APPLICATION OF TEXT MINING TO TECHNOLOGY MANAGEMENT DOMAIN TO EXTRACT TOPICS AND TRENDS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YAŞAR TEKİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR PHILOSOPHY
IN
THE DEPARTMENT OF SCIENCE AND TECHNOLOGY POLICY STUDIES

JANUARY 2022

Approval of the thesis:

APPLICATION OF TEXT MINING TO TECHNOLOGY MANAGEMENT DOMAIN TO EXTRACT TOPICS AND TRENDS

submitted by **YAŞAR TEKİN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Science and Technology Policy Studies, the Graduate School of Social Sciences of Middle East Technical University** by,

Prof. Dr. Yaşar KONDAKÇI
Dean
Graduate School of Social Sciences _____

Prof. Dr. Mehmet Teoman PAMUKÇU
Head of Department
Department of Science and Technology Policy Studies _____

Prof. Dr. Pınar KARAGÖZ
Supervisor
Department of Computer Engineering _____

Examining Committee Members:

Prof. Dr. M. Teoman PAMUKÇU (Head of the Examining Committee)
Middle East Technical University
Department of Science and Technology Policy Studies _____

Prof. Dr. Pınar KARAGÖZ (Supervisor)
Middle East Technical University
Department of Computer Engineering _____

Prof. Dr. Nursal ARICI
Gazi University
Department of Computer Engineering _____

Prof. Dr. Ahmet COŞAR
Çankaya University
Department of Computer Engineering _____

Prof. Dr. İ. Hakkı TOROSLU
Middle East Technical University
Department of Computer Engineering _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Yaşar TEKİN

Signature:

ABSTRACT

APPLICATION OF TEXT MINING TO
TECHNOLOGY MANAGEMENT DOMAIN
TO EXTRACT TOPICS AND TRENDS

TEKİN, Yaşar
Ph.D., The Department of Science and Technology Policy Studies
Supervisor: Prof. Dr. Pınar KARAGÖZ

January 2022, 120 pages

Topic modeling is a widely used technique to extract latent topics from large document collections. One of the most remarkable uses of it is its application to scientific fields. If topic modeling is applied to all articles published in a specific scientific field, it provides an overall view of topics and trends for the time period under consideration. If it is applied to a single conference or journal, it reveals differences from global trends.

The most popular method used for topic modeling is Latent Dirichlet Allocation (LDA). Although LDA is used in many different fields, the problems of how to optimize model parameters and how to eliminate topic instability have not been fully solved yet.

This thesis consists of two main parts: 1) An empirical investigation is conducted: a) to investigate the level of topic instability in ordered documents, b) to search for methods to eliminate (if not possible, to alleviate) the effects of the topic instability, c) to evaluate the use of word vector representations to optimize LDA parameters. It

is found out that: a) the level of instability is high even in ordered documents, b) average scores of replicated topic models can be used to alleviate the effects of topic instability, c) Skip-gram similarity score is an acceptable measure in optimizing LDA parameters.

2) By using the method proposed, topic modeling is applied to Technology Management (TM) domain. Top topics, the most studied industries, the most used methods and surprising topics of TM literature are identified.

Keywords: Technology Management, Topic Modeling, Latent Dirichlet Allocation, Parameter Optimization, Word Vector Representation.

ÖZ

KONULARIN VE EĞİLİMLERİN BULUNMASI AMACIYLA METİN MADENCİLİĞİNİN TEKNOLOJİ YÖNETİMİ ALANINA UYGULANMASI

TEKİN, Yaşar
Doktora, Bilim ve Teknoloji Politikası Çalışmaları Bölümü
Tez Yöneticisi: Prof. Dr. Pınar KARAGÖZ

Ocak 2022, 120 sayfa

Konu modelleme, büyük belge topluluklarındaki gizli konuların tespitinde yaygın olarak kullanılan bir yöntemdir. En dikkat çekici kullanımlarından birisi bilimsel alanlardaki uygulamalarıdır. Konu modelleme, bir bilimsel alanda yayınlanan tüm makalelere uygulandığında, incelenen dönem için konuların ve eğilimlerinin genel bir görünümünü ortaya koyar. Yalnızca bir konferans veya dergiye uygulandığında ise, söz konusu konferans veya derginin global eğilimlerden farklılıklarını ortaya çıkarır.

Konu modellemede kullanılan en popüler yöntem Gizli Dirichlet Ayrımı (GDA)'dır. Birçok farklı alanda kullanılıyor olmasına rağmen model parametrelerinin nasıl eniyileneceği ve kararsızlığın nasıl giderileceği soruları tam olarak yanıtlanabilmiş değildir.

Bu tez iki ana bölümden oluşmaktadır: 1) Birinci bölümde bir deneysel çalışma gerçekleştirilerek a) sıralı dokümanlarda kararsızlık seviyesi, b) kararsızlığı önlemek (mümkün değilse hafifletmek) için kullanılabilecek yöntemler ve c) GDA

parametrelerinin eniyilenmesinde sözcük vektör temsillerinin kullanımı araştırılmıştır. Araştırma sonucunda: a) sıralı dokümanlarda da kararsızlık seviyesinin yüksek olduğu, b) konu kararsızlığını hafifletmek için çoğaltılmış konu modellerinin ortalama puanlarının kullanılabileceği ve c) GDA parametrelerinin eniyilenmesinde Skip-gram benzerlik puanının kabul edilebilir bir ölçü olduğu tespit edilmiştir.

2) İkinci bölümde, ilk bölümde önerilen yöntem kullanılarak, konu modelleme Teknoloji Yönetimi (TY) alanına uygulanmıştır. TY literatürünün zirvedeki konuları, en çok çalışılan endüstri dalları, en çok kullanılan yöntemleri ve şaşırtıcı konuları belirlenmiştir.

Anahtar Kelimeler: Teknoloji Yönetimi, Konu Modelleme, Gizli Dirichlet Ayrımı, Parametre Eniyileme, Sözcük Vektör Temsili.

To My Family

ACKNOWLEDGMENTS

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

I would like to thank to the editors of the Technology Management specialty journals used in the thesis study for their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DE	Differential Evolution
EA	Evolutionary Algorithm
GA	Genetic Algorithm
LDA	Latent Dirichlet Allocation
MOT	Management of Technology
NPD	New Product Development
NPMI	Normalised Pointwise Mutual Information
PDMA	Product Development and Management Association
PMI	Pointwise Mutual Information
R&D	Research and Development
TIE	Technology/Innovation Management and Entrepreneurship
TIM	Technology Innovation Management
TM	Technology Management

CHAPTER 1

INTRODUCTION

Technology Management (TM) is a new scientific field that emerged at the second half of the 20th century. Different authors have used different names for TM, such as Management of Technology (MOT), Technology Innovation Management (TIM) and Technology/Innovation Management and Entrepreneurship (TIE). National Research Council (1987) defines TM as an emerging field that "links engineering, science, and management disciplines to plan, develop, and implement technological capabilities to shape and accomplish the strategic and operational objectives of an organization". Gregory (1995) proposes a process framework for TM which has the following activities: 1) Identification, 2) Selection, 3) Acquisition, 4) Exploitation, 5) Protection. Cetindamar et al. (2009a) proposes a framework which groups TM activities into two categories: 1) core activities (Identification, Selection, Acquisition, Exploitation, Protection, Learning) and 2) supporting activities (Knowledge Management, Project Management, Innovation Management).

As can be seen from the literature, TM is a new and interdisciplinary field whose definition and scope continue to be researched. The most difficult problem in an interdisciplinary field is to define its scope which also forms the basis of relations with other fields. Such a scope definition is important as, without this information, the scope of the research to be conducted is hard to determine. This information is also crucial for the editors to fully set the aim and scope of the conference/journal they manage.

One way of solving this problem is to use text mining (i.e., analysis to extract hidden patterns from the text). Topic modeling, a form of text mining, is a technique used to identify thematic structures in a corpus. Its purpose is to automatically extract latent document-topic and topic-term distributions from observable document-term

distributions. One of the most remarkable uses of topic modeling is its application to scientific fields. If topic modeling is applied to all articles published in a specific scientific field, it provides an overall view of topics and trends for the time period under consideration. If it is applied to a single conference or journal, topics and trends extracted from the conference/journal can be used by its editor to examine differences from global trends of research which are necessary to align the conference/journal with. In both cases, researchers can make use of this knowledge to plan their future studies.

Hence, the purpose of this study is to contribute to literature by revealing the topics studied in the TM domain. In order to achieve this goal, topic modeling is applied to a corpus of 16,180 research articles/abstracts published in 13 top specialty journals of TM between 1997 and 2019. The time interval is determined on the basis of the subscription periods of the journals.

Latent topics are extracted by using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the most popular topic model. Despite its broad use in many different fields, it still has two problems which are open to improvement: 1) how to optimize model parameters and 2) how to eliminate topic instability.

There are three parameters that need to be optimized in LDA: 1) the number of topics (K), 2) a Dirichlet prior on document-topic distributions (α) and 3) a Dirichlet prior on topic-word distributions (β) (Binkley et al., 2014). When the selected value for K is small, the range of topics becomes broader, and when it is large, the range becomes more specific. If the value given to α or β is small, only a few topics in a document or only a few words in a topic can have high probabilities. Conversely, when the value is large, all topics in a document or all words in a topic have similar probabilities.

Topic instability can be defined as obtaining different topics from the same corpus at each run. Mehta et al. (2014) explain the reason as the stochastic nature of LDA algorithm while Agrawal et al. (2018) explain it as order effect. Order effect is defined as the generation of different topics from different orders of input documents.

In this study, firstly, an empirical investigation is conducted to evaluate possible solutions to the aforementioned problems. Specifically, three subjects are investigated:

1) the level of instability in ordered documents, 2) methods to eliminate (if not possible, to alleviate) the effects of topic instability, 3) methods to optimize the LDA parameters. For this purpose, experiments are conducted on two corpora with different properties: the first being a corpus of news articles, and the second being a corpus of research articles. As a result of the experiments, it is found out that: 1) the level of instability is high even in time ordered documents like research articles, 2) average scores of replicated topic models can be used to alleviate the effects of topic instability, 3) semantic similarity score is a suitable measure in determining the parameters of LDA.

Then, by using the method proposed for parameter optimization, topic modeling is applied to Technology Management (TM) domain at two levels. Top topics, the most studied industries, the most used methods, and surprising topics of TM literature are identified. Trends of the topics are extracted on an annual basis which can capture annual fluctuations such as special issues and editor changes. The topics presented in this study can be thought as the main topics studied in the TM community and provide valuable information for editors, authors, and policy makers.

The editors of Technology Management specialty journals used in this study were informed about the use of full-text articles for text mining and no negative response was received.

CHAPTER 2

RELATED WORK

As stated above, LDA has two problems which are open to improvement: how to optimize model parameters and how to eliminate topic instability. In this section, the articles will be reviewed published both about LDA parameter optimization and topic instability. However, there is another subject that needs to be considered in the analysis of problems in topic modeling: topic coherence. It is a measure used to compare human interpretability of topics. For this reason, articles published about topic coherence are also reviewed.

Since TM is a new and interdisciplinary field, it is of utmost importance to determine topics studied in this field. However, to fulfill this, it is necessary to identify top specialty journals in TM. Therefore, in addition to the three subjects given above, we review the articles published about TM's top specialty journals and topics studied in TM.

Before going into the details of previous studies on the subjects, it would be appropriate to give a list of journal abbreviations to be used in the rest of the paper. Table 1 presents titles and abbreviations of journals used in this study.

Table 1. Journal Abbreviations

Name	**Abbreviation**
Engineering Management Journal	EMJ
Entrepreneurship Theory and Practice	ETP
European Journal of Innovation Management	EJIM
IEEE Transactions on Engineering Management	IEEETEM

Table 1 *(cont'd)*

Industrial and Corporate Change	ICC
Industry and Innovation	II
Innovation: Management, Policy & Practice (in 2018, renamed to Innovation: Organization & Management)	IMPP (IOM)
International Journal of Innovation Management	IJIM
International Journal of Technology Management	IJTM
Journal of Business Venturing	JBV
Journal of Engineering and Technology Management	JETM
Journal of High Technology Management	JHTM
Journal of Product Innovation Management	JPIM
Journal of Small Business Management	JSBM
Journal of Technology Transfer	JTT
R&D Management	RDM
Research Evaluation	RE
Research Policy	RP
Research Technology Management	RTM
Science and Public Policy	SPP
Small Business Economics	SBE
Technology Analysis & Strategic Management	TASM
Technological Forecasting & Social Change	TFSC
Technology Review	TR
Technovation	T

2.1. LDA Parameter Optimization

To the best of our knowledge, the first proposal to detect the optimum number of topics through LDA method is made by Griffiths and Steyvers (2004). The authors propose to keep the parameters α (50/K) and β (0.1) constant while changing the parameter K. They calculate posterior probability distributions of models for different values of K and select the model with the highest posterior probability.

Teh et al. (2006) introduce Hierarchical Dirichlet Process to model document collections with multiple groups. In this model, which is described as a distribution over a set of random probability measures, the number of topics is a random variable and any topic used by a group can be reused by the other groups.

Steyvers and Griffiths (2007) use generalization performance, namely perplexity from computational linguistics, as the selection criterion. The perplexity is related to entropy and calculated as the inverse probability which is normalized by the number of words (Jurafsky and Martin, 2008). However, in a later study, Chang et al. (2009) show that the perplexity is negatively correlated with human interpretability of topics.

Zavitsanos et al. (2008) employ an iterative process which increments the number of topics at each iteration. The process computes average symmetric KL-Divergence of topics at successive iterations and stops when the contribution of new topics added at the last iteration is sufficiently small.

Cao et al. (2009) propose to detect the optimum number of topics based on topic density which shows the number of similar topics. The aim of this method is to maximize intra-topic similarity while minimizing inter-topic similarity. It sequentially computes the model's cardinality, which is the number of topics whose densities are less than a given threshold and re-estimates the optimum number of topics based on the cardinality.

Arun et al. (2010) utilize a matrix factorization mechanism. The method factorizes the document-term matrix into document-topic and topic-term matrices and uses a new measure which is computed in terms of symmetric KL-Divergence.

Panichella et al. (2013) introduce LDA-GA which uses Genetic Algorithm (GA) to optimize the LDA parameters in software engineering tasks. The authors define the concept of dominant topic which is specific to software documents. It computes mean Silhouette coefficient of documents to measure the accuracy of LDA.

Mehta et al. (2014) propose to cluster topics from multiple models. The authors distinguish strong and weak topics by examining cluster quality. The optimum number

of topics is detected as the number of clusters that maximizes the aggregate mean silhouette.

Greene et al. (2014) employ a term-centric stability approach. The method uses a modified version of the Jaccard index, namely Average Jaccard Index, to measure the agreement between two topics. The stability for a given number of topics is measured by calculating the agreements between topics of a complete corpus and topics of its randomly selected subsets.

Agrawal et al. (2018) show that improving topic stability makes classification more accurate and therefore propose LDADE to find the parameters that make LDA more stable. The method uses Differential Evolution (DE) algorithm, which measures cross-run similarity that counts overlaps of top terms across replicated models, to optimize the LDA parameters.

Krasnov and Sen (2019) use Additive Regularization of Topic Models to separate noise topics defined as topics with uniform distributions across documents. The authors transform topics to dense vectors and use a new metric, namely Cosine Davies Bouldin Index which is a modified version of the cluster validation metric Davies Bouldin index, to find the optimum number of topics.

2.2. Topic Instability

Topic instability is another unresolved problem for LDA. As will be discussed in the following sections, the level of instability may be quite high in ordered documents. This shows that the problem is due to the stochastic nature of LDA model.

It is stated by Koltcov et al. (2014) that one promising solution is regularized topic models. However, it is also stated in the same study that this solution is not mature enough and requires further work.

An exact solution has not been able to be found yet but researchers suggest several strategies to alleviate its effects. Steyvers and Griffiths (2007) and Koltcov et al. (2014) select stable topics that reappear across replicated models. Greene et al. (2014) calculate the agreements between topics of a complete corpus and topics of its

randomly selected subsets. Mehta et al. (2014) and Mantyla et al. (2018) cluster topics from replicated models. Chuang et al. (2015) determine modeling consistency between replicated models. Lastly, Agrawal et al. (2018) count overlaps of top terms of topics from replicated models.

Actually, there is a single approach used by the researchers to alleviate the effects of topic instability: to generate multiple models, to process the models (such as clustering, comparing, etc.) and obtain the best model. The models can either be replicated models initialized by the same parameter values or different models obtained from different subsets of the same corpus. On the basis of the previous work, we believe that the replicated models are more feasible when compared to the models of subsets, which have the problem of selecting the right subsets.

2.3. Topic Coherence

The first studies in topic modeling used statistical methods to optimize the LDA parameters. However, in subsequent studies, it is determined that the statistical methods do not match the interpretability of topics and since then, the studies have focused on coherence of topics.

Chang et al. (2009) introduce the methods word intrusion and topic intrusion to measure the interpretability of a topic model. While word intrusion measures semantic coherence of topics by finding the word which does not belong with the others, topic intrusion measures decompositions of documents as a mixture of topics by finding the topic which does not belong with the document.

Newman et al. (2010) propose the task of topic coherence evaluation and compare a range of topic scoring models that use external text sources. They report that Pointwise Mutual Information (PMI) with Wikipedia is the single best-performing method among them. PMI of two words is calculated as given in Equation 1.

$$PMI(w_i, w_i) = \log \frac{p(w_i, w_j)}{p(w_i)p(\quad_j)} \qquad (1)$$

where w_i and w_j are among the top terms. In addition, the authors experiment with arithmetic mean and median to combine the PMI scores of word pairs and report that the median-based calculation produce slightly higher correlation values.

Mimno et al. (2011) develop another topic coherence measure (C) that aligns well with human judgments. This metric is calculated as given in Equation 2.

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(w_m^{(t)}, w_l^{(t)}\right) + 1}{D\left(w_l^{(t)}\right)} \qquad (2)$$

where $V^{(t)} = \left(v_1^{(t)}, \ldots, v_m^{(t)}\right)$ are the top terms of topic t and D is the document or co-document frequency of words. Unlike the others, this method gathers statistics from the corpus to be modeled.

Stevens et al. (2012) assess the quality of a topic model by measuring: 1) average coherence of all topics, and 2) entropy of the coherence for all topics. They report that the methods PMI and C they use for their evaluations often produce similar results.

Rosner et al. (2013) show that pairs of individual words can lead to poor results in measuring topic coherence and propose to score pairs of more complex word subsets.

Aletras and Stevenson (2013) propose to use distributional semantic similarity methods to measure the coherence of topics. They use Wikipedia as the reference corpus and PMI and Normalized PMI (NPMI) as weights of words. NPMI is computed as given in Equation 3.

$$NPMI\left(w_i, w_j\right) = \frac{PMI\left(w_i, w_j\right)}{-log\left(p(w_i)p\left(w_j\right)\right)} \qquad (3)$$

The authors report that distributional semantic similarity outperforms PMI and C metrics.

Chuang et al. (2013) observe that a small change in β of LDA can significantly alter the ratio of coherent topics and, in many cases, increasing the number of topics lead to more useless topics.

Röder et al. (2015) develop a framework with dimensions: 1) kind of segmentation of word sets, 2) confirmation measures, 3) word probabilities, and 4) aggregation methods. They obtain higher correlations with human scores using: 1) probabilities derived from Wikipedia, 2) 10 top terms, 3) NPMI, 4) boolean sliding window, and 5) arithmetic mean. However, the highest correlation is reached by using a new combination of the coherence measures: indirect cosine measure with NPMI and the boolean sliding window.

Nikolenko (2016) evaluates the use of distributed word representations with four different distance metrics: 1) cosine distance, 2) L1-distance, 3) L2-distance, and 4) coordinate distance. The author uses Skip-gram model to create vector representations of words. He reports that there are little differences between the new metrics but they outperform, often significantly, previously known topic evaluation approaches.

2.4. TM's Top Specialty Journals

In 1993, a subjective survey was conducted to establish a hierarchical rating of journals in the field of TIM (Liker, 1995). Top 10 journals of the TIM specialty area were given as IEEETEM, JETM, RP, JHTM, JPIM, RTM, IJTM, TASM, TFSC and TR.

Cheng et al. (1999) collected citation data from articles published between 1990 and 1994 in the top five TIM specialty journals in the Liker's study (Liker, 1995). The authors named three journals (IEEETEM, RP, JPIM) as the premium TIM specialty journals due to their high scores across all methods.

Linton and Thongpapanl (2004) used 10 TIM specialty journals (IEEETEM, RTM, TFSC, RDM, RP, T, JETM, JPIM, IJTM, TASM) as base journals for their citation analysis. Relative ranking of top 10 specialty journals in TIM were determined as follows: JPIM, RP, RTM, RDM, IEEETEM, TFSC, IJTM, T, TASM and JETM.

Merino et al. (2006) analyzed T for years between 1981 and 2004 and selected 10 journals (IEEETEM, RTM, TFSC, RDM, RP, T, JETM, JPIM, IJTM, TASM) as the most relevant journals in TIM.

Ball and Rigby (2006) selected 11 journals (EJIM, IJIM, JPIM, JETM, TASM, T, RDM, RP, IJTM, IEEETEM, RTM) to study the management of research and development, innovation, and technology. The authors analyzed author participation, affiliation, and the ratios of academic authors to non-academic authors.

Linton and Embrech (2007) gave 2006 modified impact factors of 10 TIM specialty journals (IEEETEM, IJTM, JETM, JPIM, RDM, RP, RTM, TFSC, TASM, T). The authors used a self-organizing map to identify proximities and distances between journals.

Franke and Schreier (2008) proposed a new ranking which combined existing TIE journal rankings into an integrative meta-ranking. As a result of their work, they labeled three journals (RP, JPIM, JBV) as Category A, seven journals (IEEETEM, RDM, SBE, ETP, TFSC, JSBM, IJTM) as Category B, 12 journals as Category C and 21 journals as Category D.

Linton (2009) found that average impact and change of impact of TIM journals (RP, JPIM, RDM, T, TFSC, IEEETEM, JETM, TASM, RTM, IJTM) had increased between 2006 and 2008.

Cetindamar et al. (2009b) analyzed articles published in 10 TM journals (JPIM, RP, RTM, RDM, IEEETEM, TFSC, IJTM, T, TASM, JETM) between 1995 and 2005 to understand similarities and differences between developing and developed countries.

Beyhan and Cetindamar (2011) identified intellectual structures of developing countries' TM literatures by analyzing articles published between 1998 and 2007 in 10 TIM specialty journals (JPIM, RP, RTM, RDM, IEEETEM, TFSC, IJTM, T, TASM, JETM).

Thongpapanl (2012) gave an up-to-date ranking of TIM specialty journals. The author analyzed citation data of 15 base journals (JPIM, RP, RTM, RDM, IEEETEM, TFSC,

IJTM, T, TASM, JETM, EMJ, JTT, SPP, ICC, II) published between 2006 and 2010. The top 10 specialty journals were determined as RP, JPIM, RTM, T, RDM, ICC, IEEETEM, JTT, TFSC and JETM.

