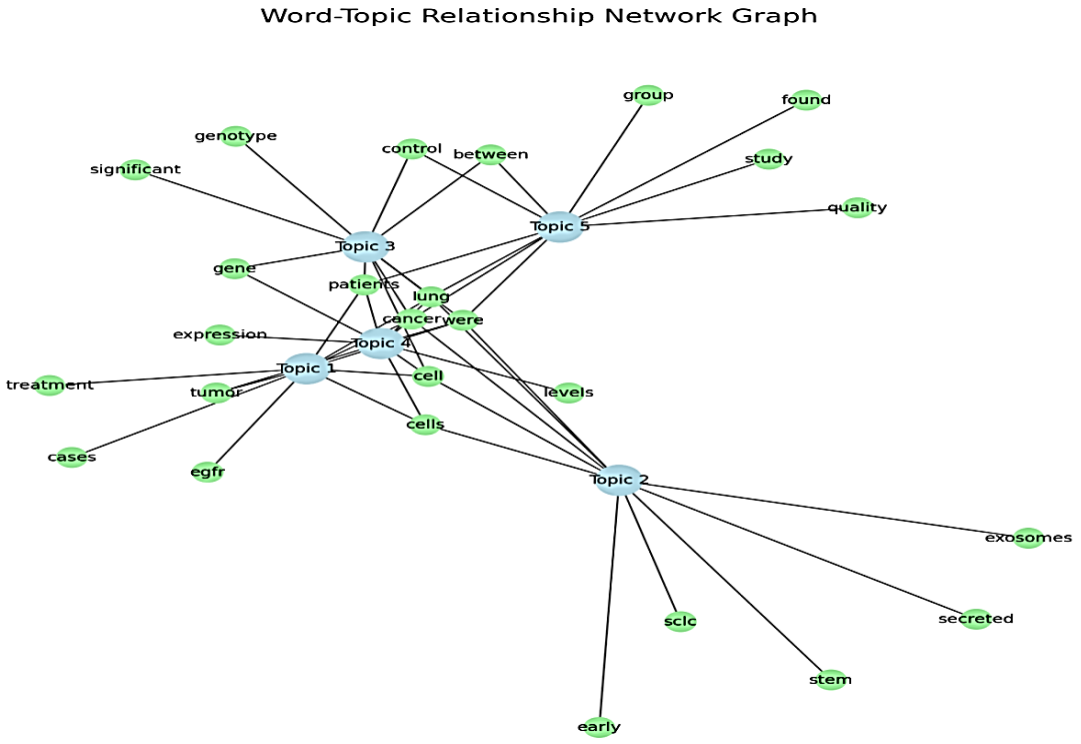
Upon examining Figure 13, the words listed in the table are ranked as follows: "control," "quality," "life," "found," "treatment," "value," "experimental," "test," "score," "scale," "year," "applied," "intervention," "cost," "symptom," "level," "determined," "difference," "individual," and "effect." The table also displays the weights of the words.

A close examination of the highest-frequency words in Topic 5 reveals the prominence of the terms "control," "quality," and "life." This finding suggests a direct correlation between the topic and health management, with a particular emphasis on health-related quality of life (HRQoL) and disease control. Such terms are congruent with health management objectives, including the monitoring of clinical outcomes, the evaluation of treatment effectiveness, and the improvement of care processes. Moreover, the repeated utilisation of the term "found"

signifies that this subject is being approached within the framework of research findings and evidence-based assessments as presented in theses. Consequently, Topic 5 exhibits a notable correlation with policies and methodologies for evaluating patient care quality and healthcare performance.