Choi et al. (2012) studied national characteristics of MOT research in 10 technology and innovation specialty journals (JETM, JPIM, RTM, TASM, RDM, IEEETEM, TFSC, T) published between 2000 and 2009.

Lee (2015) applied journal citation network analysis to TM to provide an overview of the field. The author selected 10 TM specialty journals (IEEETEM, IJTM, JETM, JPIM, RDM, RP, RTM, TASM, TFSC, T) as base journals and constructed a network of citations between journals for years 2007 through 2011. The author determined three additional journals (IMPP, JTT, RE) that deserved to be included in TM specialty journals.

Lee and Kang (2018) used LDA to model topics in a corpus of 11,693 article abstracts published in 11 TIM journals (IEEETEM, IJTM, JETM, JPIM, RDM, RP, RTM, TASM, TFSC, T, JTT) between 1997 and 2016.

Huang et al. (2019) performed a bibliometric analysis that included cocategory, cocitation, cokeyword and coauthor analyses to achieve an overall understanding of TIM. The authors reviewed 16,801 articles published in 13 journals (RP, JPIM, T, JTT, TFSC, RDM, RTM, JETM, ICC, TASM, IEEETEM, IJTM, II) between 1997 and 2017. They added two journals (ICC and II) to the list of TM specialty journals given by Lee (2015) and dropped two journals (IMPP and RE) from the list.

2.5. Topics Studied in TM

Allen and Sosa (2004) scanned the contents of 50 years of the IEEETEM between 1954 and 2004. The authors categorized papers into 11 topical groups and analyzed their trends on a decade basis. The topical groups used in the study were given as follows: 1) Planning and Control, Project Selection, and Mathematical Modeling 2) Manufacturing, CAD/CAM, and Supply Chain Management 3) Human Resources and Staffing 4) Organizational Structure and Issues 5) Organizational and Major Program Management 6) Product Development and Project Management 7) Technology

Transfer, Technical Communication, and University/Industry Relations 8) Strategy and Policy 9) Entrepreneurship and New Ventures 10) Marketing 11) Other.

Pilkington and Teichert (2006) extracted citation and co-citation data from papers published in T between 1996 and 2004. As a result of a factor analysis of the co-citations, the authors identified seven sub-fields in TM: 1) strategy and technology, 2) national systems and differences, 3) sources of competitive strategies, 4) manufacturing/operations/new product development (NPD), 5) knowledge management and inventors, 6) patents, 7) life-cycles/change/discontinuity.

Merino et al. (2006) analyzed T in terms of bibliometric indicators, thematic progressions, authors and institutional affiliations for years between 1981 and 2004. The authors grouped twelve keywords (1. Displacement of existing products, 2. NPD and introduction, 3. Governmental and industrial policy which inhibit or stimulate technological innovation, 4. Process of technological innovation, 5. Technological trends and breakthroughs, 6. Management of entrepreneurial ventures, 7. The innovator as an individual and as a personality type, 8. Investment strategies related to new science or technology based enterprises, 9. Management of innovation in medium sized and large organizations, 10. Organizational structures intended to facilitate innovation, 11. Technology transfer to developing nations, 12. Others) into three thematic areas (1. technology innovation, 2. entrepreneurship 3. technology management).

Biemans et al. (2007) analyzed citations to and from JPIM. The authors used Product Development and Management Association (PDMA) Body of Knowledge Classification which had the following subject areas: 1) process execution and metrics 2) strategy, planning and decision making 3) customer and market research 4) people, teams and culture 5) technology and intellectual property 6) industry context and alliances. The current version of this classification is as follows: 1) Strategy 2) Portfolio management 3) New products process 4) Culture, organization, and teams 5) Tools and metrics 6) Market research 7) Life cycle management.

Page and Schirr (2008) conducted a content analysis on 815 articles written on NPD in 10 different journals (three Research and Development (R&D), three marketing,

four management) between 1989 and 2004. For classification of articles, the authors used JPIM Scheme which had the following subject categories: 1) Organizing for Innovation 2) Product Development 3) Strategy 4) New Product Planning 5) Technology Innovation 6) Market Analysis 7) Forecasting 8) Creativity 9) Concept and 10) Industry Analysis. In addition to the subject categories listed above, they developed a more detailed classification system with 42 streams of research to analyze trends.

Çetindamar et al. (2009) performed a content analysis on TM journals to understand similarities and differences between developing and developed countries in terms of TM topics, approaches, research focus, and methods. In order to achieve this purpose, they identified 22 groups of topics and analyzed primary and secondary topics studied in the articles.

Beyhan and Çetindamar (2011) conducted research to identify intellectual pillars of developing countries' TM literatures. Some of the topics identified in the study were given as: 1) innovation management, 2) R&D management, 3) product development through R&D, 4) networks and collaborations for innovation, 5) emerging technologies (biotechnology and information technology) and innovation, 6) determinants of innovation, 7) diffusion of innovations, 8) absorptive capacity, 9) organizational change and innovation, 10) R&D performance, 11) innovation performance, 12) industrial innovation, 13) technology management, 14) innovation strategy, 15) innovation in developing countries.

Choi et al. (2012) classified papers in MOT research into the following 13 research domains: 1) technology innovation 2) technology strategy 3) technology policy, technology analysis, forecast, and roadmap, 4) research and development 5) technology transfer and commercialization 6) NPD 7) entrepreneurship 8) organization learning, culture and human resource development 9) project management 10) knowledge management 11) intellectual property rights 12) social change 13) no specific classification.

Antons et al. (2016) used LDA to extract 57 topics from a full-text corpus of 1,008 JPIM articles published between 1984 and 2013. The authors analyzed topic trends by computing yearly number of research articles with topic rates above 10%.

Gudanowska (2017) analyzed research trends in TM between 2011 and 2016 by using 68 keywords defined in publications referring to the issues of technology management. The author prepared a map of research trends based on co-occurrence of the keywords and identified 9 clusters in the map.

Lee and Kang (2018) extracted 50 topics by using LDA from 11,693 article abstracts published in 11 TIM journals between 1997 and 2016. The authors investigated subspecialties of the journals and analyzed topic trends by examining topic rankings over time.

Kim and Chen (2018) used LDA to extract 40 topics from 922 full-text IEEETEM articles published between 1998 and 2017. The authors investigated topic trends by counting the number of articles in four different time periods. The numbers of articles were determined based on the rule that each article could be on only one topic.

CHAPTER 3

METHODS

In this section, the methods used within this study are presented. Subsection 1 presents an overview on LDA method. Subsection 2 describes DE method which is used for LDA parameter optimization. Subsection 3 presents word vector representation employed within DE. Subsection 4 concludes with the selection of the right model.

3.1. Latent Dirichlet Allocaton (LDA)

LDA (Blei et al**.,** 2003) is a three-level hierarchical and generative probabilistic model in which the following assumptions are made:

1) each document is a mixture of a finite number of topics,
2) each topic is a mixture of a finite number of words, and
3) a document is generated by the following process:
    a. a topic distribution is drawn for the document,
    b. for each word of the document:
        i. a topic is drawn from the document's topic distribution,
        ii. a word is drawn from the topic's word distribution.

LDA reverses this generation process and extracts latent document-topic and topic-term distributions from observable document-term distributions.

Joint distribution of latent and observable variables in LDA model is given as in Equation 4 (Blei, 2012).

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \quad (4)$$

where K is the number of topics, D is the number of documents, N is the number of words, $\beta_{1:K}$ is the ratio of words in topics, $\theta_{1:D}$ is the ratio of topics in documents, $z_{1:D}$ is the topic assignments for documents and $w_{1:D}$ is the observed words in documents.

By using this notation, posterior is written as in Equation 5.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \quad (5)$$

However, the marginal probability in the denominator is intractable to compute. Therefore, it is necessary to use inference algorithms to approximate the posterior. There are two types of algorithms that can be used for this purpose: sampling-based and variational (Blei, 2012). While sampling-based algorithms use samples from the posterior, variational algorithms use parameters to find the distribution that is closest to the posterior.

The most widely used sampling-based algorithm is Gibbs sampling. Griffiths and Steyvers (2004) describe how to apply Gibbs sampling to LDA by introducing a new prior in the model. The final probability model is given as follows:

$$\begin{aligned} w_{d,n}|z_{d,n}, \phi^{(z_{d,n})} &\sim \text{Discrete}\left(\phi^{(z_{d,n})}\right) \\ \phi &\sim \text{Dirichlet}(\beta) \\ z_{d,n}|\theta^{(d)} &\sim \text{Discrete}\left(\theta^{(d)}\right) \\ \theta &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

where α is the Dirichlet parameter of document-topic distributions (θ) and β is the Dirichlet parameter of topic-word distributions (ϕ). The plate notation of LDA model with Gibbs sampling is given in Figure 1.

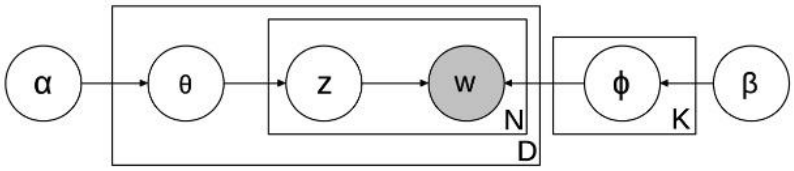


Figure 1. LDA Model with Gibbs Sampling (Wikipedia, 2021)

In this study, we prefer to use Mallet (McCallum, 2002) topic model package which includes a Gibbs sampling implementation of LDA.

3.2. Differential Evolution (DE)

DE is a stochastic search algorithm proposed by Storn and Price (1997) as a heuristic approach for continuous space optimization problems. It is a typical Evolutionary Algorithm (EA) and close to GA. The difference between DE and GA is in the mutation operation (Feoktistov and Janaqi, 2004). In this study, DE algorithm is preferred but any other multimodal optimization algorithms can also be used for this purpose.

EAs are population-based algorithms and evolve the parameters' values at each generation like organisms in the nature. Similarly, the aim of DE algorithm is to repetitively optimize a population of parameter vectors at each generation. Parameters are randomly initialized within a range between their predefined minimum and maximum values. Whether a parameter vector will be transmitted to the next generation depends on the value returned by the cost function.

DE algorithm is composed of four stages, three of which are iterative as given in Figure 2.

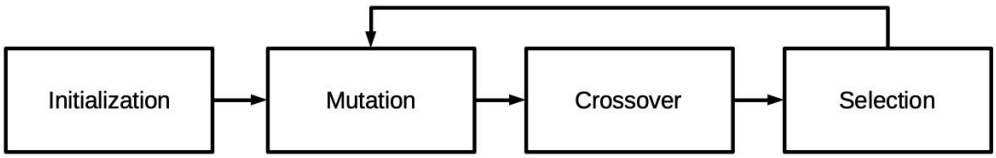


Figure 2. Stages of DE Algorithm (Das and Suganthan, 2011)

3.2.1. Initialization

Minimum $(x_{d,min})$ and maximum $(x_{d,max})$ values are defined separately for each parameter (d) of the problem. For the first generation $(g = 1)$ of the population, the parameters are set as given in Equation 6.

$$x_{np,d,g=1} = x_{d,min} + rand_{np,d}[0,1] \times (x_{d,max} - x_{d,min}) \qquad (6)$$

where $d = 1,2,3,\dots,D$ in which D is the number of parameters, $g = 1,2,3,\dots,G$ in which G is the number of generations, $np = 1,2,3,\dots,NP$ in which NP is the number of populations and $rand_{np,d}[0,1]$ is a random number regenerated for each parameter of each population vector.

3.2.2. Mutation

In biology, mutation refers to a change in the sequence of DNA. Similarly, mutations in DE algorithm are changes in the values of parameter vectors. In this stage, a mutant vector is generated for each target parameter vector $(x_{np,g})$ as given in Equation 7.

$$m_{np,g+1} = x_{r_1,g} + F \times (x_{r_2,g} - x_{r_3,g}) \qquad (7)$$

where $r_1, r_2, r_3 \in 1,2,3,\dots,NP$ are random indexes different from np and from each other and $F \in [0,2]$ is the amplification factor.

Mutation strategies define the way of generating a mutant vector. The notation for the mutation strategies is $DE/x/y/z$ where x is the selection method (rand, best, etc.) of the target parameter vector, y is the number of difference vectors and z is the crossover method (binomial, exponential, etc.).

In addition to the $DE/rand/1/bin$ strategy described above, many other strategies have been proposed such as $DE/best/1/bin$, $DE/best/2/bin$, $DE/rand-to-best/1/bin$ and $DE/rand/2/bin$. Among them, $DE/rand/1/bin$ is the most suitable strategy for multimodal problems because of its stronger exploration capability (Qin et al., 2009). To this end, in our study we implemented DE algorithm by using $DE/rand/1/bin$ strategy.

3.2.3. Crossover

The aim of this stage is to increase diversity. A trial vector is generated by applying a crossover operation to the mutant vector as given in Equation 8.

$$t_{np,d,g+1} = \begin{cases} m_{np,d,g+1} & if\left(d = rand_{np}[1,D]\right) \\ m_{np,d,g+1} & if\left(rand_{np,d}[0,1] \le CR\right) \\ x_{np,d,g} & otherwise \end{cases} \quad (8)$$

where $CR \in [0,1]$ is the crossover constant, $rand_{np,d}[0,1]$ is a random number regenerated for each parameter and $rand_{np}[1,D]$ is a random number that provides the trial vector with at least one parameter from the mutant vector.

3.2.4. Selection

In this stage, it is decided whether the trial vector will be transmitted to the next generation or not. The target parameter vector $\left(x_{np,g}\right)$ and the trial vector $\left(t_{np,g+1}\right)$ are compared in terms of their cost function (cf) values and the one that has the lower (or higher) cost is transmitted to the next generation as given in Equation 9.

$$x_{np,g+1} = \begin{cases} t_{np,g+1} & if\left(cf\left(t_{np,g+1}\right) \le cf\left(x_{np,g}\right)\right) \\ x_{np,g} & if\left(cf\left(t_{np,g+1}\right) > cf\left(x_{np,g}\right)\right) \end{cases} \quad (9)$$

3.2.5. Termination Condition

DE algorithm can be terminated in four different ways:

1) By specifying the maximum number of generations (Storn and Price, 1997)

2) When the difference between the best and the worst cost function values is less than a pre-specified threshold (Ali and Törn, 2004)

3) When the best cost function value does not change for a number of successive iterations (Das and Suganthan, 2011)

4) When the best cost function value is equal to or greater than a pre-specified threshold (Das and Suganthan, 2011)

In this study, the maximum number of generations is specified to terminate DE algorithm.

3.3. Word Vector Representations

As stated before, various methods have been proposed to optimize the LDA parameters. Griffiths and Steyvers (2004), Teh et al. (2006), Steyvers and Griffiths (2007), Zavitsanos et al. (2008), Cao et al. (2009), Arun et al. (2010) and Mehta et al. (2014) propose to use statistical methods. Agrawal et al. (2018) consider overlaps of top terms as the optimization criterion. Greene et al. (2014) use agreements between term rankings, whereas Krasnov and Sen (2019) transform topics into dense vectors to be used in assessing the quality of clusters. Only in the studies conducted by Stevens et al. (2012) and Krasnov and Sen (2019), the dependency between average coherence scores of topic models and the number of topics is evaluated and reported to be stable/monotonous which does not allow determination of the optimal number of topics.

Hence we see that none of the previous studies has considered optimizing the parameters of LDA using a cost function that maximizes the semantic relationship scores of parameter vectors. We use average similarity scores of parameter vectors as the selection criterion of an optimization algorithm. Our intuition is that a model with the highest score: 1) has no intruder words, 2) has the highest interpretability, and 3) has the highest distinction between topics, and this situation occurs at the right parameter values.

The reason for using distributed word vector representations in this study is their success in measuring the coherence of topics, as found in the studies of Aletras and Stevenson (2013) and Nikolenko (2016). There are several ways to create word vector representations but the most popular ones are Word2Vec (Mikolov et al., 2013) and GloVE (Pennington et al., 2014). Since Naili et al. (2017) report that Word2Vec performs better than GloVE, we use Word2Vec for word representations in this work.

Word2Vec has two model architectures: Continuous Bag-of-Words model (CBOW) and Continuous Skip-gram model (Skip-gram). While CBOW predicts the current word based on the context, Skip-gram predicts the context based on the current word. Graphic representation of Word2Vec is given in Figure 3.

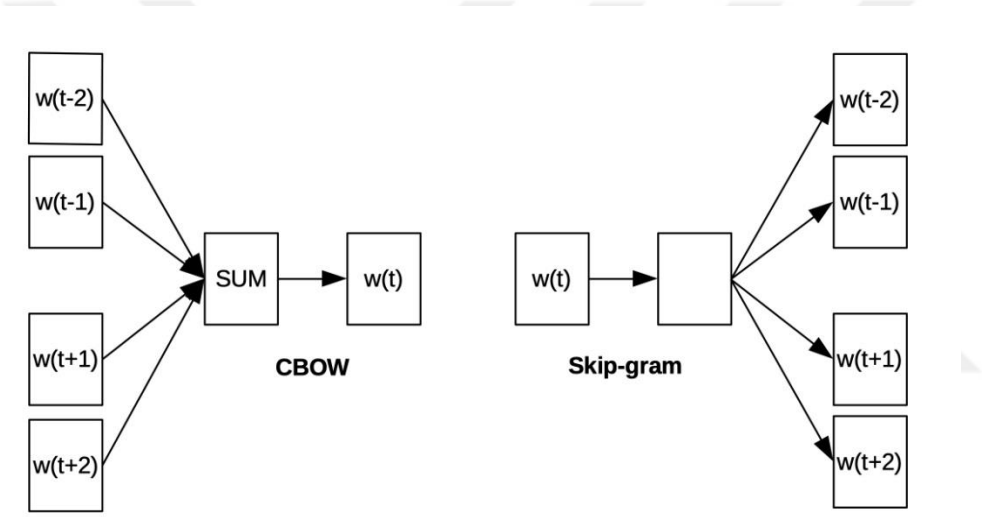


Figure 3. Word2Vec model architectures (Mikolov et al., 2013)

In the paper in which Mikolov et al. (2013) introduce Word2Vec, the authors compare CBOW and Skip-gram model architectures by using the same dimensionality and the same training data. They report that while CBOW model achieves more accurate results in syntactic relationship, Skip-gram model achieves more accurate results in semantic relationship. In this study, we use both methods to compare their effectiveness in optimizing the LDA parameters.

In the method we propose, similarity score of a topic (k) is calculated as given in Equation 10.

$$\text{sim}_k = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} similarity\left(w_i, w_j\right) \quad (10)$$

where T is the number of top terms, w is a top term and $similarity()$ is a function that returns the similarity score of two words which is calculated by measuring the similarity (or distance) of corresponding word vectors.

Since we aim to include outlier values, we use mean as the average score and calculate similarity score of an LDA model (m) as given in Equation 11.

$$\text{sim}_m = mean\{\text{sim}_k, k \in \{1, \dots, K\}\} \quad (11)$$

where K is the number of topics.

In order to increase the consistency in the scores of the cost function, this process is repeated for all the replicated models of a parameter vector, and the mean value of the model similarity scores is calculated as the similarity score of the parameter vector (pv) as given in Equation 12.

$$\text{sim}_{pv} = mean\{\text{sim}_m, m \in \{1, \dots, M\}\} \quad (12)$$

where M is the number of replicated models.

In this study, vector representations of words are generated by using Gensim (Řehůřek and Sojka, 2010) library with default values for parameters. The model is trained on the collection of Wikipedia articles downloaded on December 08, 2019, which has a vocabulary size of 4,252,721 words and a corpus size of 2,418,101,826 tokens. We set the number of top terms and the number of replicated models at 10.

3.4. Presentation of Topics

Optimum values of the LDA parameters for any corpus can be found by using the method given above. However, an additional method is required to select the right topics to be presented to the end users. Mantyla et al. (20189 argue that using the

results of a single LDA run does not reflect the topics properly and propose to cluster topics from replicated models. Chuang et al. (2015) introduce an interactive tool which allows users to assess model stability by comparing replicated models.

The way we prefer to select the right topics is to use the results of the model with the highest similarity score among replicated ones. This method has the advantages of using the model with the highest semantic interpretability and requiring no additional work to create presentable topics.

CHAPTER 4

EXPERIMENTS ON LDA PARAMETER OPTIMIZATION

4.1. Corpora

In this part of the study, two corpora with different properties are used. The first one is a corpus of 2,283 articles published in open access Turkish Journal of Electrical Engineering and Computer Sciences between 1995 and 2019. The second one is Reuters-21578 collection, which is a corpus of 21,578 documents that appeared on Reuters newswire in 1987.

Syed and Spruit (2017) compare the extraction of topics from abstracts and full-text articles and detect that full-text data is less affected by noise terms which decrease the coherence. Therefore, in the analysis, full-text articles are used in the first corpus, except for the first 37 articles, only the abstracts of which are publicly accessible.

In the Reuters corpus, there are a total of 135 economic subject categories, 120 of which appear in at least 1 of the documents and 57 of which appear in at least 20 of the documents.

4.2. Parameter Settings

In the model training, we added several extra words into the Mallet's default stop words list. We set the parameters max-idf and min-idf at 7.0 and 1.0, respectively. These parameter values remove all the words with a document frequency of less than 19 or greater than 7,939 in the Reuters corpus and less than 2 or greater than 840 in the Journal corpus. We filtered words with less than four characters and with non-alphabetical characters. The final vocabulary size and corpus size are 5,139 and 1,039,200 for the Reuters corpus and 27,836 and 2,865,417 for the Journal corpus, respectively.

In the DE algorithm, we set the number of parameter vectors at 30, amplification factor at 0.9, crossover constant at 0.9, the number of generations at 50, the minimum and maximum values of the parameters K, α and β at 10-100, 0-1, and 0-1, respectively.

We also investigate the effects of text preprocessing on the optimization process by using different values for max-idf and min-idf parameters and by merging named entities into single tokens. We used Stanford Named Entity Recognizer (Manning et al., 2014) to identify named entities.

4.3. Analysis on Topic Instability

In order to determine whether the cause of topic instability is order effect, we conduct an experiment such that we ordered the articles in the Journal corpus by publication date and determine the level of instability in topics extracted from the ordered articles. To this aim, we compare the topics of two replicated models of a parameter vector ($\mathrm{K} = 34,\ \alpha = 50/34,\ \beta = 0.1$) by counting overlaps of 10 top terms. We set the number of topics at 34 because, as will be seen later, the optimum number of topics is found to be 34 in the Journal corpus. The results obtained from the comparison are given in Table 2.

Table 2. Number of Overlaps of Top Terms

Overlaps	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
Number of Topics	2	6	8	6	3	3	2	4	0	0

The low number of overlaps between top terms of topics indicates high level of instability. Based on this, since the level of instability is high in the articles ordered in the same way and it would not be appropriate to change the order of the articles with known publication dates in any other way, it can be concluded that the reason for topic instability in LDA is not (or not only) the order effect.

This means that topic instability is a problem related to the stochastic nature of LDA. This suggests a need for a way to eliminate its effects on the semantic and syntactic

relationship scores of topic models. However, as stated before, it does not have an exact solution, there are only several strategies proposed to alleviate its effects.

In order to see the effect of topic instability on the semantic relationship scores of topic models, we calculated Skip-gram similarity scores of 10 replicated models of a parameter vector $(K = 57, \alpha = 50/57, \beta = 0.1)$ extracted from the Reuters corpus. We set the number of topics at 57 which is the number of economic subject categories that appeared in at least 20 of the documents in the Reuters corpus. Scores obtained from two runs are given in Table 3 and Table 4.

Table 3. Similarity Scores of Replicated Models - First Run

Model	Mean Score	Min Score	Max Score
1	19.367801	11.208375	27.617414
2	19.197771	11.682178	30.197244
3	19.191798	10.518077	25.388879
4	18.905101	10.518077	26.164223
5	18.948176	12.487615	27.617414
6	19.183293	11.718470	27.759172
7	19.188607	10.518077	26.533242
8	18.911677	12.184703	26.527101
9	19.390160	13.701636	27.751282
10	19.317189	11.983537	27.785325

Table 4. Similarity Scores of Replicated Models - Second Run

Model	Mean Score	Min Score	Max Score
1	19.139778	10.540663	27.756282
2	19.162742	10.542663	27.617414
3	19.367723	12.420408	26.206174
4	19.480483	12.635156	28.092780

Table 4 *(cont'd)*

5	19.223068	11.298941	27.617414
6	19.273109	10.518077	30.915678
7	19.162590	12.107647	27.617414
8	19.445265	10.518077	29.247615
9	19.156673	12.674428	27.617414
10	19.084820	10.832464	27.617414

In the tables, the same scores in the Max Score and Min Score columns belong to the topics which have the same top terms in different orders.

It can be seen from Table 3 and Table 4 that the difference between the highest and the lowest mean scores of the replicated models has the potential to negatively affect the success of an optimization algorithm. In order to alleviate this difference, we have decided to use a similar way to that used by most of the researchers and computed average scores of replicated topic models. Although it has a negative effect on the running time of the algorithm, the method has two positive effects: 1) it increases the consistency in the similarity scores of a parameter vector, and 2) even if the top terms of a topic change due to topic instability, changes with the words belonging to the same semantic field do not cause a significant change in the similarity score.

For the above given example, mean value of the parameter vector is 19.160157 in the first run and 19.249625 in the second run, which means a difference of 0.089468 between two runs. For cross-validation, we run the algorithm two more times and obtain the scores of 19.203078 and 19.217540. These results suggest that the average score becomes more stable as the number of replicated models increases, and the method we use can reduce the differences between similarity scores of a parameter vector to acceptable levels.

As another analysis of the order effect, we calculate average Skip-gram similarity scores of the parameter vectors given above for different orders of input documents: 1) default ordered, 2) well ordered (the Reuters corpus by alphabetical order, the

Journal corpus by publication date), and 3) reverse ordered. Scores obtained from four replicated runs (a total of 40 replicated models) on the Reuters corpus and the Journal corpus for each ordering are given in Table 5 and Table 6, respectively.

Table 5. Similarity Scores in the Reuters corpus

Order	Run 1	Run 2	Run 3	Run 4
Default	19.160157	19.203078	19.217540	19.249625
Well	19.149234	19.190832	19.204757	19.292303
Reverse	19.201709	19.211751	19.232542	19.330780

Table 6. Similarity Scores in the Journal corpus

Order	Run 1	Run 2	Run 3	Run 4
Default	19.168598	19.213635	19.246590	19.336062
Well	19.207929	19.227536	19.294607	19.306033
Reverse	19.257453	19.280972	19.301260	19.335961

As can be seen from Table 5 and Table 6, the scores are close to each other. This shows that the order of input documents does not have an effect on the semantic relationship scores of topic models.

4.4. Analysis on Topic Extraction with Parameter Optimization

The optimal parameter values obtained by changing the cost function while keeping the parameters max-idf (7.0) and min-idf (1.0) constant are given in Table 7.

Table 7. Scores Obtained by Using Different Cost Functions

Corpus	Function	K	A	β	Score
Reuters	CBOW	41	0.365681	0.564030	12.291904
	Skip-gram	52	0.946186	0.021711	19.409072
Journal	CBOW	11	0.254452	0.117408	16.459047
	Skip-gram	34	0.118924	0.005694	19.547783

As mentioned in Section 4.1, in the Reuters corpus, there are 120 economic subject categories that appear in at least 1 of the documents and 57 economic subject categories that appear in at least 20 of the documents. Since we removed the words with a document frequency of less than 19 or greater than 7,939, we expected the optimum number of topics to be around 57. Under the Skip-gram model, the number of topics is determined as 52, which is close to the expected number.

In the Journal corpus, under the Skip-gram model, the optimum number of topics is found as 34. This result can be considered plausible since the Journal corpus has a narrow scope in comparison with the Reuters corpus.

Top terms of 10 topics with the highest and the lowest scores in the models extracted using the optimum parameter values from the Reuters corpus and the Journal corpus are given in Table 8 and Table 9, respectively.

Table 8. Topics with the Highest and the Lowest Scores in the Reuters corpus

No	Top Terms	Score	Rank
50	securities financial companies foreign investment market firms investors capital markets	28.092780	1
27	president chairman chief executive board vice officer director named company	27.617414	2
17	tonnes sugar production export exports total tonne wheat stocks imports	24.276302	3
0	joint services company venture corp companies telephone international service communications	22.931765	4
13	systems products computer system technology corp data line equipment business	22.467803	5
48	federal system reserve funds reserves statement market expected period agreements	16.329517	48
39	loss profit note includes gain year operations share sales extraordinary	15.904988	49

Table 8 *(cont'd)*

51	issue bond lead bonds manager priced coupon date payment selling	15.276265	50
1	dollar currency west exchange german central germany mark rates paris	15.275254	51
3	rose fell rise figures increase department adjusted revised orders index	14.725310	52

Table 9. Topics with the Highest and the Lowest Scores in the Journal corpus

No	Top Terms	Score	Rank
0	measurement sensor measurements temperature sample samples device sensors surface tests	25.600047	1
19	converter switching inverter harmonic grid circuit switches currents capacitor voltages	23.687377	2
1	graph node tree nodes path query edges networks edge graphs	23.544503	3
24	cell cells flow pressure water blood pump fluid concentration velocity	22.588380	4
32	nonlinear stability delay chaotic discrete stable matrices dynamics lyapunov theorem	22.258865	5
20	controller response controllers mode loop sliding gain observer damping disturbance	16.466181	30
14	robot position robots vehicle motion force tracking velocity joint road	16.397793	31
13	fuzzy rule agent rules membership agents logic learning decision inference	16.293034	32
7	optimization search population fitness particle solutions swarm iteration objective local	15.780419	33
11	flow loss active generation transmission constraints reactive optimization units buses	15.676127	34

In order to label the topics extracted from the Reuters corpus, we match the extracted topics to the given economic subject categories. For the sake of objectivity in the matching, we determine the overlapping words between the top terms of the topics and the names of the economic subject categories. As a result, 27 of 52 topics are labelled by using 37 categories. Top terms of the topics and the corresponding economic subject categories are given in Table 10. The numbers given in parentheses denote the number of documents that the given economic subject categories appear in.

Table 10. Economic Subject Categories as Topic Labels

No	**Top Terms**	**Categories**
1	dollar currency west exchange german central germany mark rates paris	D-Mark (0)
3	rose fell rise figures increase department adjusted revised orders index	Industrial Production Index (65)
7	union workers strike spokesman airlines work national general employees aircraft	Unemployment (76)
8	trade exports south imports united states foreign taiwan products import	Trade (552)
11	production energy crude opec petroleum barrels prices barrel natural output	Crude Oil (634), Natural Gas (130)
15	pacific areas steel area southern weather north early people normal	iron-steel (67)
16	debt banks brazil interest foreign payments billion country bank creditors	Balance of Payments (116), Brazilian Cruzado (1)
17	tonnes sugar production export exports total tonne wheat stocks imports	sugar (184), wheat (306)
18	quarter year earnings share company expects income reported profits results	Personal Income (18), Earnings and Earnings Forecasts (3987)

Table 10 *(cont'd)*

19	exchange trading stock futures options market contracts york contract chicago	Money/Foreign Exchange (801)
20	canada canadian statement british announced made dome continental added full	Canadian Dollar (3)
21	credit interest financing basis facility paper program offered international years	Instalment Debt/Consumer Credit (7)
25	japan japanese officials trade tokyo ministry official open united states	Japanese Yen (69)
28	department agriculture program wheat corn farm usda farmers grain land	corn (254), grain (628), wheat (306)
30	analysts market industry analyst major kong hong time stocks street	Hong Kong Dollar (1)
31	gulf iran saudi iranian united military states shipping agency iraq	Shipping (305)
33	market dealers traders dollar close buying early closed trading selling	U.S. Dollar (217)
35	offer share merger company tender takeover corp shareholders cash board	Mergers/Acquisitions (2448)
36	coffee international meeting agreement producers export cocoa stock council market	coffee (145), cocoa (76)
38	growth economy year economic rise inflation report domestic demand expected	Gross National/Domestic Product (163)
41	gold resources australia production mining mine tons australian copper silver	gold (135), copper (78), silver (37), Australian Dollar (4)
42	european community commission french europe ministers france sources proposals west	French Franc (0)

Table 10 *(cont'd)*

43	world countries economic international development meeting baker nations developing major	Leading Economic Indicators (17)
45	sales corp cars general drug stores motors chrysler food company	Retail Sales (27)
47	bank rate rates money interest market central bills treasury england	Interest Rates (513)
48	federal system reserve funds reserves statement market expected period agreements	Reserves (84)
51	issue bond lead bonds manager priced coupon date payment selling	Wholesale Price Index (32)

Some of the surprising results from the labeling are listed below with their examples:

1) Some of the topics could not be labelled despite they appear to be economic subjects with manual checking (Topic 50 in Table 8)

2) Some of the topics are labelled with more than one economic subject category (Topics 11, 16, 17, 18, 28, 36, 41)

3) Some of the topics could not be labelled since they are not economic subjects (Topics 0, 13, 27 in Table 8)

4) Some of the topics are labelled with economic subject categories that appear in less than 20 of the documents (Topics 1, 20, 21, 30, 42, 43)

These results show that the corpus has two problems related to topic modeling: 1) all the given economic subject categories are not in the same level, and this prevents the extraction of all of them by a single topic model, and 2) some documents do not have economic subject category assignments, and this causes the extraction of topics either not given in the economic subject categories or appear in less than 20 of the documents.

These problems are an indication that a complete one-to-one match is not possible. However, as will be seen below, by changing the parameter values, topics can be extracted at different levels. With this method, it becomes possible to extract the economic subject categories that cannot be extracted by a single model.

Another result we obtain in this study is that optimum values of parameters α and β highly vary with different corpora. This shows that these parameters need to be optimized for each corpus, as opposed to what is applied in the study of Griffiths and Steyvers (2004).

In our analysis, it is seen that Skip-gram model provides higher similarity scores than CBOW model in both corpora. This result is compatible with the study of Mikolov et al. (2013) in which Skip-gram model achieved more accurate results in semantic relationship. Additionally, as expected, it gives higher similarity scores than the models extracted using the default settings of LDA, scores of which are given in Table 5 and Table 6.

Since it was determined that the statistical methods did not match the interpretability of topics, we compared the results only with those obtained by the method given in the study of Agrawal et al. (2018). As a result of the optimization done by using LDADE, the optimum number of topics was found to be 10 for both corpora.

In the light of the results given above, it can be concluded that the optimization with Skip-gram model is more successful due to its higher coherence scores. Therefore, we use Skip-gram similarity scores to optimize topic models.

As stated before, we investigate the effects of the text preprocessing on the optimization process. To this aim, firstly, we optimize the LDA parameters with merged named entities. We extract the named entities with Location, Person, Organization and Misc types, filter those with more than four words and merge the remaining ones into single tokens. Scores obtained by using this method are given in Table 11, from which it can be seen that the method increased the optimum number of topics for both corpora, although the increase is not very marked in the Journal corpus.

Table 11. Effects of Merged Named Entities

Corpus	K	α	β	Score
Reuters	59	0.897372	0.010262	18.805330
Journal	35	0.229822	0.021636	18.615432

Secondly, we optimize the LDA parameters for different values of max-idf and min-idf parameters. Scores obtained from the Reuters corpus and the Journal corpus are given in Table 12 and Table 13, respectively. It can be seen from the tables that the changes in the parameter values also lead to changes in the optimum number of topics. This is an indication that changing the parameter values results in the extraction of topics at different levels.

Table 12. Effects of Different Parameter Values on the Reuters Corpus

max-idf	min-idf	K	α	β	Score
7.0	4.0	38	0.082931	0.077762	17.727794
7.0	5.5	52	0.043958	0.195187	16.329127
6.5	1.0	49	0.864662	0.017281	19.315531
6.0	1.0	37	0.919898	0.004740	19.320366

Table 13. Effects of Different Parameter Values on the Journal Corpus

max-idf	min-idf	K	α	β	Score
7.0	4.0	10	0.675458	0.130764	17.001204
7.0	5.5	22	0.551891	0.020239	11.683725
6.5	1.0	25	0.021310	0.118982	19.695736
6.0	1.0	27	0.176330	0.125691	19.766908

As an example of this situation, we set max-idf at 7.0 and min-idf at 3.4 and reduce the vocabulary size to 4,946 in the Reuters corpus. The optimum values for the parameters K, α and β are found to be 18, 0.097496 and 0.317216, respectively. Top

terms of one of the topics extracted by using the optimum values are *wheat, sugar, grain, agriculture, corn, crop, usda, soybean, farmers, soviet*. This topic is a combination of the topics 17 and 28 given in Table 10 and adds a new item, namely *soybean*, to the list of economic subject categories used as labels.

However, as seen from the example and the values given in Table 12 and Table 13, the changes in the parameter values do not have a stable effect on the number of topics. It appears that the effects of changes in the parameter values on the optimization results depend on the distribution of words in the corpus. We consider this analysis as a subject for future studies.

CHAPTER 5

APPLICATION OF TOPIC MODELING TO TM DOMAIN

An overview of the process followed in this section is given in Figure 4.

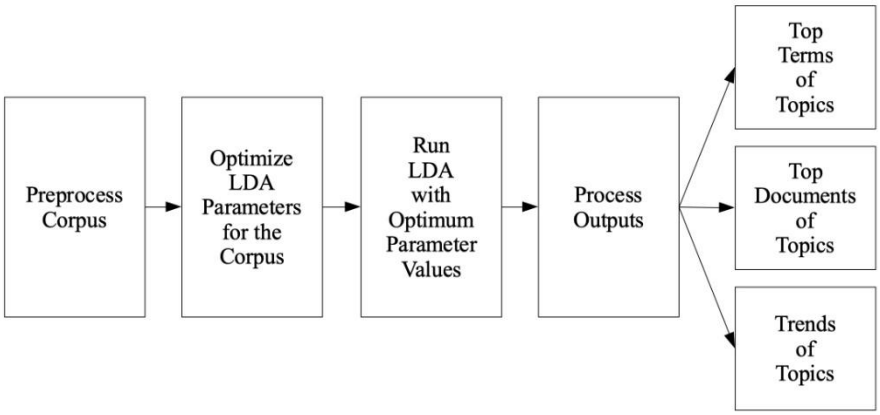


Figure 4. Flow Diagram of the Process

5.1. Corpus

In the previous studies on TM, it is observed that the analyses are conducted on a certain set of journals. Most of the studies rank the journals and use top 10 of them in their analyses. One of the studies (Lee, 2015) is remarkable in that it applies journal citation network analysis on 10 TM specialty journals and determines three additional journals that deserve to be included in the list. Therefore, in our study, 13 journals given by Lee (2015) are selected as the TM's top journals.

Antons et al. (2016) and Syed and Spruit (2017) report that using full-text articles generate more coherent topics than abstracts. Therefore, in this study, full-text articles are used to create the corpus. However, since the articles of IJTM are not publicly

accessible, only abstracts are used from this journal. As a result, 14,471 full-text research articles from 12 journals and 1,709 abstracts from one journal, a total of 16,180 research articles/abstracts are downloaded from the top journals specialized on TM. Table 14 presents the journal names, and the counts and types of articles used in this study.

Table 14. Numbers and Types of Articles

No	Name	# of articles	Type
1	IEEE Transactions on Engineering Management	1,048	Full-text
2	Innovation: Organization & Management	493	Full-text
3	International Journal of Technology Management	1,709	Abstract
4	Journal of Engineering and Technology Management	435	Full-text
5	Journal of Product Innovation Management	1,024	Full-text
6	Journal of Technology Transfer	936	Full-text
7	R&D Management	838	Full-text
8	Research Evaluation	649	Full-text
9	Research Policy	2,494	Full-text
10	Research Technology Management	925	Full-text
11	Technological Forecasting and Social Change	3,026	Full-text
12	Technology Analysis & Strategic Management	1,186	Full-text
13	Technovation	1,417	Full-text

Figure 5, generated by using Matplotlib library (Hunter, 2007), shows the number of articles published in the selected journals per year. As seen in the figure, there is a steady increase in the total number of research articles published in the TM journals which shows a gradually increasing research interest in TM.

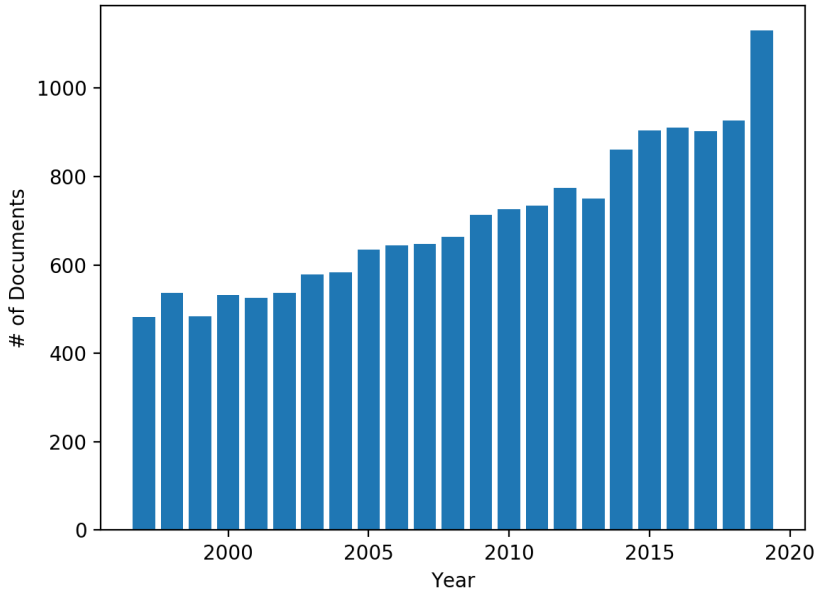


Figure 5. Number of Documents per Year

In the preprocessing of the corpus, stop words are filtered out by using Mallet's default stop words list. This default list is further extended for filtering as follows. Words shorter than four characters, or containing non-alphabetical characters are filtered out. Names of journals, authors, countries, cities and companies are also discarded. Words that still existed in the top terms are replaced with coded texts. Named entities including Location, Person, Organization and Misc types are identified by using Stanford Named Entity Recognizer (Manning et al. 2014) and those consisting of up to four words are added to the vocabulary.

5.2. Results

As stated in the previous sections, LDA parameters are optimized by using average semantic similarity scores of replicated topic models. One advantage of this method is its ability to determine the optimum number of topics at different levels. This feature of the technique is exploited and topics are extracted at two levels by using two groups of parameters at the preprocessing stage. Table 15 presents optimum LDA parameter values obtained by each Mallet parameter group.

Table 15. Optimum LDA Parameter Values

	Preprocessing Parameters			LDA Parameters		
Level	**Min Docs**	**Max Docs**	**Terms**	**K**	**α**	**β**
High	14	5,953	64,549	13.0	0.1420528	0.1867320
Low	14	3,611	63,887	80.0	0.5725732	0.3653631

In Table 15, *Min Docs* and *Max Docs* columns denote the count of documents that are used to filter out frequent and rare terms from the corpus. They show that all the words that occurred fewer than 14 and more than 5,953 documents at high-level topics, and fewer than 14 and more than 3,611 documents at low-level topics are filtered out. *Terms* column shows the vocabulary size of the corpus after the removal of frequent and rare terms. As seen in the table, the optimum number of topics is found as 13 for the high-level topics and 80 for the low-level topics. Top terms and labels, top documents, and trends of the topics are given in Table 16, Appendix A and Appendix B, respectively, for the high-level topics, and in Table 5, Appendix C and Appendix D, respectively, for the low-level topics.

In Table 16, *Topic* column shows topic identifiers assigned automatically by Mallet toolkit, *Top Terms* column shows top 10 terms of the topics and *Label* column shows labels of the topics manually determined by checking the top terms of the topics. As an example from the table, the topic given in the first row has the topic id 0, it has the top 10 terms *patent, patents, patenting, inventors, licensing, invention, inventions, pharmaceutical, biotechnology, intellectual*, and since all the top terms are patent related, the topic is manually labeled as *Patents (IPRs).*

Table 16. Top Terms and Labels For High-Level Topics

Topic	**Top Terms**	**Label**
0	patent patents patenting inventors licensing invention inventions pharmaceutical biotechnology intellectual	Patents (IPRs)

Table 16 *(cont'd)*

1	team new_product items teams orientation innovativeness respondents creativity perceived constructs	NPD Teams
2	users user internet adoption online consumers consumer mobile platform community	ICT
3	supplier suppliers customer team product_development manager phase tools improvement equipment	Product Development
4	universities technology_transfer entrepreneurship entrepreneurial entrepreneurs funding commercialization ventures venture incubator	Academic Entrepreneurship
5	energy environmental sustainability water electricity emissions fuel green sustainable scenarios	Sustainability
6	foresight scenario experts scenarios nanotechnology expert cluster matrix forecasting trends	Foresight & Forecasting
7	price diffusion optimal parameters probability simulation forecasting profit parameter option	Optimization
8	regional foreign cluster regions region clusters domestic enterprises economies smes	Regional Policies
9	intensity dummy hypothesis coefficient estimation observations estimates spillovers estimated probability	Policy Evaluation
10	funding scientists universities publications papers publication journals articles students faculty	Scholarly Publications
11	actors political stakeholders governance organisations community society organisation health participants	Governance
12	partners alliances absorptive alliance cooperation partner exploration exploitation assets collaborative	Alliances

The issues that attract attention or require explanation in the high-level topics are as follows:

- Six topics can be grouped into three semantically intra-related clusters, each of which is composed of two topics as follows: Product Development (NPD Teams (1) and Product Development (3)), Universities' Role (Academic Entrepreneurship (4) and Scholarly Publications (10)) and Policy Making (Policy Evaluation (9) and Governance (11)). These clusters can be thought as the top three topics of TM literature.