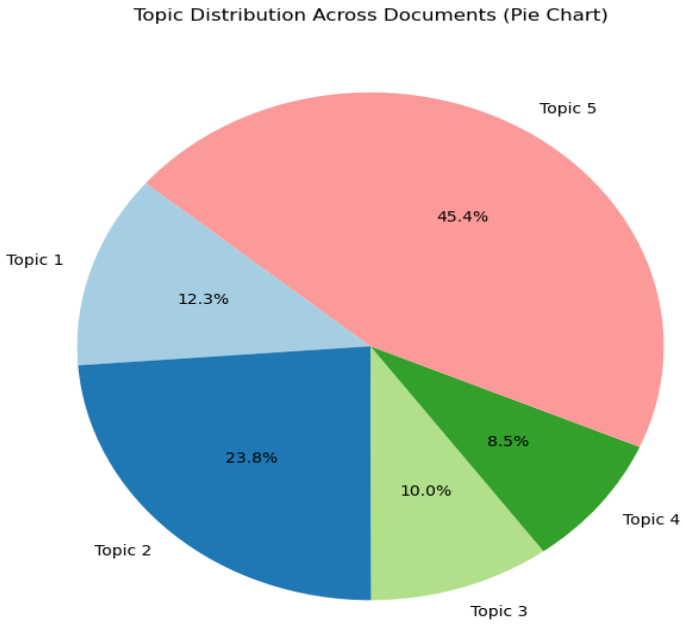
To examine the relationships between the five topics used in the study, a word-topic relationship network was analyzed using a software package. The resulting graph of the word-topic relationship network is presented in detail in Figure 13.

Figure 14. Word-Subject Relationship Network Graphic for Documents



Word-topic relationship network graphs are employed to visually represent the connections between different topics and the associated keywords in text mining analysis. Based on the word-topic relationship network, Topic-4 occupies a central position with the strongest connections to other topics. This centrality can be attributed to the presence of key terms such as "gene," "expression," and "cell" within Topic-4, which are shared across multiple topics. In contrast, Topics 1 and 2 exhibit relatively weaker interconnections, though their focus on more specialized areas within their respective domains is notable. For example, terms like "sclc" and "exosomes" play a prominent role in shaping the content of these topics. Further analysis of the word-subject relationship network highlights a strong association between the word "cell" and both Topics 1 and 4, suggesting that these topics are largely centered around biological themes. Additionally, the terms "cancer" and "lung" reinforce the concentration of cancer-related themes within Topic-4. Meanwhile, the terms "quality" and "found" suggest that Topic-5 is primarily concerned with issues surrounding the quality of life and overall health. The central positioning of Topic-4, along with its ability to attract numerous keywords and links, emphasizes its key role in the network, whereas the other topics are more focused on distinct and specialized areas.

Figure 15. Percentage of Subject Distribution Among Documents



When the distribution percentages of the topics across the documents are analyzed in Figure 15, it is observed that Topic 1 represents 12.3%, Topic 2 represents 23.8%, Topic 3 represents 10%, Topic 4 represents 8.5%, and Topic 5 represents 45.4%. This distribution suggests that while Topic 5 is the dominant theme, other topics are also represented in the text.

4. DISCUSSION AND CONCLUSION

The results of this study are anticipated to serve as a valuable reference for students and thesis supervisors in the selection of future thesis topics and the identification of distinctive topics. The application of Latent Dirichlet Allocation (LDA) facilitates the identification of key topics and contributes to evidence-based studies.

A review of the abstracts of the theses indicates that topic modeling with Latent Dirichlet Allocation (LDA) analysis is an effective method for identifying the topics that are frequently discussed in the theses. The analysis yielded an optimal model comprising five topics. The analysis of the five topics identified reveals that Topic 1 focuses on "Treatment Approaches Based on Genetic Variations of Cancer." The second topic pertains to the "Biological Structure of Lung Cancer Cells," the third topic addresses the "Genetic Basis of Lung Cancer," the fourth topic is about the "Physiopathological Processes of Lung Cancer" and the fifth topic is related to the "Treatment Process of Lung Cancer and Its Impact on Quality of Life." The identification of these topics is predicated on the statistical analysis of frequently mentioned words across all identified topics. A comparison of the number of documents related to each topic reveals that the topic 'Treatment Process of Lung Cancer and its Impact on Quality of Life' dominated the discussions of the doctoral theses analyzed in the study. Future endeavors must focus on refining the model by extracting additional thesis abstracts and leveraging interpretable artificial intelligence methods to enhance the model's predictions. A comparative analysis of LDA with other topic modeling methods could provide further insights.

The prominent topics in doctoral theses are separated by LDA analysis. The LDA approach was employed to examine all theses on lung cancer in the National Thesis Centre database of the Council of Higher Education (YÖK). The analysis yielded five distinct topics and keywords related to each topic. The topics were assigned word names, with keywords given priority. The keywords within these topics were then subjected to a thorough evaluation by the researchers, who classified them into relevant headings and conducted quantitative analyses. The analysis of the subject summaries of the doctoral theses on the YÖK-TEZ website revealed five predominant themes. The planning of the subject status and objectives of the documents revealed values of 0.21 in Topic 1, "0.07" in Topic 2, "0.21" in Topic 3, "0.38" in Topic 4, and 0.11 in Topic 5.

The analysis of the five identified topics reveals that Topic 1 focuses on *"Treatment Approaches Based on Genetic Variations of Cancer"* In this context, the doctoral theses "Investigation of Oncogene Changes in Peripheral Blood cfDNA Specimens After First-Line Therapy of Lung Adenocarcinoma Patients" and "Investigation of the Effect of EGFR on PD-L1 and Related Pathways in Non-Small Cell Lung Cancer Cells" serve as concrete exemplars representing genetically based treatment approaches. The second topic, entitled "*Biological Structure of Lung Cancer Cells*" is clearly related to the field, as demonstrated by the following papers: "The Role of HGF/c-Met Signal Transduction Pathway in Non-Small Cell Lung Cancer" and "Investigation of Expression and Methylation Profiles of KIFC1 Protein in Patients with Lung Cancer." The third topic focuses on the "*Genetic Basis of Lung Cancer*" and is illustrated by "Critical Gene Variants Related to PD-1/PD-L1 Signal Path in Non-Small Cell Lung Cancer and Investigation of Related Gene" and "Wide-Spectrum Analysis of KRAS and NRAS Mutations in Lung Cancers." The fourth topic focuses on the "*Physiopathological Processes of Lung Cancer*" and includes two case studies. The first is entitled "The Effect of CXCL7-Mediated Neutrophil Infiltration on Tumor Progression and Immunology in an Experimental Lung Adenocarcinoma Model", and the second is entitled "Investigation of Non-Small Cell Lung Cancer Patient Samples Using Advanced Proteomic Analysis Methods". Finally, Topic 5 addresses the "*Treatment Process of Lung Cancer and Its Impact on Quality of Life*" with "Cost Effectiveness Analysis of Pemetrexed and Gemcitabine Treatment for Advanced Non-Small Cell Lung Cancer in Türkiye" and "The Effect of Acupressure on Quality of Life and Dyspnea Level of Patients with Lung Cancer" serving as typical examples. The identification of these subjects is founded upon a statistical analysis of frequently occurring keywords within each theme. The sample thesis titles demonstrate that the themes derived from LDA align closely with the actual research focus.

The findings of this study, which identified five thematic clusters in Turkish doctoral theses on lung cancer using Latent Dirichlet Allocation (LDA), are consistent with prior applications of topic modelling in health-related domains. In a manner analogous to Altıntaş et al. (2021), who implemented LDA to analyse cancer-related discourse on social media, the present study demonstrates the efficacy of the method in systematically extracting predominant thematic domains. However, in contrast to the short-form user-generated texts utilised in their study, the present analysis employed long-form, peer-reviewed doctoral theses. This methodological difference ensures a more robust and nuanced thematic structure, along with more stable topic distributions.

In the context of bibliometric research, studies such as Ekinci et al. (2020) have employed LDA on large-scale biomedical literature to reveal emerging themes in medical research. The findings of this study are consistent with those of previous research in demonstrating the capacity of LDA to identify thematic trends. However, the present study's emphasis on lung cancer in the Turkish doctoral thesis corpus addresses a lacuna in national-level thematic mapping. To date, bibliometric studies have predominantly focused on journal articles or international corpora.

Moreover, the predominance of Topic 5: Treatment Process and Its Impact on Quality of Life is in alignment with qualitative studies in oncology (Schellekens, 2017; Criswell, 2012), which underscore the significance of symptom management, patient-centred care, and psychosocial well-being as pivotal elements of lung cancer care. This thematic overlap indicates that the results of LDA topic modelling on academic theses demonstrate parallels with the key concerns identified in qualitative studies on patient experience. This shows that there is convergence of findings, even though the present study did not conduct qualitative interviews directly.