- In line with expectations, Patents or Intellectual Property Rights (IPRs) (0) is determined as one of the main pillars of TM. The top terms *pharmaceutical* and *biotechnology* are an indication that many articles studied biotechnology and pharmaceutical patents. It shows a horizontal trend in both the high-level and the low-level topics (57), but its trend might increase in the future with the increasing use of text mining techniques on patents.

- NPD Teams (1) is a topic that addresses issues such as creativity, innovativeness, performance, intuitive skills, leadership, and fairness. Like Governance (11), it has a higher trend than other topics for the whole period. It seems that it will continue to have a higher trend due to the studies to be conducted to increase the success of NPD teams.

- Extraction of a topic labelled ICT (2), an abbreviation for Information and Communication Technology, among high-level topics is parallel to the growth in the ICT sector which continues to be a key component of the world economy. The topic shows an increasing trend and this trend will probably continue in the future in parallel with the increase in the ICT sector.

- Product Development (3) is related to NPD process. Some of the issues studied in its top documents are Modularity, Stage-Gate, Agile-Stage-Gate and Lean Principles. While it was a widely studied topic in 2000s, it lost its popularity over time.

- Although Entrepreneurship is extracted in some of the previous studies, Academic Entrepreneurship (4) is not mentioned in any of them. In this study, the cluster composed of Academic Entrepreneurship (4) and Scholarly

Publications (10), namely Universities' Role, is determined as one of the top three topics of TM. From this result, it can be deduced that researchers are more interested in founding their own companies than working with a company. Both researchers and policy makers should address this issue and investigate its effects on universities and university-industry relations.

- The fact that Sustainability (5) is determined as one of the most studied topics of TM is an indication of the importance given to it in the field of TM. It shows a slightly increasing trend in both the high-level and the low-level topics (53 and 55), and this trend is likely to continue in the future.

- Foresight & Forecasting (6) are determined as the most widely used methods of the field. This result is compatible with the fact that one of the top concerns of the community is the identification of new technologies and trends as early as possible. Horizontal trend that it had in the past indicates that studies on, or using, the methods will continue in the future. Additionally, it can be inferred from the top term *nanotechnology* that these methods were used mostly in this field.

- Optimization (7) and Regional Policies (8) are two of the high-level topics extracted from the TM corpus. While Optimization (7) is related to diffusion models, substitution models, growth models, capital investment analysis, production control and scheduling, Regional Policies (8) is related to regional innovation systems, technology transfer, foreign innovation, foreign R&D and regional characteristics (e.g. prerequisites for innovation, innovation barriers and absorption capabilities). They both show decreasing trends for the time period under consideration. From the slopes of the trends, it can be predicted that their trends will continue to decrease in the future.

- Policy Evaluation (9) is not extracted in any of the previous studies in TM. However, in this study, together with the topic labeled Governance (11), it is determined as one of the top three topics of TM. Some of the issues that the topic addresses are: a) public R&D and innovation policies such as subsidies, tax incentives and low-interest loans, b) crowding-out between public and

private spending, and c) the link between R&D, innovation and productivity. Policy Evaluation (9) shows an increasing trend and reaches to a level close to Governance (11). This demonstrates the researchers' increasing interest in evaluating the effectiveness and efficiency of policies. It is likely that technology policies and their evaluation will continue to be one of the top topics of TM. The top term *health* in Governance (11) is an indication that health (e.g. health technology assessment) is one of the main concerns of governance.

- Alliance (12) is a widely preferred way of acquiring new technologies. It shows an increasing trend in both the high-level and the low-level topics (47). In the future, the world of business will be more competitive and knowledge-intensive than today's. Hence, the companies will have to form more alliances in order to acquire new knowledge and technology. This means that the trend of this topic will continue to increase in the future.

Table 17. Top Terms and Labels for Low-Level Topics

Topic	Top terms	Label	New
0	forecasting forecast curve forecasts parameter logistic prediction cycles substitution estimation	Forecasting	
1	domestic developing_countries catch-up upgrading latecomer indigenous institutes provinces governments state-owned	Technological Upgrading	+
2	health medical healthcare care patients hospital clinical patient hospitals health_care	Healthcare	
3	business_model value_creation providers provider proposition delivery revenue co-creation offering offerings	Business Model	+
4	conflict boundary interdependence identity interpersonal collective cultures virtual champions cohesion	NPD Teams	

Table 17 *(cont'd)*

5	emissions scenario scenarios carbon climate emission climate_change consumption coal fuel	Climate Change	+
6	cluster clusters proximity spatial geographical spillovers geographic clustering agglomeration geography	Clusters	
7	centers federal defense military laboratory laboratories mission agencies agency sbir	Defense-related R&D	+
8	crisis safety security resilience disaster emergency adaptive crises event threat	Crisis Management	+
9	century modern revolution book principle mind societies economists mass living	Technological Revolution	+
10	city cities smart urban smart_city citizens housing smart_cities buildings residents	Smart City	+
11	absorptive tacit competencies knowledge_transfer routines knowledge_management dynamic_capabilities organizational_learning new_knowledge codification	Knowledge Management	
12	disruptive incumbent incumbents entrants discontinuous dominant_design designs technological_change mature inertia	Disruptive Technologies	
13	social_media big_data company_A tourism media company_B sites news websites website	Social Media	+
14	delphi round consensus panel opinion statements opinions rounds forecasting delphi_method	Delphi Method	+
15	students women faculty gender career student female teaching male graduate	Academic Productivity	+

Table 17 *(cont'd)*

16	cross-functional project_management npd_process front-end front_end launch npd_projects project_team proficiency execution	NPD Projects	
17	option cash valuation real_options asset profitability intangible volatility market_value earnings	Project Valuation	+
18	fuzzy algorithm weights optimization alternatives optimal scheduling weight simulation criterion	Scheduling	+
19	designers virtual objects prototype prototypes designer prototyping designs language object	Design	
20	biotechnology pharmaceutical drug biotech drugs clinical trials discovery pharmaceuticals gene	Biotechnology	
21	communities crowdsourcing open_source innovators code developers contest contests crowd solvers	Community Support	+
22	academics patenting faculty knowledge_transfer commercialization academia ttos engagement university-industry scientist	University-Industry Relations	
23	licensing commercialization license invention inventions licenses faculty royalty disclosure licensed	Licensing	
24	intention usage acceptance attitude usefulness attitudes satisfaction intentions privacy ease_of_use	Technology Acceptance	
25	institutes basic_research centres public_research co-operation applied_research laboratories labs budget profile	Research and Development	
26	incubator incubators spin-offs spin-off incubation ntbfs start-ups science_parks parks start-up	Incubation	
27	platform ecosystem platforms intermediaries ecosystems intermediary digital developers providers business_ecosystem	Ecosystems	+

Table 17 *(cont'd)*

28	workers mobility jobs labor skill human_capital skilled workforce wage employee	Labor Market	+
29	nanotechnology nano physics cern nanotechnologies nanotech fusion emerging_technologies sensor properties	Emerging Technologies	
30	agents adopters simulation agent adopter simulations system_dynamics technology_adoption modelling fitness	Diffusion (Technology Adoption)	
31	programme programmes commission public_sector agencies governments priorities oecd priority council	Policy-Making	
32	vehicle vehicles transport fuel cars automotive hybrid hydrogen mobility road	Motor Vehicle Industry	+
33	brand launch advertising brands reputation segment segments selling signals pricing	Brand	+
34	protection appropriability imitation protect appropriation iprs secrecy trademarks litigation patenting	Intellectual Property Rights (IPRs)	
35	legitimacy identity discourse collective logics logic norms framing tensions politics	(Politics of) Technology Assessment	
36	spillovers export expenditure spillover domestic expenditures exports goods imports machinery	International Technology Diffusion	+
37	diversification survival acquisitions acquired mergers exit hazard acquiring relatedness merger	Mergers and Acquisitions	+
38	organisations organisational organisation behaviour analysed analyse characterised commercialisation specialised specialisation	Organisational Behaviour	

Table 17 *(cont'd)*

39	smes small_firms innovators large_firms firm_size innovate innovating kibs manufacturing_firms turnover	Firm Size	
40	convergence configuration configurations causal quadrant negotiation converging configurational archetypes convergent	Technology Convergence	+
41	entrepreneurial entrepreneurs ventures entrepreneur venture venturing founders new_venture start-up start-ups	Entrepreneurship	
42	exploitation openness exploratory ambidexterity breadth focal exploitative slack depth moderating	Organisational Ambidexterity	+
43	ownership family board executives owners directors ceos shareholders corporate_governance owner	Corporate Governance	+
44	citations citation cited disciplines interdisciplinary bibliometric library disciplinary discipline sciences	Bibliometrics	
45	e-business alignment e-commerce agility lean deployment certification delivery automation adopting	Strategic Alignment	+
46	game digital media music video players games film movie player	Digital Media	+
47	alliance partner collaborations partnerships cooperative agreements partnership relational interorganizational consortia	Alliances	
48	distance class novelty classes similarity cloud computing pairs distant pair	Classification (Supervised Learning)	+
49	centrality nodes density node connections embeddedness network_analysis holes actor network_structure	Network Analysis	

Table 17 *(cont'd)*

50	electricity wind solar renewable_energy grid biomass heat renewable deployment nuclear	Renewable Energy	+
51	banks bank banking rural icts credit income poverty inclusive developing_countries	Subsistence Economies	+
52	evolutionary mode sectoral trajectories technological_change modes trajectory paradigm systemic accumulation	Technological Change	
53	green sustainability regulatory regulation iste recycling pollution regulations eco-innovation sustainable_development	Sustainability	
54	supplier outsourcing buyer contract contracts client buyers sourcing purchasing procurement	Buyer-Supplier Relationships	
55	regime transitions niche sustainability regimes niches socio-technical water landscape societal	Sustainability Transitions	+
56	subsidiaries subsidiary internationalization home mncs overseas abroad mnes multinational domestic	Multinational Enterprises (MNEs)	+
57	patenting inventors citations invention inventions inventor granted citation number_of_patents inventive	Patents (IPRs)	
58	percent corporation business_units spending metrics breakthrough revenue executives executive business_unit	Portfolio Management	
59	innovativeness constructs item scales moderating antecedents correlations latent loadings indirect	Organizational Innovativeness	
60	investors financing venture_capital finance equity crowdfunding startups venture fund investor	Startup Financing	

Table 17 *(cont'd)*

61	oecd capita income nations developing_countries world_bank country_A human_capital expenditure country_B	National Systems	
62	consumers preferences preference purchase attribute subjects experiment resistance consumption hedonic	Consumer Evaluation	
63	plant aircraft plants machine steel engine machines maintenance mechanical metal	Metallurgical Industry	+
64	roadmap roadmapping roadmaps intelligence maturity technology_management technology_development technology_roadmapping layer technology_roadmap	Roadmapping	
65	subsidies subsidy matching treatment additionality grants treated credits credit control_group	Subsidies	+
66	food agricultural water farmers agriculture land crop forest farm soil	Food Industry	
67	creativity cognitive personality emotional cognition style psychology memory intuition traits	Creativity	
68	architecture modularity modular modules module interfaces interface subsystems architectural architectures	Product Architecture	
69	dummy estimation regression dummies regressions specification explanatory control_variables standard_errors heterogeneity	Estimation Methods	+
70	score scores metrics ranking outputs rank rating rankings proposals evaluations	Decision-Making	
71	semiconductor electronics computers chip optical memory devices manufacturers hardware company_C	Semiconductor Industry	+

Table 17 *(cont'd)*

72	interview interviewees meetings interviewed in-depth documents coding interviewee contact decided	Data Collection Methods	+
73	mining topics text documents clusters keyword cluster clustering frequency discovery	Text Mining	+
74	stakeholders stakeholder engagement societal outputs triple_helix health dialogue library practitioners	Stakeholder Management	+
75	satisfaction employee rewards climate autonomy leader career intrinsic professionals reward	Career System	+
76	mobile standardization telecommunications broadband standardisation phone wireless providers switching mobile_phone	Telecommunications Industry	+
77	supply_chain manufacturer chain manufacturers inventory rfid logistics retailer retail supply_chains	Supply Chain Management	
78	optimal utility equilibrium proposition profits incentive marginal prices revenue game	Pricing	
79	foresight scenario scenarios futures exercise uncertainties workshop workshops backcasting visions	Foresight	

In Table 17, *Topic*, *Top Terms* and *Label* columns have the same semantics as in Table 16. In this table, the additional *New* column denotes whether the given low-level topic was extracted in any of the previous studies or not. The issues that attract attention or require explanation in the low-level topics are as follows:

- According to the *New* column, the total number of topics that are extracted for the first time is 38.

- The most studied industries are Defense Industry (7), Motor Vehicle Industry (32), Digital Media Industry (46), Metallurgical Industry (63), Food Industry (66), Semiconductor Industry (71) and Telecommunications Industry (76).

- The most used methods are Bibliometrics (44), Classification (48), Network Analysis (49), Estimation (69) and Text Mining (73). Delphi Method (14) is not only a method used in the studies but also a topic covered in studies on the method itself.

- Community Support (21) and Classification (48) are determined as the surprising topics of the study. As a subjective assessment, it is an unexpected result to have these low-level topics among the 80 most studied topics of the TM domain.

- There are several closely related topics like Labor Market (28) and Career System (75), Diffusion (30) and International Technology Diffusion (36), and Sustainability (53) and Sustainability Transitions (55). These topics can be considered as clusters as in the ones clustered in the high-level topics.

- Ecosystems (27) includes issues such as platform ecosystems, business ecosystems and innovation ecosystems, and shows a sharply increasing trend in the last period.

- The topic with the label Defense-related R&D (7) means that defense industry is one of the main drivers of R&D. However, it shows a decreasing trend which is compatible with the decrease in the trend of Research and Development (25).

- There are three topics with words from different domains in their top terms:

    - Social Media (13) with *tourism* denotes that there are many articles which studied social media in tourism industry.

    - Firm Size (39) with words related to innovation shows that many of the researchers investigated the relationship between innovation and firm size.

    - Subsistence Economies (51) with *icts* denotes that various studies analyzed the role or contribution of ICT in subsistence economies.

- The issues related to trends of topics are as follows:

- Topics with sharply increasing trends are Smart City (10), Ecosystems (27) and Organizational Ambidexterity (42).
- Topic with a sharply decreasing trend is Strategic Alignment (45).
- Topics with increasing trends are University-Industry Relations (22), Entrepreneurship (41), Sustainability (53), Sustainability Transitions (55), Estimation Methods (69), Text Mining (73) and Stakeholder Management (74).
- Topics with decreasing trends are Defense-related R&D (7), NPD Projects (16), Project Valuation (17), Research and Development (25), Technological Change (52), Portfolio Management (58), Metallurgical Industry (63), Semiconductor Industry (71) and Data Collection Methods (72).

Lastly, as a note for those who are interested in topic modeling, types of the top 100 documents of each topic in the low-level topics (a total of 8,000 top documents) are examined and it is determined that there are only three abstracts in the list. This shows that adding abstracts to a corpus composed of full-text articles does not make substantial contribution to the topics extracted from the corpus.

CHAPTER 6

CONCLUSION AND FUTURE WORK

TM is a new scientific field that emerged at the second half of the 20th century. National Research Council (1987) defines TM as an emerging field that "links engineering, science, and management disciplines to plan, develop, and implement technological capabilities to shape and accomplish the strategic and operational objectives of an organization". It is a new and interdisciplinary field whose definition and scope continue to be researched.

The most difficult problem in an interdisciplinary field is to define its scope. One way of solving this problem is to use topic modeling. Its purpose is to automatically extract latent topics from document collections. Topic modeling can be applied to a specific scientific field to obtain an overall view of its topics and trends.

Hence, the purpose of this study is to reveal the topics studied in the TM domain by using a topic model, namely LDA (Blei et al., 2003). It is applied to a corpus of 16,180 research articles/abstracts published in 13 top specialty journals of TM between 1997 and 2019.

However, LDA has two problems: how to optimize model parameters and how to eliminate topic instability. So, firstly, an empirical investigation is conducted to evaluate possible solutions to these problems. As a result of the experiments conducted, it is found out that semantic similarity measures can be used to optimize the LDA parameters. However, instead of using a single LDA model, average scores of replicated models must be used to avoid the negative effects of topic instability.

Then, by using the method proposed, topic modeling is applied to Technology Management (TM) domain. Topics are extracted at two levels, namely the high-level

and the low-level topics. The optimum number of topics are found to be 13 for the high-level topics, and 80 for the low-level topics. In the high-level topics, three basic clusters are determined: Product Development, Universities' Role and Policy Making which could be thought as the top three topics of TM literature. In the low-level topics, topics that a not extracted in any of the previous studies, the most studied industries, the most used methods and surprising topics are identified. For both levels, trends of topics are extracted on an annual basis and the topics with increasing and decreasing trends are reported.

The results presented in this study provide valuable information for editors, authors and policy makers in TM domain since with this information, the following actions are facilitated: 1) for editors to define the scope of the conference/journal they manage, 2) for researchers to conveniently plan their future studies, and 3) for policy makers to plan their future strategies.

This study contributes to the understanding of the TM domain by resolving the confusion on its scope. With the topics identified in the study, the boundaries of the discipline have become more salient.

The basic limitation of the study is about the coverage of full texts. Since the articles published in IJTM are not publicly accessible, only the abstracts from this journal could be downloaded and used. The results of the conducted analysis show that adding abstracts to our corpus composed of full-text articles does not provide substantial contribution to the topics extracted from the corpus. Therefore, in our study, the use of abstracts from one of the journals may have prevented the extraction of several topics.

The results obtained in this study pave the way for future research studies in TM in several directions. Researchers can investigate the topics studied in each of the journals and make a comparative analysis with the topics and trends extracted in this study. This will contribute to the policies and specializations of the journals. As a second research direction, country-based comparative analyses can be made to assess and contribute to regional policies.

Policy Recommendation on Academic Entrepreneurship

Since the 1990s, universities have been given a "third mission" which aims at contributing to economic growth and development, in addition to their teaching and research missions (Guimón, 2013). The reason is explained as the “failures in the market of knowledge, suggesting that inventor entrepreneurship is a second-best solution to the commercialization of new technology” (Shane, 2002) This new mission led to the emergence of a new concept: entrepreneurial university, which is “a dynamic system, which includes special inputs (such as Resources, Rules and Regulations, Entrepreneurial capabilities, and Expectations), processes (such as Teaching, Research and Development, Innovation, Commercialization, Funding, Networking) and outputs (such as Entrepreneur Human Resources, Entrepreneurial Networks, Entrepreneurial Centers)" (Salarnzadehl et al. 2011).

One of the most important roles in an entrepreneurial university, maybe the most important one, is the academic entrepreneur. It is a term used to describe academics who go beyond the production of knowledge and become active participants in designing new products and take role in commercialization (Henrekson and Rosenberg, 2000). Despite the importance of this role, there is an ongoing debate among academics about the impacts (the good, the bad and the challenging) of activities undertaken by academic entrepreneurs on academic culture (Baycan and Stough, 2013). Some examples from “the bad” are given as: 1) conflicts in values (a shift from “public good” to “academic capitalist”), 2) differences in culture and perspectives (production of knowledge and scientific excellence versus valorization of knowledge and generation of wealth), 3) conflicts of interest (openness versus secrecy) and 4) conflicts in the commercialization of knowledge (ethics).

There are different ways to transfer knowledge to industry such as academic entrepreneurship, licensing, consultation, and collaborative research. In this study, three of them, namely academic entrepreneurship (4 in high level topics), licensing (23 in low level topics) and research and development (can also be labelled as collaborative research, 25 in low level topics), are determined as three topics of TM.

Looking at the trends, it is seen that academic entrepreneurship showed a slightly increasing trend while licensing and research and development showed decreasing trends. The trends are consistent in themselves since the level of licensing and collaboration is decreasing due to the increasing entrepreneurial activities of academics. However, considering "the bad" given in the previous paragraph, "encouraging faculty members to become economic entrepreneurs may not be the best way to bolster university-industry collaboration" (Konaç, 2018). Thus, keeping a balance between the different ways of knowledge transfer and taking measures for the prevention of the bad might be more beneficial for the long-term success of the university and the industry. One way of achieving this may be to support alumni (or graduate) entrepreneurship (Beyhan and Findik, 2018), which benefits from the strong relationship between alumni and universities.

Comments on the Reason for the Downtrend of Product Development

Product development is defined as the process of converting needs into a technical and commercial solution. The aim is the production of products or services with higher quality, lower price, and shorter response times. Although all product development processes are different from each other, they also have some similar features or elements that can be used to manage the process. (Smith and Morrow, 1999). Here is where the product development process models come into play: managing the shared functions of product development process. There are two types of models used in the process (Li et al., 2019): 1) sequential such as Stage-gate (Cooper, 1994) which is a composition of a series of stages and gates, and 2) spiral such as Agile (Beck et al., 2001) which is an iterative process of development activities.

Product development process models were very popular in the early 2000s, especially in software industry. Initially, Waterfall model, which is the equivalent of the Stage-gate model in the software industry, was used for the development of software products. Later, a new methodology called Agile Software Development was developed to overcome the shortcomings of the Waterfall model. Lastly, Agile was combined with Stage-gate. After seeing the successful applications of the models in the software industry, they have begun to be applied in other industries.

Since product development process models have been developed, used and succeeded in the software industry, they have been adapted from the software industry and there has been no need for a new model in other domains. Additionally, due to the increasing speed of technological development, the interest has shifted from how-to-develop to what-to-develop. As a result of these factors, trend of the topic has decreased gradually from a very high level to the level of the others.

REFERENCES

Agrawal, A., Fu, W., Menzies, T. “What is wrong with topic modeling? and how to fix it using search-based software engineering”, Information and Software Technology, 2018.

Aletras, N., Stevenson, M. “Evaluating topic coherence using distributional semantics,”, Proceedings of the The 10th International Conference on Computational Semantics, pp. 13–22, 2013.

Ali, M., Törn, A. “Population set-based global optimization algorithms: some modifications and numerical studies”, Computers and Operations Research, vol. 31, no. 10, pp. 1703–1725, 2004.

Allen, T. J., Sosa, M. L. “50 Years of Engineering Management Through the Lens of the IEEE TRANSACTIONS”, IEEE Transactions On Engineering Management, 2004.

Antons, D., Kleer, R., Salge, T. O. “Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories”, The Journal of Product Innovation Management, 2016.

Arun, R., Suresh, V., Madhavan, C. E. V., Murty, M. N. “On finding the natural number of topics with latent dirichlet allocation: Some observations”, Advances in Knowledge Discovery and Data Mining (PAKDD), M. Zaki, J. Yu, B. Ravindran, and V. Pudi, Eds. Springer, Berlin, Heidelberg, 2010.

Ball, D. F., Rigby, J. “Disseminating research in management of technology: journals and authors”, R&D Management, 2006.

Baycan, T., Stough, R. R. “Bridging knowledge to commercialization: the good, the bad, and the challenging”, The Annals of Regional Science, 2013.

Beck, K., Grenning, J., Martin, R. C., Beedle, M., Highsmith, J., Mellor, S., Bennekum, A., Hunt, A., Schwaber, K., Cockburn, A., Jeffries, R., Sutherland, J., Cunningham, W., Kern, J., Thomas, D., Fowler, M., Marick, B. “Manifesto for Agile Software Development”, http://agilemanifesto.org/, Retrieved 1 November 2021.

Beyhan, B., Cetindamar, D. “No escape from the dominant theories: The analysis of intellectual pillars of technology management in developing countries”, Technological Forecasting & Social Change, 2011.

Beyhan, B., Findik, D. “Student and graduate entrepreneurship: ambidextrous universities create more nascent entrepreneurs”, The Journal of Technology Trasnsfer, 2018.

Biemans, W., Griffin, A., Moenaert, R. “Twenty Years of the Journal of Product Innovation Management: History, Participants, and Knowledge Stock and Flows”, The Journal of Product Innovation Management, 2007.

Binkley, D., Heinz, D., Lawrie, D., Overfelt, J. “Understanding lda in source code analysis”, Proceedings of the 22nd International Conference on Program Comprehension, 2014.

Blei, D. M., Ng, A. Y., Jordan, M. I. “Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

Blei, D. M. “Probabilistic topic models” Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S. “A density-based method for adaptive lda model selection”, Neurocomputing, 2009.

Cetindamar, D., Phaal, R., Probert, D. “Understanding technology management as a dynamic capability: A framework for technology management activities”, Technovation, 2009a.

Cetindamar, D., Wasti, S.N., Ansal, H., Beyhan, B. “Does technology management research diverge or converge in developing and developed countries?”, Technovation, 2009b.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D. M. “Reading tea leaves: How humans interpret topic models”, Advances in Neural Information Processing Systems, 22, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009.

Chuang, J., Gupta, S., Manning, C. D., Heer, J. “Topic model diagnostics: Assessing domain relevance via topical alignment”, Proceedings of the 30th International Conference on Machine Learning, 2013.

Cheng, C. H., Kumar, A., Motwani, J. G., Reisman, A., Mada, M. S. “A Citation Analysis of the Technology Innovation Management Journals”, IEEE Transactions On Engineering Management, 1999.

Choi, D. G., Lee, Y., Jung, M., Lee, H. “National characteristics and competitiveness in MOT research: A comparative analysis of ten specialty journals, 2000–2009”, Technovation, 2012.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., Heer, J. “Topiccheck: Interactive alignment for assessing topic model stability”, The 2015 Annual Conference of the North American Chapter of the ACL, pp. 175–184, 2015.

Das, S., Suganthan, P. N. “Differential evolution: A survey of the state-of-the-art” IEEE Transactions on Evolutionary Computation, vol. 15, no. 1, pp. 4–31, 2011.

Cooper, R. G. “PERSPECTIVE: Third-Generation New Product Processes” Journal of Product Innovation Management, 1994.

Feoktistov, V., Janaqi, S. “Generalization of the strategies in differential evolution”, 18th International Parallel and Distributed Processing Symposium, 2004.

Franke, N., Schreier, M. “A Meta-Ranking of Technology and Innovation Management / Entrepreneurship Journals”, Die Betriebswirtschaft (DBW), 2008.

Greene, D., OCallaghan, D., Cunningham, P. “How many topics? stability analysis for topic models”, Machine Learning and Knowledge Discovery in Databases (ECML PKDD), T. Calders, F. Esposito, E. Hllermeier, and R. Meo, Eds. Springer, Berlin, Heidelberg, 2014.

Gregory, M. J. “Technology management: a process approach”, Proceedings of the Institution of Mechanical Engineers, 1995.

Griffiths, T. L., Steyvers, M. “Finding scientific topics”, Proceedings of the National Academy of Sciences (PNAS), 2004.

Gudanowska, A. E. “A Map of Current Research Trends within Technology Management in the Light of Selected Literature”, Management and Production Engineering Review, 2017.

Guimón, J. “Promoting University-Industry Collaboration in Developing Countries”, World Bank, 2013.

Henrekson, M., Rosenberg, N. “Incentives for Academic Entrepreneurship and Economic Performance: Sweden and the United States”, IUI Working Paper, No. 530, The Research Institute of Industrial Economics (IUI), 2000.

Huang, Y., Ding, X., Liu, R., He, Y., Wu, S. “Reviewing the Domain of Technology and Innovation Management: A Visualizing Bibliometric Analysis”, SAGE Open, 2019.

Hunter, J. D. “Matplotlib: A 2D Graphics Environment”, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Jurafsky, D., Martin, J. H. “Speech and Language Processing”, 2nd ed. Prentice Hall, 2008.

Kim, J. H., Chen, W. “Research Topic Analysis in Engineering Management Using a Latent Dirichlet Allocation Model”, Journal of Industrial Integration and Management, 2018.

Koltcov, S., Koltsova, O., Nikolenko, S. “Latent dirichlet allocation: Stability and applications to studies of user-generated content”, Proceedings of the 2014 ACM conference on Web science, pp. 161–165, 2014.

Konaç, E.H. “Academic entrepreneurs: motivational aspects, challenges and success criteria in technology development zones in Ankara”, Thesis (M.S.), Middle East Technical University, 2018.

Krasnov, F., Sen, A. “The number of topics optimization: Clustering approach”, Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 416–426, 2019.

Lee, H. “Uncovering the multidisciplinary nature of technology management: journal citation network analysis”, Scientometrics, 2015.

Lee, H., Kang, P. “Identifying core topics in technology and innovation management studies: a topic model approach”, Journal of Technology Transfer, 2018.

Li, Y. Roy, U., Saltz J. S. “Towards an integrated process model for new product development with data-driven features (NPD3)”, Research in Engineering Design, 2019.

Liker, J. “Results of Survey of Management Journals for TIM Research”, TIM Newsletter, 1995.

Linton, J. D., Thongpapanl, N. “PERSPECTIVE: Ranking the Technology Innovation Management Journals”, The Journal of Product Innovation Management, 2004.

Linton, J. D., Embrechts, M. “MOT TIM journal rankings 2006”, Technovation, 2007.

Linton, J. D. “Technology innovation management’s growing influence and impact”, Technovation, 2009.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D. “The stanford corenlp natural language processing toolkit”, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, 2014.

Mantyla, M. V., Claes, M., Farooq, U. “Measuring lda topic stability from clusters of replicated runs”, Proceedings of the 12th ACM/IEEE International

Symposium on Empirical Software Engineering and Measurement, pp. 1–4, 2018.

McCallum, A. K. “Mallet: A machine learning for language toolkit”, http://mallet.cs.umass.edu 2002.

Mehta, V., Caceres, R. S., Carter, K. M. “Evaluating topic quality using model clustering”, IEEE Symposium on Computational Intelligence and Data Mining, 2014.

Merino, M. T. G., Carmo, M. L. P., Alvarez, M. V. S. “25 Years of Technovation: Characterisation and evolution of the journal”, Technovation, 2006.

Mikolov, T., Chen, K., Corrado, G., Dean, J. “Efficient Estimation of Word Representations in Vector Space", Workshop of the International Conference on Learning Representations (ICLR), 2013.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., Mc-Callum, A. “Optimizing semantic coherence in topic models”, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272, 2011.

Naili, M., Chaibi, A. H., Ghezala, H. H. B. “Comparative study of word embedding methods in topic segmentation,” Procedia Computer Science, vol. 112, pp. 340–349, 2017.

National Research Council, “Management of Technology The Hidden Competitive Advantage”, National Academy Press, 1987.

Newman, D., Lau, J. H., Grieser, K., Baldwin, T. “Automatic evaluation of topic coherence”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 100–108, 2010.

Nikolenko, S. I. “Topic quality metrics based on distributed word representations”, The annual international Special Interest Group on Information Retrieval (SIGIR) conference, pp. 1029–1032, 2016.

Page, A. L., Schirr, G. R. “Growth and Development of a Body of Knowledge: 16 Years of New Product Development Research, 1989–2004”, The Journal of Product Innovation Management, 2008.

Panichella, A., Dit, B., Oliveto, R., Penta, M. D., Poshynanyk, D., Lucia, A. D. “How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms”, 35th International Conference on Software Engineering (ICSE), 2013.

Pilkington, A., Teichert, T. “Management of technology: themes, concepts and relationships”, Technovation, 2006.

Pennington, J., Socher, R., Manning, C. D. “Glove: Global vectors for word representation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.

Qin, A. K., Huang, V. L., Suganthan, P. N. “Differential evolution algorithm with strategy adaptation for global numerical optimization”, IEEE Transactions On Evolutionary Computation, vol. 13, no. 2, pp. 398–417, 2009.

Rder, M., Both, A., Hinneburg, A. “Exploring the space of topic coherence measures”, International Conference on Web Search and Data Mining, 2015.

Řehůřek, R., Sojka, P. “Software framework for topic modelling with large corpora”, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, 2010.

Rosner, F., Hinneburg, A., Rder, M., Nettling, M., Both, A. “Evaluating topic coherence measures”, Topic Models: Computation, Application and Evaluation workshop at the Neural Information Processing Systems conference, 2013.

Salarnzadehl, A., Salarnzadeh, Y., Daraei, M. R. “Toward a Systematic Framework for an Entrepreneurial University: A Study in Iranian Context with an IPOO Model”, Global Business and Management Research, 2011.

Shane, S. “Selling University Technology: Patterns from MIT”, Management Science, Vol. 48, No. 1, pp. 122-137, 2002.

Smith, R. P., Morrow, J.A. “Product development process modeling”, Design Studies, 1999.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D. “Exploring topic coherence over many models and many topics”, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961, 2012.

Steyvers, M., Griffiths, T. “Latent Semantic Analysis: A Road to Meaning”, Laurence Erlbaum, 2007, chapter Probabilistic Topic Models.

Storn, R., Price, K. “Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", Journal of Global Optimization, 1997.

Syed, S., Spruit, M. “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation", International Conference on Data Science and Advanced Analytics, IEEE, 2017.

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. “Hierarchical dirichlet processes”, Journal of the American Statistical Association, 2006.

Tekin, Y. “Optimization of LDA parameters”, IEEE Conference on Signal Processing and Communications Applications (SIU), 2020.

Thongpapanl, N. “The changing landscape of technology and innovation management: An updated ranking of journals in the field”, Technovation, 2012.

Wikipedia “Latent dirichlet allocation”, [Online] Available: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation, 2021.

Zavitsanos, E., Petridis, S., Paliouras, G., Vouros, G. A. “Determining automatically the size of learned ontologies”, 18th European Conference on Artificial Intelligence, M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. Avouris, Eds. IOS Press, 2008.

APPENDICES

A. TOP DOCUMENTS FOR HIGH-LEVEL TOPICS

Table A.1. Documents and Publication Dates

Topic	Top Document	Year
0	Do applicant patent citations matter?	2013
1	Antecedents of Team Intuition and Its Impact on the Success of New Product Development Projects	2011
2	User preferences of social features on social commerce websites: An empirical study	2015
3	Modularity as a Strategy for Supply Chain Coordination: The Case of U.S. Auto	2007
4	Determinants and consequences of university spinoff activity: a conceptual framework	2008
5	Water demands for electricity generation in the U.S.: Modeling different scenarios for the water–energy nexus	2015
6	Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology	2019
7	Predicting the diffusion of LCD TVs by incorporating price in the extended Gompertz model	2013
8	The Role of Multinational Corporations and National States in the Globalization of Innovatory Capacity: The European Perspective	2000
9	Do public subsidies stimulate private R&D spending?	2008
10	A bibliometric study of productivity and impact of modern language and literature research	2011

Table A.1 *(cont'd)*

11	Thinking parliamentary technology assessment politically: Exploring the link between democratic policy making and parliamentary TA	2019
12	Trading knowledge for status: Conceptualizing R&D alliance formation to achieve ambidexterity	2019

B. TRENDS FOR HIGH-LEVEL TOPICS

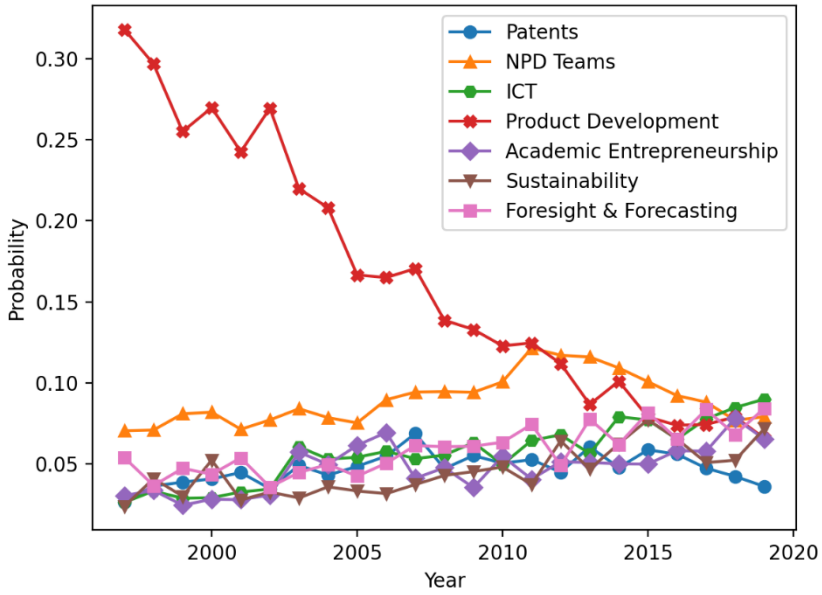


Figure B.1. Trends for High-Level Topics (0-6)

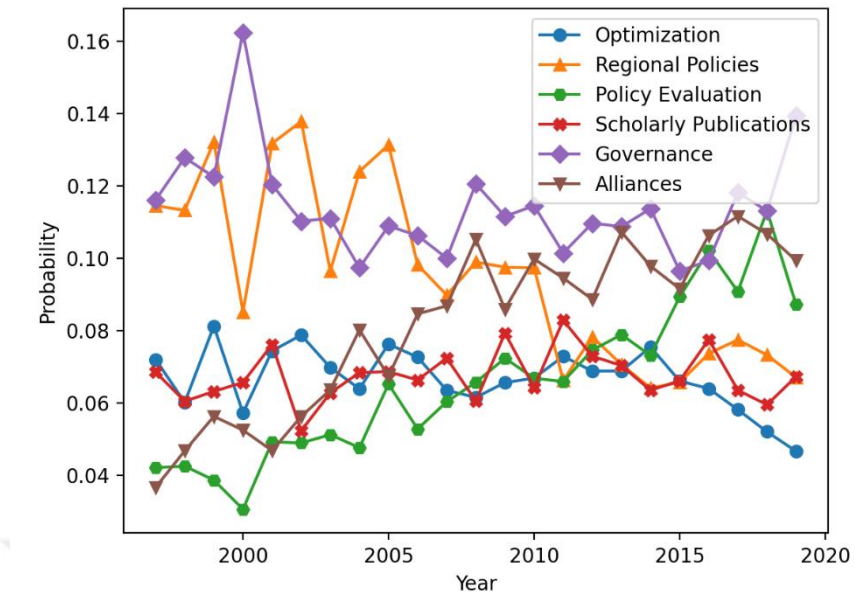


Figure B.2. Trends for High-Level Topics (7-12)

C. TOP DOCUMENTS FOR LOW-LEVEL TOPICS

Table C.1. Documents and Publication Dates

Topic	Top Document	Year
0	Predicting the diffusion of LCD TVs by incorporating price in the extended Gompertz model	2013
1	Differences between learning processes in small tigers and large dragons Learning processes of two color TV (CTV) firms within China	2003
2	Early diagnostics and Alzheimer's disease: Beyond ‘cure’ and ‘care’	2015
3	Servitization and Industry 4.0 convergence in the digital transformation of T product firms: A business model innovation perspective	2019
4	When Team Identity Helps Innovation and When It Hurts: Team Identity and Its Relationship to Team and Cross-Team Innovative Behavior	2018
5	Making or breaking climate targets: The AMPERE study on staged accession scenarios for climate policy	2015
6	Clusters, convergence, and economic performance	2014
7	Defense-related R&D as a model for “Grand Challenges” technology policies	2012
8	A holistic framework for building critical infrastructure resilience	2016
9	Global Commons in the Global Brain	2017
10	Strategic principles for smart city development: A multiple case study analysis of European best practices	2019
11	Constructing a strategy on the creation of core competencies for African companies	2018

Table C.1 *(cont'd)*

12	Ex-ante evaluation of disruptive susceptibility in established value networks—When are markets ready for disruptive innovations?	2013
13	Polarization and acculturation in US Election 2016 outcomes – Can twitter T analytics predict changes in voting preferences	2019
14	Biases in future-oriented Delphi studies: A cognitive perspective	2016
15	Academic outcomes among principal investigators, co-principal investigators, and non-PI researchers	2014
16	Measuring the Knowns to Manage the Unknown: How to Choose the Gate Timing Strategy in NPD Projects	2017
17	Real options valuation: the new frontier in R&D project evaluation?	1999
18	On the Robust and Stable Flowshop Scheduling Under Stochastic and Dynamic Disruptions	2017
19	Strategies of Innovation and Imitation of Product Languages	2007
20	The myth of the biotech revolution: An assessment of technological, clinical and organisational change	2007
21	Monetary donations to an open source software platform	2009
22	Determinants and public policy implications of academic- industry knowledge transfer in life sciences: a review and a conceptual framework	2016
23	Objectives, Characteristics and Outcomes of University Licensing: A Survey of Major U.S. Universities	2001
24	Chat now... Examining the variables influencing the use of online live chat	2019
25	The point of view of firms in Minas Gerais about the contribution of universities and research institutes to R&D activities	2012
26	The role of incubator interactions in assisting new ventures	2010
27	Platform design framework: conceptualisation and application	2018
28	Unintended consequences on gender diversity of high-tech growth and labor market polarization	2018

Table C.1 *(cont'd)*

29	Nanotechnology innovation system: Understanding hidden dynamics of nanoscience fusion trajectories	2009
30	MADness in the method: On the volatility and irregularity of technology diffusion	2016
31	Policy-making in science policy: The 'OECD model' unveiled	2013
32	The emergence of hybrid-electric cars: Innovation path creation through co-evolution of supply and demand	2010
33	Assessing Consequences of Component Sharing across Brands in the Vertical Product Line in the Automotive Market	2012
34	Risk factors and mechanisms of technology and insignia copying—A first empirical approach	2012
35	Thinking parliamentary technology assessment politically: Exploring the link between democratic policy making and parliamentary TA	2019
36	Use Tables for Imported Goods and Valuation Matrices for Trade Margins—an Integrated Approach for the Compilation of the Belgian 1995 Input–Output Tables	2004
37	Acquisitions of small high-tech firms as a mechanism for external knowledge sourcing: The integration-autonomy dilemma	2017
38	Organisational conditions for service encounter-based innovation	2013
39	Making sense of innovation by R&D and non-R&D innovators in low technology contexts: A forgotten lesson for policymakers	2011
40	Patent-based semantic measurement of one-way and two-way technology convergence: The case of ultraviolet light emitting diodes (UV-LEDs)	2019
41	Properties of opportunity creation and discovery: Comparing variation in T contexts of innovativeness	2019
42	Organizational Learning Ambidexterity, Strategic Flexibility, and New Product Development	2014
43	The determinants of financial fraud in Chinese firms: Does corporate governance as an institutional innovation matter?	2017

Table C.1 *(cont'd)*

44	Scholarly book publishing: Its information sources for evaluation in the social sciences and humanities	2017
45	Key Dimensions of Inhibitors for the Deployment of Web-Based Business-to-Business Electronic Commerce	2006
46	Hardware gimmick or cultural innovation? Technological, cultural, and social foundations of the Japanese video game industry	2003
47	Formalization, Communication Quality, and Opportunistic Behavior in R&D Alliances between Competitors	2014
48	Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records	2015
49	The firm's knowledge network and the transfer of advice among corporate inventors—A multilevel network study	2017
50	The bumpy road of biomass gasification in the Netherlands: Explaining the rise and fall of an emerging innovation system	2008
51	Toward A Theory on the Reproduction of Social Innovations in Subsistence Marketplaces	2019
52	The market failure and the systemic failure rationales in technological innovation systems	2013
53	Product recovery decisions within the context of Extended Producer Responsibility	2014
54	Collaborating with Suppliers in Product Development: A U.S. and Japan Comparative Study	1999
55	The co-evolution of policy mixes and socio-technical systems: Towards a conceptual framework of policy mix feedback in sustainability transitions	2019
56	Decentralised R&D and strategic competitiveness: globalised approaches to generation and use of technology in multinational enterprises (MNEs)	1999
57	Patent claims and patent scope	2019

Table C.1 *(cont'd)*

58	Perspective: The Stage-Gates Idea-to-Launch Process—Update, What's New, and NexGen Systems	2008
59	Strategic Orientation and Product Innovation: Exploring a Decompositional Approach	2012
60	The evolution of equity crowdfunding: Insights from co-investments of angels and the crowd	2019
61	Patterns of restructuring in research, development and innovation activities in central and eastern European countries: an analysis based on S&T indicators	1999
62	Innovation Aesthetics: The Relationship between Category Cues, Categorization Certainty, and Newness Perceptions	2013
63	Open versus closed innovation: development of the wide strip mill for steel in the United States during the 1920s	2009
64	An empirical analysis of the determinants of technology roadmap utilization	2011
65	Additionality or crowding-out? An overall evaluation of public R&D subsidy on private R&D expenditure	2016
66	Why new crop technology is not scale-neutral-A critique of the expectations for a crop-based African Green Revolution	2016
67	Using Intuition in Fuzzy Front-End Decision-Making: A Conceptual Framework	2013
68	Capturing the Degree of Modularity Embedded in Product Architectures	2006
69	Long-run versus short-run decisions: R&D and market structure in Spanish firms	2009
70	Examining the value added by committee discussion in the review of applications for research awards	2007
71	Changes in the technology spillover structure due to economic paradigm shifts: A driver of the economic revival in Japan's material industry beyond the year 2000	2009

Table C.1 *(cont'd)*

72	Handbooks as a tool for organizational learning: a case study	1998
73	Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology	2019
74	Exploring transdisciplinary integration within a large research program: Empirical lessons from four thematic synthesis processes	2017
75	A Contemporary Justice Perspective on Dual Ladders for R&D Professionals	2016
76	An assessment of Gigabit Ethernet technology and its applications at the NASA Glenn Research Center: a case study	2003
77	In-Store Pickup and Returns for a Dual Channel Retailer	2017
78	Trade-In Rebates for Price Discrimination and Product Recovery	2016
79	A comprehensive scenario intervention typology	2019