From a methodological standpoint, while Van Nieuwenhove (2017) highlighted stability and parameter sensitivity as limitations of LDA, this study mitigated such concerns through parameter tuning, consistency checking, and expert validation of topic labels. Nevertheless, future work could incorporate alternative models (e.g., Labeled LDA, NMF) for comparison, thereby aligning with recommendations from previous comparative topic modelling research.

In conclusion, this research validates LDA as a tool for thematic extraction in long-form academic texts and also bridges computational text mining with insights typically emerging from qualitative oncology studies. It contributes a novel, nationally focused bibliometric perspective to the lung cancer literature.

A thorough analysis of the initial four subjects reveals that the research on lung cancer is predominantly clinical and medical in nature, with findings being interpreted within the medical context. In Topic 5, the future is predicted in the field of Health Management and Health Economics, where there are mainly general findings. The utilization of keywords related to emerging topics facilitates the direction of research towards the "Treatment Process of Lung Cancer and the Effect of Quality of Life," as evidenced by the most cited research published at both the national

and international levels. A more thorough examination of the keywords "control, quality, life, cost, cost, found, treatment, value" in lung cancer under Topic 5 illuminates the focal points within the domain of Health Economics. A close examination of these keywords reveals that Topic 5 encompasses a range of themes, including quality of life measurements, diagnosis and treatment processes, cost effectiveness, and general health. Consequently, countries can identify solutions and health policies for health economics and health technology assessment (HTA) by leveraging these keywords.

LDA analysis facilitates the comparison of articles and studies from disparate databases (e.g., PubMed, Scopus, Web of Science, etc.). In accordance with the aforementioned analyses, the subject density of field theses conducted in Türkiye can be ascertained. This analysis can serve as a guide for researchers in selecting suitable topics. In this context, effective inferences can be made by determining the importance levels of the labels emerging in the LDA. Additionally, the following topics may be suggested for future research and evaluations related to the subject:

- ***In-Depth Analysis of the Distribution of Topics in Theses:*** Future studies may involve a detailed examination of the topics and subtopics of doctoral theses on lung cancer. The areas where topics are most concentrated (such as genetics, treatment methods, quality of life, biomarkers, etc.) can be analyzed, highlighting how these areas align with current research trends. This can help identify gaps in the field for future research.
- ***Exploring the Connections Between Treatment Methods and Molecular Research:*** Research could be conducted to investigate the frequency of studies related to lung cancer treatment methods (such as surgery, chemotherapy, immunotherapy, etc.) in doctoral theses in Türkiye. Furthermore, numerical analyses of theses focused on molecular-level research, genetic mutations, and biomarkers could be undertaken. This would provide insight into which treatment methods are more extensively researched and which genetic targets are emphasized.
- ***Research on Social and Psychological Aspects:*** Studies could be carried out to increase research on the social and psychological dimensions of lung cancer. In addition to biological and treatment-focused research, it is important to include studies related to patients' quality of life, psychological conditions, and societal impacts in doctoral theses. Increasing the number of theses on these topics would allow for a more holistic approach to patient care.
- ***Developing Data Mining and Text Analysis Methods:*** Further work could be done to improve text mining and topic modeling methods used in the analysis of doctoral theses on lung cancer. A more detailed examination of these techniques, such as how they can be used to conduct in-depth analyses of research and enhance the reliability and validity of research outcomes, would be beneficial.
- ***Comparisons with National and International Literature:*** A comparison could be made between doctoral theses on lung cancer in Türkiye and similar studies in the international literature. This would help assess the position of Turkish research within the global context. Additionally, it would be possible to draw conclusions about how factors unique to Türkiye, such as geographical, cultural, or health system-specific factors, play a role in the content of these theses.
- ***Research on Educational and Policy Recommendations:*** Based on the findings of research derived from doctoral theses in Türkiye, suggestions could be made regarding healthcare policies and educational programs. For instance, recommendations could be developed to increase government funding for lung cancer research, strengthen scholarships, and enhance project support for doctoral students.
- ***Establishing Directions for Future Research:*** Drawing from the findings of doctoral theses on lung cancer in Türkiye, recommendations could be made on which research areas should be prioritized in the future. For example, more research could be emphasized on new treatment methods, the integration of biomarkers into clinical applications, genetic analyses, and personalized treatment planning.
- ***Promoting Multidisciplinary Research:*** It could be suggested to further diversify doctoral theses on lung cancer through interdisciplinary approaches encompassing biological, clinical, social, and psychological fields. The importance of a multidisciplinary approach in the treatment process of cancer could be highlighted, and evidence could be presented showing how this approach has positive effects on patients' overall well-being.