D. TRENDS FOR LOW-LEVEL TOPICS

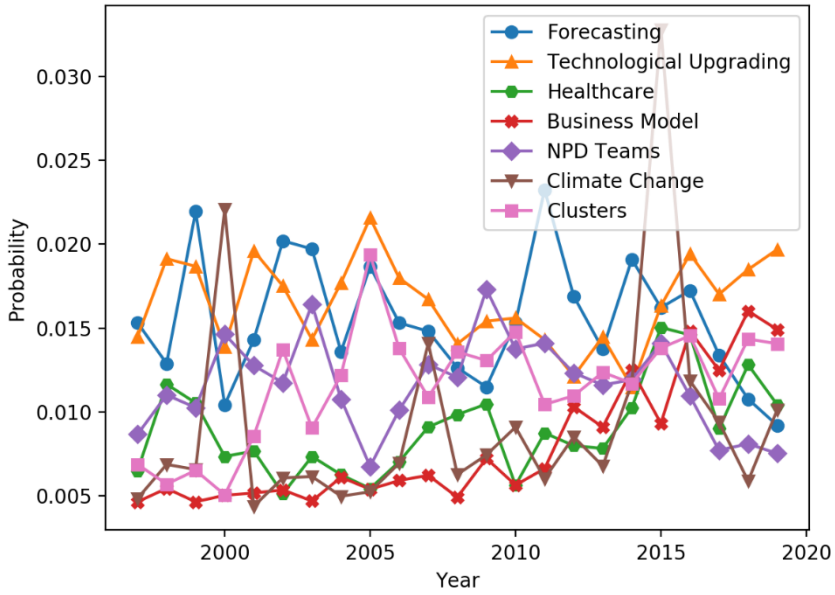


Figure D.1. Trends For Low-Level Topics (0-6)

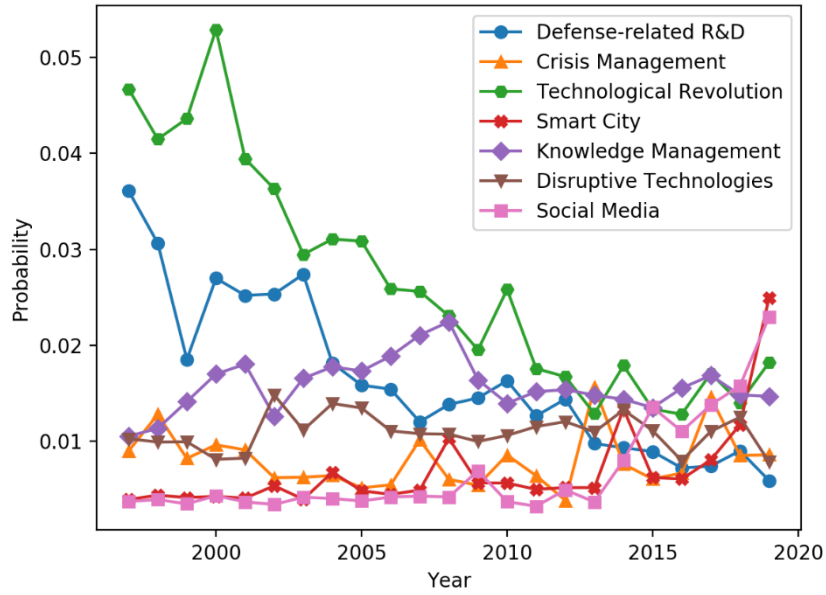


Figure D.2. Trends For Low-Level Topics (7-13)

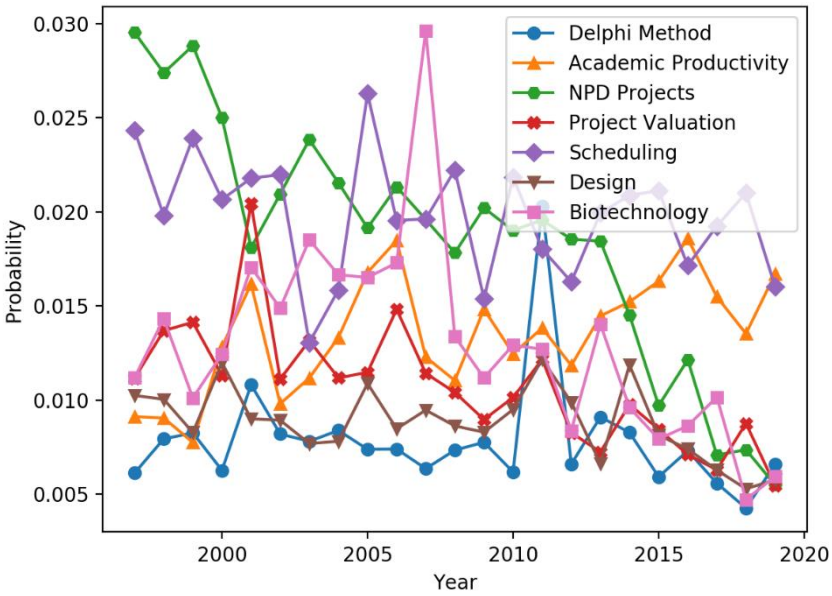


Figure D.3. Trends For Low-Level Topics (14-20)

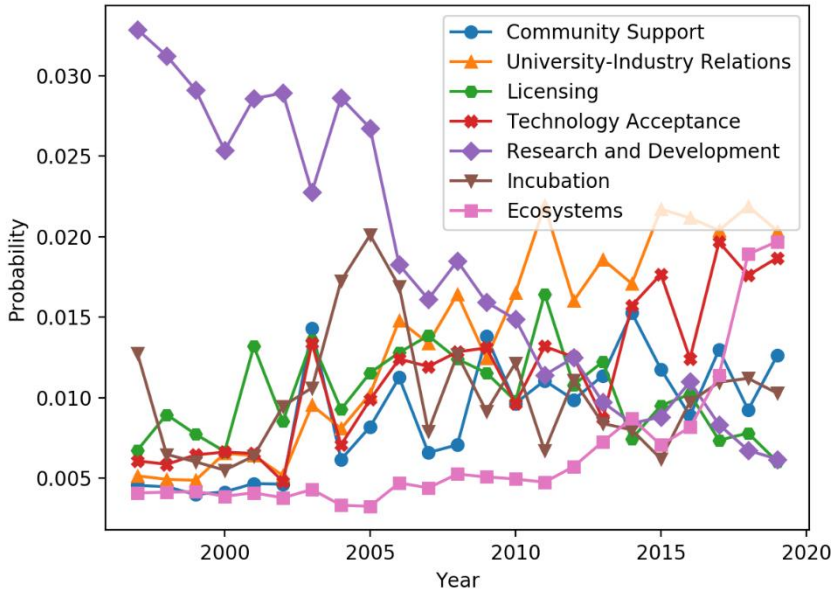


Figure D.4. Trends For Low-Level Topics (21-27)

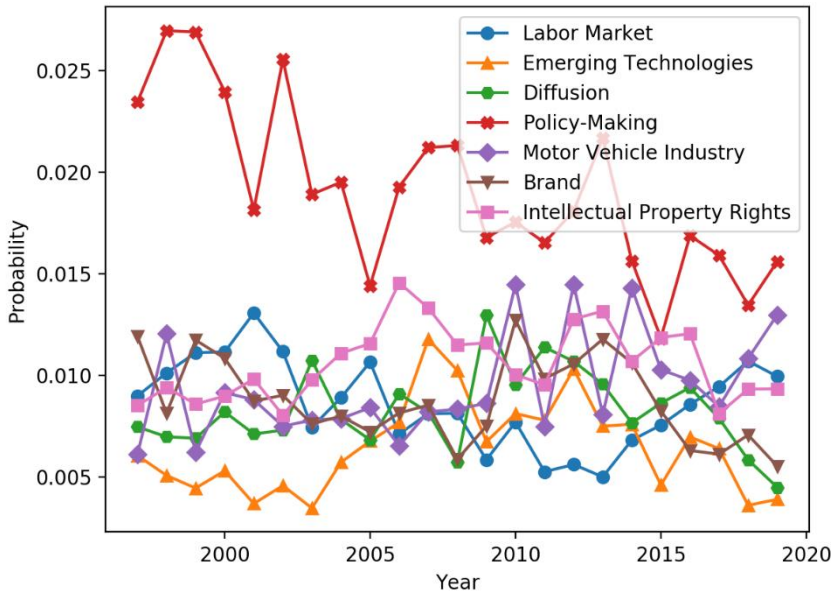


Figure D.5. Trends For Low-Level Topics (28-34)

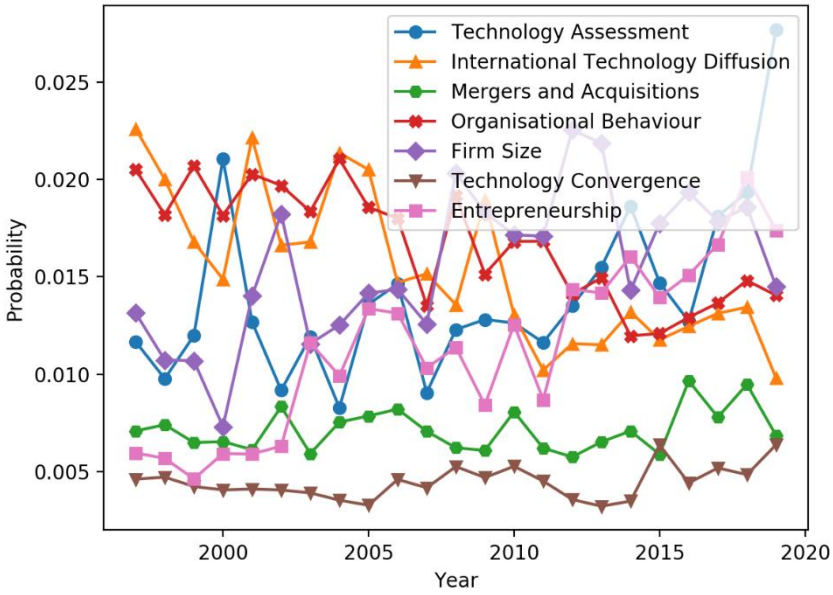


Figure D.6. Trends For Low-Level Topics (35-41)

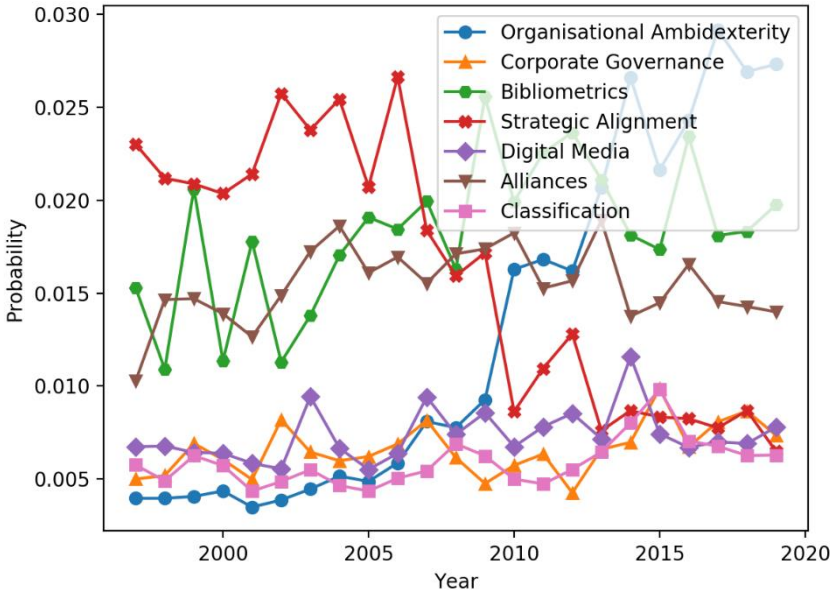


Figure D.7. Trends For Low-Level Topics (42-48)

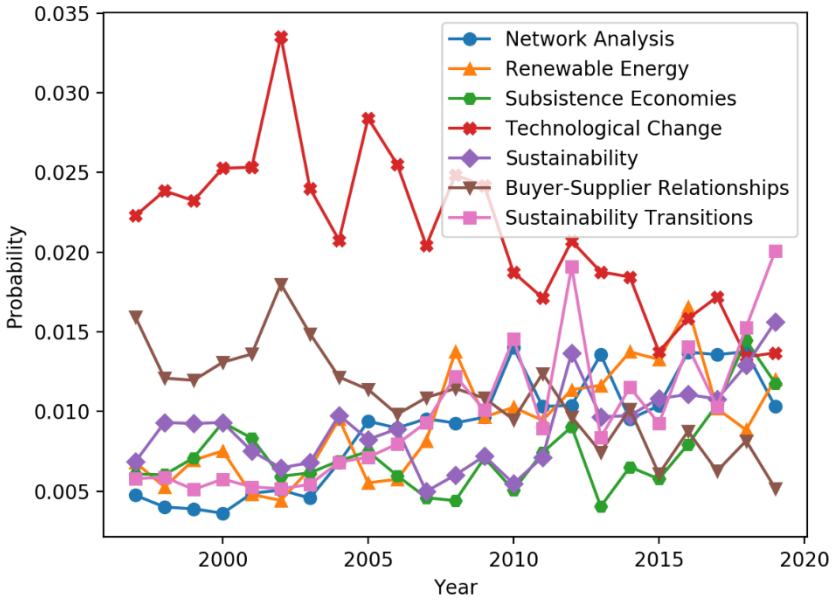


Figure D.8. Trends For Low-Level Topics (49-55)

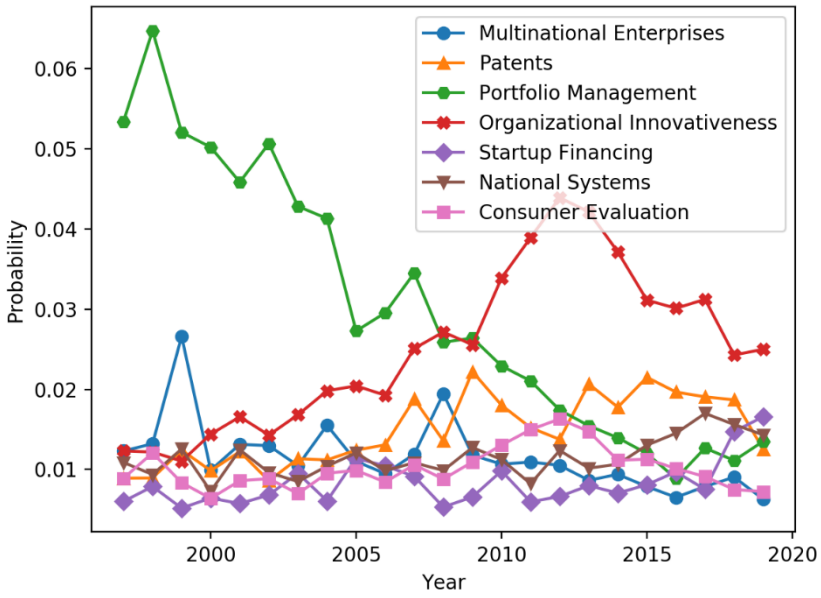


Figure D.9. Trends For Low-Level Topics (56-62)

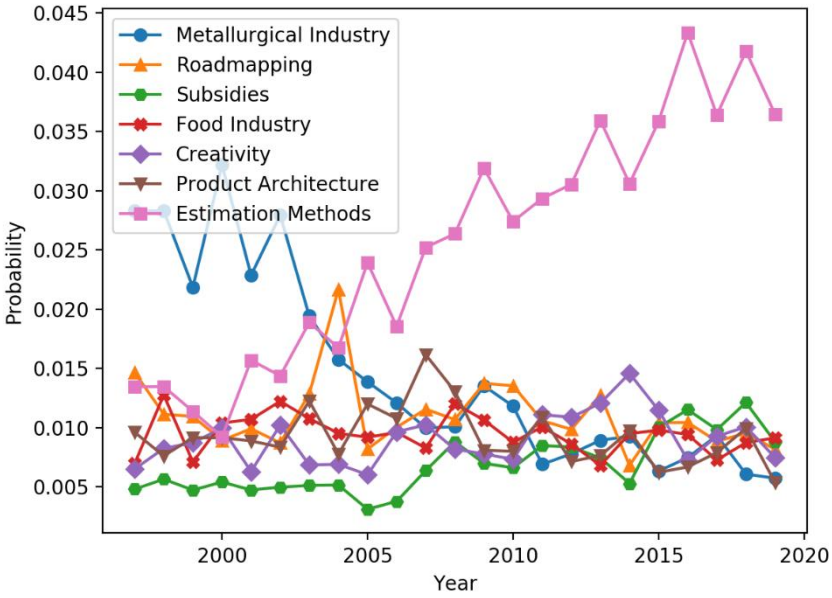


Figure D.10. Trends For Low-Level Topics (63-69)

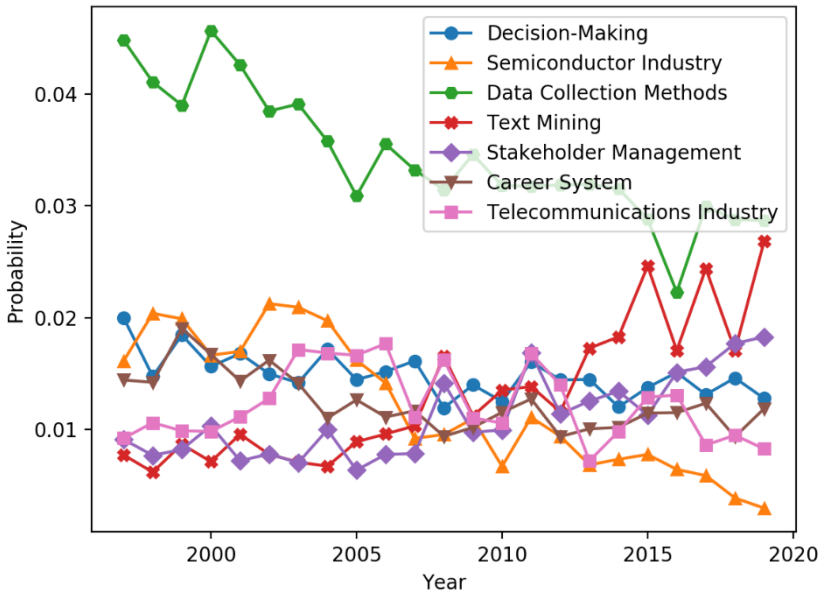


Figure D.11. Trends For Low-Level Topics (70-76)

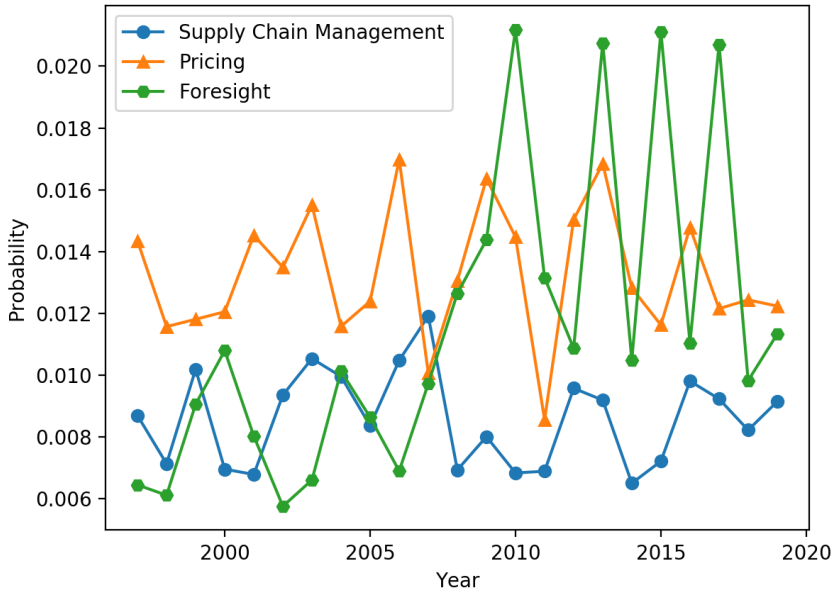


Figure D.12. Trends For Low-Level Topics (77-79)

E. CURRICULUM VITAE

Personal Information

Surname, Name	: Tekin, Yaşar
Nationality	:
Date and Place of Birth	:
Email	:

Education

Degree	Institution	Year of Graduation

Work Experience

Year	Place	Enrollment

Foreign Languages

Publications

F. TURKISH SUMMARY / TÜRKÇE ÖZET

GİRİŞ

Teknoloji Yönetimi (TY), 20. yüzyılın ikinci yarısında ortaya çıkan yeni bir bilim dalıdır. National Research Council (1987) TY'yi "bir organizasyonun stratejik ve operasyonel hedeflerini şekillendirmek ve gerçekleştirmek için teknolojik yetenekleri planlamak, geliştirmek ve uygulamak amacıyla mühendislik, bilim ve yönetim disiplinlerini birbirine bağlayan" ve henüz gelişmekte olan bir alan olarak tanımlamıştır. Gregory (1995), TY için 1) Tanımlama, 2) Seçim, 3) Edinme, 4) Sömürme, 5) Koruma faaliyetlerine sahip bir süreç çerçevesi önermiştir. Çetindamar ve arkadaşları (2009a) TY faaliyetlerini iki kategoride gruplandıran bir çerçeve önermişlerdir: 1) temel faaliyetler (Tanıma, Seçme, Edinme, Sömürme, Koruma, Öğrenme) ve 2) destekleyici faaliyetler (Bilgi Yönetimi, Proje Yönetimi, Yenilik Yönetimi).

Literatürden de anlaşılacağı üzere TY tanımı ve kapsamı halen araştırılmaya devam eden yeni ve disiplinler arası bir alandır. Disiplinler arası bir alandaki en önemli sorun, diğer alanlarla ilişkilerinin de temelini oluşturan kapsamın tanımlanmasıdır. Bu bilgi olmadan yapılacak bir araştırmanın çerçevesini belirlemek zor olacaktır. Ayrıca bu bilgi, editörlerin yönettikleri konferansın/derginin amacını ve kapsamını tam olarak belirlemeleri açısından da önemlidir.

Bu sorunu çözmenin bir yolu, metin madenciliğini kullanmaktır. Bir metin madenciliği yöntemi olan konu modelleme, bir derlemde ele alınan konuları tanımlamak için kullanılan bir tekniktir. Amacı, gözlemlenebilir belge-sözcük dağılımlarından gizli belge-konu ve konu-sözcük dağılımlarının otomatik olarak elde edilmesidir. Konu modellemenin en dikkat çekici kullanımlarından birisi, bilimsel alanlara uygulanmasıdır. Belirli bir bilimsel alanda yayınlanan tüm makalelere uygulandığında, konuların ve eğilimlerin genel bir görünümü elde edilir. Tek bir konferansa veya dergiye uygulandığında ise o konferanstan/dergiden elde edilen

konular ve eğilimler, editörü tarafından konferansı/dergiyi küresel araştırmalarla uyumlu hale getirmek için kullanılabilir. Her iki durumda da araştırmacılar bu bilgiyi gelecekteki çalışmalarını planlamak için kullanabilirler.

Bu çalışmanın amacı, TY alanında çalışılan konuların konu modellemesi yöntemiyle ortaya çıkartılmasıdır. Bu amaçla, TY alanında yayın yapan 13 dergide 1997-2019 yılları arasında yayınlanmış 16.180 araştırma makalesi/özetinden oluşan bir derlem oluşturulmuş ve bu derleme konu modelleme uygulanmıştır.

Gizli konular, en popüler konu modeli olan GDA (Blei ve arkadaşları, 2003) kullanılarak elde edilmiştir. Birçok farklı alanda yaygın olarak kullanılmasına rağmen, GDA'nın uygulanışı ile ilgili halen çözümlenememiş iki sorun bulunmaktadır: model parametrelerinin nasıl eniyileneceği ve konu kararsızlığının nasıl ortadan kaldırılacağı.

GDA'da eniyilenmesi gereken üç parametre bulunmaktadır: konu sayısı (K), belge-konu dağılımı Dirichlet önceli (α) ve konu-sözcük dağılımı Dirichlet önceli (β) (Binkley ve arkadaşları, 2014).

Konu kararsızlığı, aynı derlemden her uygulamada farklı konuların elde edilmesi olarak tanımlanabilir. Mehta ve arkadaşları (2014) bunun nedenini GDA algoritmasının olasılıksal doğası olarak açıklarken, Agrawal ve arkadaşları (2018) sıra etkisi olarak açıklamışlardır. Sıra etkisi, uygulamada girdi olarak kullanılan belgelerinin farklı sıralamalarından farklı konuların elde edilmesi olarak tanımlanmıştır.

Bu çalışmada, öncelikle yukarıda bahsedilen sorunlara çözüm bulabilmek amacıyla deneysel bir araştırma yapılmıştır. Bu kapsamda üç konu araştırılmıştır: 1) sıralı belgelerdeki kararsızlık seviyesi, 2) konu kararsızlığının etkilerini ortadan kaldırma (mümkün değilse, hafifletme) yöntemleri, 3) GDA parametrelerini eniyileme yöntemleri. Bu amaçla, ilki haber makalelerinden, ikincisi ise araştırma makalelerinden oluşan farklı özelliklere sahip iki derlem üzerinde deneyler yapılmıştır. Gerçekleştirilen deneyler sonucunda: 1) araştırma makaleleri gibi zaman sıralı belgelerde bile kararsızlık seviyesinin yüksek olduğu, 2) konu kararsızlığının etkilerini hafifletmek için tekrarlanan konu modellerinin ortalama puanlarının

kullanılabileceği, 3) Skip-gram anlamsal benzerlik puanının GDA parametrelerinin belirlenmesinde kullanılabilecek bir ölçü olduğu tespit edilmiştir.

Daha sonra, bir önceki bölümde önerilen yöntem kullanılarak, TY alanına iki farklı düzeyde konu modelleme uygulanmıştır. TY literatüründe öne çıkan konular, en çok çalışılan sektörler, en çok kullanılan yöntemler ve şaşırtıcı konular belirlenmiştir. Ek olarak yıllık bazda konu eğilimleri tespit edilerek grafik haline getirilmiştir.

KULLANILAN YÖNTEMLER

Gizli Dirichlet Ayrımı (GDA)

GDA, Blei ve arkadaşları (2003) tarafından önerilen üç seviyeli hiyerarşik bir modeldir. Şu varsayımlara dayanmaktadır:

1) Her belge sınırlı sayıda konudan oluşmaktadır,
2) Her konu sınırlı sayıda sözcükten oluşmaktadır,
3) Bir belge aşağıda verilen şekilde oluşturulmaktadır:
    a) Belge için bir konu dağılımı seçilir,
    b) Belgenin her sözcüğü için:
        i) Belgenin konu dağılımından bir konu seçilir,
        ii) Konunun sözcük dağılımından bir sözcük seçilir.

GDA, bu üretim sürecini tersine çevirerek gözlemlenebilir belge-sözcük dağılımlarından gizli belge-konu ve konu-sözcük dağılımlarını ortaya çıkarır.

Bu çalışmada, GDA uygulaması için MALLET (McCallum, 2002) yazılım paketi kullanılmıştır.

Diferansiyel Evrim (DE) Algoritması

DE, sürekli uzay eniyileme problemleri için sezgisel bir yaklaşım olarak Storn ve Price (1997) tarafından önerilmiş olasılıksal bir arama algoritmasıdır. Genetik Algoritmaya (GA) yakın tipik bir Evrimsel Algoritmadır (EA). DE ve GA arasındaki fark, mutasyon

aşamasında gerçekleştirilen işlemlerdeki farklılıktan oluşmaktadır (Feoktistov ve Janaqi, 2004). Bu çalışmada eniyileme algoritması olarak DE tercih edilmiş olmasına rağmen bu amaçla başka herhangi bir çokdoruklu eniyileme algoritması da kullanılabilir.

EA'lar popülasyona dayalı algoritmalardır ve doğadaki organizmalara öykünerek her nesilde değerlerini geliştirirler. Benzer şekilde DE algoritmasının amacı da her nesilde bir parametre vektör nüfusunu eniyilemektir. Parametreler, önceden tanımlanmış asgari ve azami değerleri arasındaki bir aralıkta rastgele atanır. Bir parametre vektörünün bir sonraki nesle aktarılıp aktarılmayacağı, maliyet fonksiyonu tarafından döndürülen değere bağlıdır.

DE algoritması, üçü yinelemeli toplam dört aşamadan oluşmaktadır. Bu aşamalar şunlardır:

a) Başlangıç kuşağının oluşturulması

Problemin her bir değişkeninin alabileceği alt ve üst sınır değerler belirlenir. Başlangıç kuşağı için belirlenen adette parametre vektörünün her bir parametresi için bir parametre değeri üretilir.

b) Mutasyon

Her bir hedef parametre vektörü için, mutant (mutasyon geçirmiş) bir parametre vektörü oluşturulur.

c) Çaprazlama

Mutant parametre vektörleri çaprazlanarak deneme parametre vektörleri oluşturulur.

ç) Seçim

Her bir deneme parametre vektörünün bir sonraki kuşağa aktarılıp aktarılmayacağı, deneme parametre vektörü ile karşılık gelen hedef parametre vektörünün bir maliyet fonksiyonu kullanılarak karşılaştırılması sonucunda belirlenir. Maliyet fonksiyon değeri daha küçük (veya daha büyük) olan parametre vektörü bir sonraki kuşağa aktarılır.

d) Sonlandırma Koşulu

DE algoritması dört farklı şekilde sonlandırılabilir:

- Azami nesil sayısı belirlenerek (Storn ve Price, 1997)
- En iyi ve en kötü maliyet fonksiyonu değerleri arasındaki fark önceden belirlenmiş bir eşiğin altında olduğunda (Ali ve Törn, 2004)
- En iyi maliyet fonksiyonu değeri birkaç ardışık yineleme boyunca değişmediğinde (Das ve Suganthan, 2011)
- En iyi maliyet fonksiyonu değeri önceden belirlenmiş bir eşiğe eşit veya daha büyük olduğunda (Das ve Suganthan, 2011)

Bu çalışmada DE algoritmasını sonlandırmak için azami nesil sayısı belirlenmiştir.

Sözcük vektör gösterimi

GDA parametrelerini eniyilemek için çeşitli yöntemler önerilmiştir. Griffiths ve Steyvers (2004), Teh ve arkadaşları (2006), Steyvers ve Griffiths (2007), Zavitsanos ve arkadaşları (2008), Cao ve arkadaşları (2009), Arun ve arkadaşları (2010) ve Mehta ve arkadaşları (2014) istatistiksel yöntemlerin kullanılmasını önermişlerdir. Agrawal ve arkadaşları (2018) en yüksek olasılıklı sözcüklerin örtüşmesini eniyileme kriteri olarak kabul etmişlerdir. Green ve arkadaşları (2014) sözcük sıralamaları arasındaki benzerlikleri kullanırken, Krasnov ve Sen (2019) konuları oluşturan en yüksek olasılıklı sözcükleri sözcük vektörlerine dönüştürerek kümelerin kalitesini değerlendirmede kullanmışlardır. Yalnızca iki çalışma, Stevens ve arkadaşları (2012) ve Krasnov ve Sen (2019), konu modeli ortalama tutarlılık puanı ile konu sayısı arasındaki bağımlılığı değerlendirmiş ve ilişkinin durağan/monoton olduğunu ve en uygun konu sayısının belirlenmesinde kullanılamayacağını tespit etmişlerdir.

Özetle, önceki çalışmaların hiçbirisinde GDA parametrelerinin eniyilenmesi için parametre vektörlerinin anlamsal ilişki puanlarını enbüyüten bir maliyet fonksiyonu kullanılmadığını görüyoruz. Bu çalışmada seçim kriteri olarak parametre vektörlerinin ortalama benzerlik puanlarını kullanmamızın nedeni en yüksek benzerlik puanına

sahip modelin: 1) ilgisiz sözcük içermeyeceğini, 2) en yüksek yorumlanabilirliğe sahip olacağını ve 3) konuları arası ayrımın en yüksek olacağını düşünmemizdir.

Bu çalışmada sözcük vektör gösterimlerinin kullanılmasının nedeni, Aletras ve Stevenson (2013) ve Nikolenko (2016)'nun çalışmalarında gösterildiği üzere konuların tutarlılığını ölçmedeki başarılarıdır. Sözcük vektör gösterimlerini oluşturmanın birkaç yolu olmakla birlikte en popüler yöntemler Word2Vec (Mikolov ve arkadaşları, 2013) ve GloVE (Pennington ve arkadaşları, 2014)'dur. Naili ve arkadaşları (2017) Word2Vec'in GloVE'dan daha iyi performans gösterdiğini tespit ettikleri için sözcük vektör gösterimleri Word2Vec kullanılarak oluşturulmuştur.

Word2Vec'te iki farklı mimari tanımlanmıştır: CBOW ve Skip-gram. CBOW mevcut sözcüğü bağlama göre tahmin ederken, Skip-gram mevcut sözcüğe göre bağlamı tahmin eder.

Mikolov ve arkadaşları (2013), CBOW ve Skip-gram model mimarilerini aynı boyut ve aynı eğitim verilerini kullanarak karşılaştırdıklarında CBOW modelinin sözdizimsel ilişkide, Skip-gram modelinin ise anlamsal ilişkide daha doğru sonuçlara ulaştığını tespit etmişlerdir. Bu çalışmada, GDA parametrelerini eniyilemedeki etkinliklerini karşılaştırmak için her iki yöntem de kullanılmıştır.

Önerdiğimiz metotta bir konunun (k) benzerlik puanı en yüksek olasılıklı sözcükleri arasındaki benzerlik puanlarının toplamıdır. İki sözcük arasındaki benzerlik puanı sözcüklerin vektör gösterimleri arasındaki mesafe hesaplanarak bulunur.

$$\mathrm{sim}_k = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} benzerlik\left(w_i, w_j\right)$$

Burada, T en yüksek olasılıklı sözcük sayısı, w en yüksek olasılıklı sözcük ve $benzerlik()$ karşılık gelen sözcük vektörlerinin benzerliğini (veya uzaklığını) ölçen bir işlevdir.

Aykırı değerler hesaplamaya dahil edilmek istendiği için model (m) ortalama puanı konuların benzerlik puanlarının aritmetik ortalaması alınarak hesaplanmıştır.

$$\text{sim}_m = mean\{\text{sim}_k, k \in \{1, \dots, K\}\}$$

Burada, K konu sayısıdır.

Maliyet fonksiyonu model puanları arası tutarlılığını artırmak için bu işlem bir parametre vektörünün (pv) çoğaltılan tüm modelleri için tekrarlanmış ve model benzerlik puanlarının aritmetik ortalaması hesaplanmıştır.

$$\text{sim}_{pv} = mean\{\text{sim}_m, m \in \{1, \dots, M\}\}$$

Burada, M çoğaltılan modellerin sayısıdır.

Bu çalışmada, sözcük vektör gösterimleri, Gensim (Řehůřek ve Sojka, 2010) kütüphanesi varsayılan parametre değerleri ile kullanılarak oluşturulmuştur. Sözcük vektör gösterimleri, 08 Aralık 2019 tarihinde indirilen, 4.252.721 farklı sözcük ve 2.418.101.826 toplam sözcükten oluşan Wikipedia makaleleri derlemi üzerinde eğitilmiştir. En yüksek olasılıklı sözcük sayısı ve çoğaltılan model sayısı 10 olarak belirlenmiştir.

GDA PARAMETRE OPTİMİZASYONU İLE İLGİLİ ANALİZLER

Derlemler

Bu çalışma kapsamında farklı özelliklere sahip iki derlem kullanılmıştır. İlki açık erişimli Tübitak Elektrik Mühendisliği ve Bilgisayar Bilimleri Dergisi'nde 1995-2019 yılları arasında yayınlanmış 2.283 makaleden oluşmaktadır. İkincisi, 1987 yılında Reuters haber bülteninde yayınlanmış 21.578 belgeden oluşan Reuters-21578 derlemidir.

Syed ve Spruit (2017), özetlerden ve tam metin makalelerden konu çıkarımını karşılaştırmış ve tam metinlerin gürültü sözcüklerden daha az etkilendiğini tespit etmişlerdir. Bu nedenle, ilk derlemde, yalnızca özetleri bulunan ilk 37 makale dışında, tam metin makaleler kullanılmıştır.

Reuters derleminde, 120'si belgelerin en az 1'inde ve 57'si belgelerin en az 20'sinde geçen toplam 135 ekonomik konu kategorisi bulunmaktadır.

Parametre Ayarları

Model eğitiminde, Mallet'in varsayılan etkisiz sözcük listesine bazı sözcükler eklenmiştir. Mallet max-idf ve min-idf parametreleri sırasıyla 7.0 ve 1.0 olarak ayarlanmış, bu parametre değerleri ile, Reuters derleminde 19'dan küçük veya 7.939'dan büyük ve Tübitak derleminde 2'den küçük veya 840'tan büyük belge sayısına sahip tüm sözcükler elenmiştir. Dört karakterden az ve alfabetik olmayan karakterlere sahip sözcükler de elenmiştir. Sözcük hazinesi büyüklüğü ve derlem boyutu, sırasıyla, Reuters derlemi için 5.139 ve 1.039.200 ve Tübitak derlemi için 27.836 ve 2.865.417 olmuştur.

DE algoritmasında parametre vektörlerinin sayısı 30, amplifikasyon faktörü 0.9, çaprazlama sabiti 0.9, nesil sayısı 50, K, α ve β parametrelerin asgari ve azami değerleri sırasıyla 10-100, 0-1, ve 0-1 olarak belirlenmiştir.

İlave olarak, max-idf ve min-idf parametreleri için farklı değerler kullanılarak ve varlık isimleri belirlenerek metin ön işlemenin eniyileme süreci üzerindeki etkileri araştırılmıştır. Varlık isimlerini tanımlamak için Stanford Varlık İsmi Tanıyıcı (Manning ve arkadaşları, 2014) kullanılmıştır.

Konu Kararsızlığına İlişkin Analiz

Konu kararsızlığının nedeninin sıra etkisi olup olmadığını belirlemek için Tübitak derlemindeki makaleler yayın tarihine göre sıralanıp bu sıra kullanılarak elde edilen konuların kararsızlık düzeyi belirlenmiştir. Bu amaçla, aynı parametre vektörü (K=34, α=50/34, β=0.1) kullanılarak oluşturulmuş iki konu modelinin konularındaki en yüksek olasılıklı 10'ar sözcüğün ne kadar örtüştükleri incelenmiştir. Konu sayısının 34 olarak belirlenmesinin sebebi Tübitak derleminde en uygun konu sayısının 34 olarak bulunmasıdır. Karşılaştırmadan elde edilen sonuçlar Tablo 1'de verilmiştir.

Tablo 1. En Yüksek Olasılıklı 10 Sözcüğün Örtüşme Sayısı

Örtüşme Sayısı	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**
Konu Sayısı	2	6	8	6	3	3	2	4	0	0

Konuların en yüksek olasılıklı sözcükleri arasındaki örtüşme sayısının düşük olması, yüksek düzeyde kararsızlık bulunduğunun göstergesidir. Yayın tarihine göre sıralanan makalelerde kararsızlık düzeyinin yüksek olması GDA'daki konu kararsızlığının nedeninin sıra etkisi olmadığını (veya sadece bu olmadığını) göstermektedir.

Bu, GDA'nın konu kararsızlığının, çıkarım algoritmalarının rastlantısal başlangıçları ile ilgili bir sorun olduğu anlamına gelir. Bu nedenle, konu kararsızlığının konu modelleri üzerindeki etkilerini ortadan kaldırmanın bir yolunun bulunması gerekmektedir. Ancak şu ana kadar bulunabilmiş kesin bir çözüm yoktur, yalnızca etkilerini hafifletmek için önerilen yöntemler bulunmaktadır.

Konu kararsızlığının konu modellerinin anlamsal ilişki puanları üzerindeki etkisini görmek için, Reuters derleminde bir parametre vektörü ile oluşturulmuş 10 modelin Skip-gram benzerlik puanları hesaplanmıştır. Reuters derlemindeki belgelerin en az 20'sinde geçen ekonomik konu kategorilerinin sayısı 57 olduğundan konu sayısı 57 (K=57, α=50/57, β=0.1) olarak belirlenmiştir. İki kez tekrarlanan bu işlem sonucunda elde edilen puanlar Tablo 2 ve Tablo 3'te verilmiştir.

Tablo 2. Çoğaltılan Modellerin Benzerlik Puanları - İlk Çalıştırma

Model	**Aritmetik Ortalama**	**Asgari Değer**	**Azami Değer**
1	19.367801	11.208375	27.617414
2	19.197771	11.682178	30.197244
3	19.191798	10.518077	25.388879
4	18.905101	10.518077	26.164223
5	18.948176	12.487615	27.617414
6	19.183293	11.718470	27.759172
7	19.188607	10.518077	26.533242
8	18.911677	12.184703	26.527101
9	19.390160	13.701636	27.751282
10	19.317189	11.983537	27.785325

Tablo 3. Çoğaltılan Modellerin Benzerlik Puanları - İkinci Çalıştırma

Model	Aritmetik Ortalama	Asgari Değer	Azami Değer
1	19.139778	10.540663	27.756282
2	19.162742	10.542663	27.617414
3	19.367723	12.420408	26.206174
4	19.480483	12.635156	28.092780
5	19.223068	11.298941	27.617414
6	19.273109	10.518077	30.915678
7	19.162590	12.107647	27.617414
8	19.445265	10.518077	29.247615
9	19.156673	12.674428	27.617414
10	19.084820	10.832464	27.617414

Tablo 2 ve Tablo 3'ten, çoğaltılan modellerin en yüksek ve en düşük ortalama puanları arasındaki farkın, bir eniyileme algoritmasının başarısını olumsuz yönde etkileme potansiyeline sahip olduğu görülmektedir. Bu farkı azaltmak için, çoğu araştırmacı tarafından kullanılana benzer bir yöntem kullanılmış ve tekrarlanan konu modellerinin ortalama puanları hesaplanmıştır. Algoritmanın çalışma süresi üzerinde olumsuz bir etkisi olmasına rağmen, bu yöntemin iki avantajı bulunmaktadır: 1) bir parametre vektörünün benzerlik puanları arasındaki tutarlılığı arttırır ve 2) bir konunun en yüksek olasılıklı sözcükleri kararsızlık sebebiyle değişse bile aynı anlam alanına ait sözcüklerin değişmesi benzerlik puanında önemli bir değişikliğe neden olmaz.

Yukarıda verilen örnek için, parametre vektörünün ortalama değeri, ilk çalıştırmada 19.160157 ve ikinci çalıştırmada 19.249625 olmuştur. Bu iki çalışma arasındaki fark 0.089468'dır. Çapraz doğrulama için algoritma iki kez daha çalıştırılmış, 19.203078 ve 19.217540 puanları elde edilmiştir. Bu sonuçlar, çoğaltılan modellerin sayısı arttıkça ortalama puanın daha kararlı hale geldiğini ve kullandığımız yöntemin bir parametre vektörünün benzerlik puanları arasındaki farkları kabul edilebilir seviyelere indirebildiğini göstermektedir.

Sıra etkisinin başka bir analizi olarak, yukarıda verilen parametre vektörlerinin 1) varsayılan sıralı, 2) düz sıralı (Reuters alfabetik, Tübitak yayın tarihi) ve 3) ters sıralı olarak ortalama Skip-gram benzerlik puanları hesaplanmıştır. Her bir sıralama için Reuters ve Tübitak derlemlerinde tekrarlanan dört çalıştırmadan (toplam 40 çoğaltılan model) elde edilen puanlar sırasıyla Tablo 4 ve Tablo 5'de verilmiştir.

Tablo 4. Reuters Derleminde Benzerlik Puanları

Sıralama	**Tekrar 1**	**Tekrar 2**	**Tekrar 3**	**Tekrar 4**
Varsayılan	19.160157	19.203078	19.217540	19.249625
Düz	19.149234	19.190832	19.204757	19.292303
Ters	19.201709	19.211751	19.232542	19.330780

Tablo 5. Tübitak Derleminde Benzerlik Puanları

Sıralama	**Tekrar 1**	**Tekrar 2**	**Tekrar 3**	**Tekrar 4**
Varsayılan	19.168598	19.213635	19.246590	19.336062
Düz	19.207929	19.227536	19.294607	19.306033
Ters	19.257453	19.280972	19.301260	19.335961

Tablo 4 ve Tablo 5'den görülebileceği üzere puanlar birbirine oldukça yakındır. Bu durum, derlemi oluşturan belgelerin sırasının konu modellerinin anlamsal ilişki puanları üzerinde bir etkisinin olmadığını göstermektedir.

Parametre Eniyileme ile Konu Çıkarma Analizi

max-idf (7.0) ve min-idf (1.0) parametrelerinin sabit tutulduğu, maliyet fonksiyonunun değiştirildiği durumlarda elde edilen en uygun parametre değerleri Tablo 6'da verilmiştir.

Tablo 6. Farklı Maliyet Fonksiyonları Kullanılarak Elde Edilen Puanlar

Derlem	Fonksiyon	K	α	β	Puan
Reuters	CBOW	41	0.365681	0.564030	12.291904
	Skip-gram	52	0.946186	0.021711	19.409072
Tübitak	CBOW	11	0.254452	0.117408	16.459047
	Skip-gram	34	0.118924	0.005694	19.547783

Daha önce bahsedildiği üzere Reuters derleminde belgelerin en az 1'inde geçen 120, belgelerin en az 20'sinde geçen 57 ekonomik konu kategorisi bulunmaktadır. Bu çalışmada, belge sıklığı 19'dan az veya 7.939'dan büyük olan sözcükler çıkarıldığı için 57 civarında bulunması beklenen en uygun konu sayısının 52 olarak tespit edildiği görülmektedir.