DECLARATION OF THE AUTHORS

Declaration of Contribution Rate: The authors have equal contributions.

Declaration of Support and Thanksgiving: No support is taken from any institution or organization.

Declaration of Conflict: There is no potential conflict of interest in the study.

REFERENCES

Altıntaş, V., Albayrak, M., & Topal, K. (2021). Kanser hastalığı ile ilgili paylaşımlar için dirichlet ayrımı ile gizli konu modelleme. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, *36*(4), 2183-2196.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, *11*(2-3), 143-296.

Budak, İ. (2024). Labeling of European environment agency waste and recycling reports with LDA analysis. In *Technical landfills and waste management: Volume 2: Municipal solid waste management* (p. 285-294). Springer Nature Switzerland.

Criswell, K. R. (2012). *A qualitative study of psychosocial needs for ındividuals with lung cancer.* [Doctoral Dissertation]. Loma Linda University.

Ekinci, E., Omurca, S. İ., Kırık, E., & Taşçı, Ş. (2020). Tıp veri kümesi için gizli dirichlet ayrımı. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *22*(64), 67-80.

Fadaka, A., Ajiboye, B., Ojo, O., Adewale, O., Olayide, I., & Emuowhochere, R. (2017). Biology of glucose metabolization in cancer cells. *Journal of Oncological Sciences*, *3*(2), 45-51.

Feldman, R. & Sanger J. (2007). *Text mining handbook*. Cambridge Universıty Press.

Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., ... & Rosell, R. (2015). Non-small-cell lung cancer. *Nature Reviews Disease Primers*, *1*(1), 1-16.

Güneş, A., & Yıldırım, B. (2022). *Eğitimde metin madenciliği ve uygulamaları.* Eğitim Yayınevi.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (p. 3-10). https://aclanthology.org/P99-1001.pdf

Hoy, H., Lynch, T., & Beck, M. (2019). Surgical treatment of lung cancer. *Critical Care Nursing Clinics*, *31*(3), 303-313.

Karakuş, L. (2021). *Eğitimde Metin Madenciliği: Türkçe metinlerde sözlük tabanlı duygu analizi*. [Master's Thesis]. Akdeniz Üniversitesi.

Kennedy, K., Hulbert, A., Pasquinelli, M., & Feldman, L. E. (2022). Impact of CT screening in lung cancer: Scientific Evidence and Literatüre Review. In *Seminars in Oncology* (p. 198-205). WB Saunders

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling. In *Computational methods for communication science* (p. 13-38). Routledge.

Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on lung cancer. *Cancer cell*, *1*(1), 49-52.

Morgan, D. L. (1996). *Focus groups as qualitative research*. Sage Publications.

Nasution, M. K. (2017). *Penelaahan literatur*. Teknik Penulisan Karya Ilmiah.

Schellekens, M. P. J. (2017). *Psychological distress in lung cancer: Mindfulness-based stress reduction for patients and partners*. [Doctoral Dissertation]. Utrecht University.

Sugiantoro, B., Humam, A. I., Fitriyani, N. L., Alfian, G., Maarif, M. R., & Syafrudin, M. (2023). *Utilizing latent dirichlet allocation for analyzing topics in undergraduate theses*. In 2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) (pp. 121-126). IEEE.

Van Nieuwenhove, Z. (2017). *Analysis of latent dirichlet allocation* [Ph.D. Dissertation]. Tilburg University.

Yeşilmen, S. (2024). *Synthesıs of dye-conjugated PD-L1 targeted peptıdes for use in PET/CT in dıagnosıs of lung cancer* [Master's Thesis]. İstanbul Technical Universitesi.