Tübitak derlemi en uygun konu sayısı Skip-gram modelinde 34 olarak bulunmuştur. Tübitak derleminin Reuters derlemine göre daha dar bir kapsama sahip olması nedeniyle bu sonuç kabul edilebilir olarak değerlendirilmiştir.

Reuters ve Tübitak derlemlerinden en uygun parametre değerleri kullanılarak elde edilen modellerde en yüksek ve en düşük puana sahip 10'ar konunun en yüksek olasılıklı sözcükleri Tablo 7 ve Tablo 8'de verilmiştir.

Tablo 7. Reuters Derleminde En Yüksek ve En Düşük Puanlı Konular

No	Top Terms	Score	Rank
50	securities financial companies foreign investment market firms investors capital markets	28.092780	1
27	president chairman chief executive board vice officer director named company	27.617414	2
17	tonnes sugar production export exports total tonne wheat stocks imports	24.276302	3
0	joint services company venture corp companies telephone international service communications	22.931765	4

Tablo 7 *(devam ediyor)*

13	systems products computer system technology corp data line equipment business	22.467803	5
48	federal system reserve funds reserves statement market expected period agreements	16.329517	48
39	loss profit note includes gain year operations share sales extraordinary	15.904988	49
51	issue bond lead bonds manager priced coupon date payment selling	15.276265	50
1	dollar currency west exchange german central germany mark rates paris	15.275254	51
3	rose fell rise figures increase department adjusted revised orders index	14.725310	52

Tablo 8. Tübitak Derleminde En Yüksek ve En Düşük Puanlı Konular

No	**Top Terms**	**Score**	**Rank**
0	measurement sensor measurements temperature sample samples device sensors surface tests	25.600047	1
19	converter switching inverter harmonic grid circuit switches currents capacitor voltages	23.687377	2
1	graph node tree nodes path query edges networks edge graphs	23.544503	3
24	cell cells flow pressure water blood pump fluid concentration velocity	22.588380	4
32	nonlinear stability delay chaotic discrete stable matrices dynamics lyapunov theorem	22.258865	5
20	controller response controllers mode loop sliding gain observer damping disturbance	16.466181	30

Tablo 8 *(devam ediyor)*

14	robot position robots vehicle motion force tracking velocity joint road	16.397793	31
13	fuzzy rule agent rules membership agents logic learning decision inference	16.293034	32
7	optimization search population fitness particle solutions swarm iteration objective local	15.780419	33
11	flow loss active generation transmission constraints reactive optimization units buses	15.676127	34

Reuters derleminden elde edilen konuları etiketlemek için, konular verilen ekonomik konu kategorileriyle eşleştirilmiştir. Eşleştirmede nesnelliği sağlayabilmek amacıyla konuların en yüksek olasılıklı sözcükleri ile ekonomik konu kategorilerinin isimleri arasında örtüşen sözcükler bulunmuştur. Sonuç olarak 52 konudan 27 tanesi 37 kategori kullanılarak etiketlenmiştir. Konuların en yüksek olasılıklı sözcükleri ve bunlara karşılık gelen ekonomik konu kategorileri Tablo 9'da verilmiştir. Parantez içinde verilen sayılar, verilen ekonomik konu kategorilerinin yer aldığı belge sayısını göstermektedir.

Tablo 9. Konu Etiketleri Olarak Ekonomik Konu Kategorileri

No	**En Yüksek Olasılıklı Sözcükler**	**Kategori**
1	dollar currency west exchange german central germany mark rates paris	D-Mark (0)
3	rose fell rise figures increase department adjusted revised orders index	Industrial Production Index (65)
7	union workers strike spokesman airlines work national general employees aircraft	Unemployment (76)
8	trade exports south imports united states foreign taiwan products import	Trade (552)

Tablo 9 *(devam ediyor)*

11	production energy crude opec petroleum barrels prices barrel natural output	Crude Oil (634), Natural Gas (130)
15	pacific areas steel area southern weather north early people normal	iron-steel (67)
16	debt banks brazil interest foreign payments billion country bank creditors	Balance of Payments (116), Brazilian Cruzado (1)
17	tonnes sugar production export exports total tonne wheat stocks imports	sugar (184), wheat (306)
18	quarter year earnings share company expects income reported profits results	Personal Income (18), Earnings and Earnings Forecasts (3987)
19	exchange trading stock futures options market contracts york contract chicago	Money/Foreign Exchange (801)
20	canada canadian statement british announced made dome continental added full	Canadian Dollar (3)
21	credit interest financing basis facility paper program offered international years	Instalment Debt/Consumer Credit (7)
25	japan japanese officials trade tokyo ministry official open united states	Japanese Yen (69)
28	department agriculture program wheat corn farm usda farmers grain land	corn (254), grain (628), wheat (306)
30	analysts market industry analyst major kong hong time stocks street	Hong Kong Dollar (1)
31	gulf iran saudi iranian united military states shipping agency iraq	Shipping (305)
33	market dealers traders dollar close buying early closed trading selling	U.S. Dollar (217)

Tablo 9 *(devam ediyor)*

35	offer share merger company tender takeover corp shareholders cash board	Mergers/Acquisitions (2448)
36	coffee international meeting agreement producers export cocoa stock council market	coffee (145), cocoa (76)
38	growth economy year economic rise inflation report domestic demand expected	Gross National/Domestic Product (163)
41	gold resources australia production mining mine tons australian copper silver	gold (135), copper (78), silver (37), Australian Dollar (4)
42	european community commission french europe ministers france sources proposals west	French Franc (0)
43	world countries economic international development meeting baker nations developing major	Leading Economic Indicators (17)
45	sales corp cars general drug stores motors chrysler food company	Retail Sales (27)
47	bank rate rates money interest market central bills treasury england	Interest Rates (513)
48	federal system reserve funds reserves statement market expected period agreements	Reserves (84)
51	issue bond lead bonds manager priced coupon date payment selling	Wholesale Price Index (32)

Etiketlemeden elde edilen bazı dikkat çekici sonuçlar örnekleriyle birlikte aşağıda listelenmiştir:

1. Bazı konular elle yapılan kontrolde ekonomik konular olarak belirlenmesine rağmen etiketlenememiştir (Tablo 7'de Konu 50)
2. Bazı konular birden fazla ekonomik konu kategorisi ile etiketlenmiştir (Konular 11, 16, 17, 18, 28, 36, 41)

3. Bazı konular eşleşen ekonomik konular bulunamadığı için etiketlenememiştir (Tablo 7'de Konular 0, 13, 27)
4. Bazı konular, 20'den az belgede geçen ekonomik konu kategorileri ile etiketlenmiştir (Konular 1, 20, 21, 30, 42, 43)

Bu sonuçlar, derlemin konu modelleme ile ilgili iki sorunu bulunduğunu göstermektedir: 1) tüm ekonomik konu kategorileri aynı seviyede değildir ve bu, verilen tüm ekonomik konu kategorilerinin tek bir konu modeli ile çıkarılmasına engel olmaktadır ve 2) bazı belgelerde ekonomik konu kategorisi bulunmamakta, bu ise ekonomik konu kategorilerinde verilmeyen veya 20'den az belgede yer alan konuların çıkarılmasına neden olmaktadır.

Bu sorunlar birebir bir eşleşmenin mümkün olmadığını göstermektedir. Ancak aşağıda da görüleceği üzere parametre değerleri değiştirilerek farklı seviyelerdeki konular elde edilebilmektedir. Bu yöntemle tek bir modelle elde edilemeyen ekonomik konu kategorilerinin elde edilebilmesi mümkün hale gelmektedir.

Bu çalışmada elde ettiğimiz bir diğer sonuç, α ve β parametrelerinin en uygun değerlerinin farklı derlemlerde oldukça farklı değerler almasıdır. Bu, Griffiths ve Steyvers'ın (2004) çalışmasında uygulanan yöntemin aksine, bu parametrelerin her derlem için eniyilenmesi gerektiğini göstermektedir.

Analizimizde her iki derlemde de Skip-gram modelinin CBOW modeline göre daha yüksek benzerlik puanlarına ulaştığı görülmektedir. Bu sonuç Mikolov ve arkadaşlarının (2013) çalışması ile uyumludur. Bu çalışmada, Skip-gram modelinin anlamsal ilişkide daha başarılı sonuçlar elde ettiği görülmüştür. Ek olarak ve beklendiği üzere, GDA'nın varsayılan ayarları kullanılarak çıkarılan modellerinden (Tablo 4 ve Tablo 5) daha yüksek benzerlik puanları elde edilmiştir.

İstatistiksel yöntemlerin konuların yorumlanabilirliği ile uyuşmadığı belirlendiğinden, bu çalışmadaki sonuçlar yalnızca Agrawal ve arkadaşlarının (2018) çalışmasında verilen yöntemle elde edilen sonuçlarla karşılaştırılmıştır. LDADE kullanılarak yapılan eniyileme sonucunda her iki derlem için en uygun konu sayısı 10 olarak bulunmuştur.

Yukarıda verilen sonuçlar ışığında Skip-gram modeli ile eniyilemenin tutarlılık puanlarının yüksek olması nedeniyle daha başarılı olduğu söylenebilir. Bu nedenle, çalışmanın bundan sonraki bölümünde konu modellerinin tutarlılığını ölçmek için Skip-gram benzerlik puanları kullanılmıştır.

İlk olarak, metin ön işlemenin eniyileme süreci üzerindeki etkileri araştırılmıştır. Bu amaçla GDA parametreleri derlemdeki varlık isimleri tespit edilerek eniyilenmiştir. Derlemlerde "Location", "Person", "Organisation" ve "Misc" türlerinde varlık isimleri tespit edilmiş, dörtten fazla sözcük içerenler filtrelenmiş ve kalanlar birleştirilerek tek sözcük haline getirilmiştir. Bu yöntem kullanılarak elde edilen puanlar Tablo 10'da verilmiştir. Tablodan, Tübitak derlemindeki artış çok belirgin olmasa da yöntemin her iki derlem için de en uygun konu sayısını arttırdığı görülmektedir.

Tablo 10. Birleştirilmiş Varlık İsimlerinin Etkileri

Derlem	**K**	**α**	**β**	**Puan**
Reuters	59	0.897372	0.010262	18.805330
Tübitak	35	0.229822	0.021636	18.615432

İkinci olarak, GDA parametreleri max-idf ve min-idf parametrelerinde farklı değerler kullanılarak eniyilenmiştir. Reuters ve Tübitak derlemlerinden elde edilen puanlar sırasıyla Tablo 11 ve Tablo 12'de verilmiştir. Tablolardan, parametre değerlerindeki değişikliklerin en uygun konu sayısında değişikliklere yol açtığı görülmektedir. Bu husus, parametre değerlerinin değiştirilmesinin farklı seviyelerdeki konuların elde edilmesini sağladığının bir göstergesidir.

Tablo 11. Farklı Parametre Değerlerinin Reuters Derlemi Üzerindeki Etkileri

max-idf	**min-idf**	**K**	**α**	**β**	**Puan**
7.0	4.0	38	0.082931	0.077762	17.727794
7.0	5.5	52	0.043958	0.195187	16.329127
6.5	1.0	49	0.864662	0.017281	19.315531
6.0	1.0	37	0.919898	0.004740	19.320366

Tablo 12. Farklı Parametre Değerlerinin Tübitak Derlemi Üzerine Etkileri

max-idf	min-idf	K	α	β	Puan
7.0	4.0	10	0.675458	0.130764	17.001204
7.0	5.5	22	0.551891	0.020239	11.683725
6.5	1.0	25	0.021310	0.118982	19.695736
6.0	1.0	27	0.176330	0.125691	19.766908

Bu duruma örnek olarak Reuters derleminde max-idf parametresi 7.0 ve min-idf parametresi 3.4 olarak girilmiş ve toplam sözcük sayısı 4.946'ya düşürülmüştür. K, α ve β parametreleri için en uygun değerler sırasıyla 18, 0.097496 ve 0.317216 olarak bulunmuştur. En uygun değerler kullanılarak çıkarılan konulardan bir tanesinin en yüksek olasılıklı sözcükleri *wheat, sugar, grain, agriculture, corn, crop, usda, soybean, farmers, soviet* olmuştur. Bu konu, Tablo 9'da verilen 17 ve 28 numaralı konuların birleşimidir ve etiket olarak kullanılan ekonomik konu kategorileri listesine *soybean* kategorisini eklemiştir.

Ancak yukarıdaki örnek ve Tablo 11 ve Tablo 12'de verilen değerlerden görüleceği üzere parametre değerlerindeki değişikliklerin konu sayısı üzerinde sabit bir etkisi yoktur. Parametre değerlerindeki değişikliklerin eniyileme sonuçları üzerindeki etkilerinin derlemlerdeki sözcüklerin dağılımına bağlı olduğu düşünülmektedir ancak bu analizin gelecekteki çalışmalarda incelenmesine karar verilmiştir.

KONU MODELLEMENİN TM ALANINA UYGULANMASI

Derlem

TM ile ilgili daha önce yapılan çalışmalar incelendiğinde, analizlerde genellikle belirli dergilerin kullanıldığı görülmektedir. Çalışmaların çoğunda bu alandaki ilk 10 dergi kullanılmıştır. Bu çalışmalardan birinde (Lee, 2015), 10 TM dergisine dergi atıf ağı analizi uygulanmış ve listeye dahil edilmeyi hak eden üç ek dergi bulunduğu belirlenmiştir. Bu nedenle, çalışmamızda Lee (2015) tarafından önerilen 13 dergi kullanılmıştır.

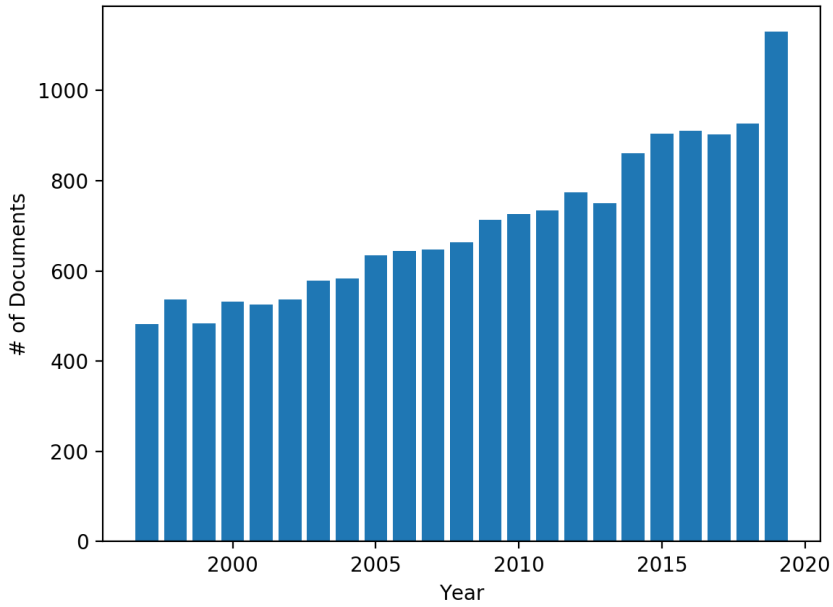
Antons ve arkadaşları (2016) ve Syed ve Spruit (2017) tarafından yapılan tam metin makalelerin daha başarılı olduklarına dair tespit nedeniyle tam metin makaleler kullanılmıştır. Ancak, IJTM'nin tam metin makaleleri erişime açık olmadığı için bu dergiden sadece özetler kullanılabilmiştir. Sonuç olarak, 12 dergiden 14.471 tam metin araştırma makalesi ve bir dergiden 1.709 özet olmak üzere TM alanında yayınlanmış toplam 16.180 araştırma makalesi/özet kullanılmıştır. Bu çalışmada kullanılan dergi adları, makale sayıları ve türleri Tablo 13'te verilmiştir.

Tablo 13. Makale Sayıları ve Türleri

No	Adı	Makele Sayısı	Türü
1	IEEE Transactions on Engineering Management	1,048	Tam metin
2	Innovation: Organization & Management	493	Tam metin
3	International Journal of Technology Management	1,709	Özet
4	Journal of Engineering and Technology Management	435	Tam metin
5	Journal of Product Innovation Management	1,024	Tam metin
6	Journal of Technology Transfer	936	Tam metin
7	R&D Management	838	Tam metin
8	Research Evaluation	649	Tam metin
9	Research Policy	2,494	Tam metin
10	Research Technology Management	925	Tam metin
11	Technological Forecasting and Social Change	3,026	Tam metin
12	Technology Analysis & Strategic Management	1,186	Tam metin
13	Technovation	1,417	Tam metin

Şekil 1, seçilen dergilerde yayınlanan yıllık makale sayılarını göstermektedir. Şekilde görüldüğü üzere, TM dergilerinde yayınlanan araştırma makalelerinin toplam sayısında istikrarlı bir artış söz konusudur ve bu TM'ye giderek artan bir ilginin varlığının göstergesidir.

Bu çalışmada, grafikleri oluşturmak için Matplotlib kütüphanesi (Hunter, 2007) kullanılmıştır.



Şekil 1. Yıllık Belge Sayısı

Derlemin ön işlemesinde, Mallet'in varsayılan etkisiz sözcüklerine ilaveler yapılmış, dört karakterden kısa veya alfabetik olmayan karakterler içeren sözcükler elenmiştir. Dergi, yazar, ülke, şehir ve şirket adları çıkarılmıştır. Buna rağmen en yüksek olasılıklı sözcükler içinde hala var olanlar kodlanmış sözcüklerle değiştirilmiştir. Varlık isimleri, Stanford Varlık İsmi Tanıyıcı (Manning ve arkadaşları, 2014) kullanılarak tanımlanmış ve dört sözcüğe kadar olanlar modellemeye dahil edilmiştir.

Sonuçlar

Önceki bölümde yapılan çalışmalar sonucunda ortaya konulduğu üzere GDA parametreleri, çoğaltılan konu modellerinin ortalama anlamsal benzerlik puanları kullanılarak eniyilenebilmektedir. Bu yöntemin bir avantajı, farklı seviyelerde en uygun konu sayısını belirleme yeteneğidir. Tekniğin bu özelliği kullanılarak konular iki farklı düzeyde çıkarılmıştır. Her bir seviye için kullanılan Mallet parametreleri ve elde edilen en uygun GDA parametre değerleri Tablo 14'te verilmiştir.

Tablo 14. En uygun GDA Parametre Değerleri

	Ön İşleme Parametreleri			**GDA Parametreleri**		
Seviye	**As Bel**	**Az Bel**	**Sözcük**	**K**	**α**	**β**
Üst	14	5,953	64,549	13.0	0.1420528	0.1867320
Alt	14	3,611	63,887	80.0	0.5725732	0.3653631

Tabloda, *As Bel* ve *Az Bel* sütunları, derlemdeki sık ve nadir sözcükleri elemek için kullanılan asgari ve azami belge sayılarını göstermektedir. Tablodaki veriler üst düzey konularda 14'ten az ve 5.953'ten fazla belgede, alt düzeyli konularda ise 14'ten az ve 3.611'den fazla belgede geçen sözcüklerin elendiğini ifade etmektedir. *Sözcük* sütunu, sık ve nadir sözcüklerin elenmesinden sonra derlemde kalan sözcük sayısını göstermektedir. Tablodan görülebileceği üzere eniyileme işlemi sonucunda konu sayısı üst düzey konular için 13, alt düzey konular için 80 olarak bulunmuştur. Konuların en yüksek olasılıklı sözcükleri, etiketleri, en yüksek olasılıklarla bulundukları belgeler ve eğilimleri, sırasıyla, üst düzey konular için Tablo 15, Ek A ve Ek B'de, alt düzey konular için Tablo 16, Ek C ve Ek D'de verilmiştir.

Tablo 15'de, *Konu* sütunu, Mallet tarafından otomatik olarak atanan konu tanımlayıcılarını, *En Yüksek Olasılıklı Sözcükler* sütunu, her konunun en yüksek olasılığa sahip 10 sözcüğünü ve *Etiket* sütunu yazarlarca belirlenen konu etiketlerini göstermektedir. Bir örnek olmak üzere, ilk satırda verilen konuya ait konu numarası 0, en yüksek olasılıklı sözcükler *patent patents patenting inventors licensing invention inventions pharmaceutical biotechnology intellectual* ve etiketi tüm en yüksek olasılıklı sözcüklerin ilgili olduğu konu olduğu için *Patents (IPRs)*'dir.

Tablo 15. Üst Düzey Konular En Yüksek Olasılıklı Sözcükler ve Etiketler

Konu	**En Yüksek Olasılıklı Sözcükler**	**Etiket**
0	patent patents patenting inventors licensing invention inventions pharmaceutical biotechnology intellectual	Patents (IPRs)

Tablo 15 *(devam ediyor)*

1	team new_product items teams orientation innovativeness respondents creativity perceived constructs	NPD Teams
2	users user internet adoption online consumers consumer mobile platform community	ICT
3	supplier suppliers customer team product_development manager phase tools improvement equipment	Product Development
4	universities technology_transfer entrepreneurship entrepreneurial entrepreneurs funding commercialization ventures venture incubator	Academic Entrepreneurship
5	energy environmental sustainability water electricity emissions fuel green sustainable scenarios	Sustainability
6	foresight scenario experts scenarios nanotechnology expert cluster matrix forecasting trends	Foresight & Forecasting
7	price diffusion optimal parameters probability simulation forecasting profit parameter option	Optimization
8	regional foreign cluster regions region clusters domestic enterprises economies smes	Regional Policies
9	intensity dummy hypothesis coefficient estimation observations estimates spillovers estimated probability	Policy Evaluation
10	funding scientists universities publications papers publication journals articles students faculty	Scholarly Publications
11	actors political stakeholders governance organisations community society organisation health participants	Governance
12	partners alliances absorptive alliance cooperation partner exploration exploitation assets collaborative	Alliances

Üst düzey konularda dikkate değer veya açıklama gerektiren hususlar şunlardır:

- Altı konu, her biri iki konudan oluşan, anlamsal olarak birbiriyle ilişkili üç kümede gruplandırılabilir: Ürün Geliştirme (Yeni Ürün Geliştirme (YÜG)

Takımları (1) ve Ürün Geliştirme (3)), Üniversitelerin Rolü (Akademik Girişimcilik (4) ve Bilimsel Yayınlar (10)) ve Politika Oluşturma (Politika Değerlendirme (9) ve Yönetişim (11)). Bu kümelerin, TM literatürünün en fazla çalışılan üç konusu olduğu düşünülebilir.

- Beklentiler doğrultusunda Patent veya Fikri Mülkiyet Hakları (IPR) (0), TM'nin temel konularından biri olarak belirlenmiştir. Hem üst hem de alt düzey (57) konularda yatay bir eğilim göstermektedir ancak patentlerde metin madenciliği tekniklerinin kullanımının artmasıyla birlikte eğiliminin artabileceği değerlendirilmektedir.
- Yeni Ürün Geliştirme (YÜG) Takımları (1) yaratıcılık, yenilikçilik, performans, sezgisel beceriler, liderlik ve dürüstlük gibi konuları ele almaktadır. Yönetişim (11) konusu ile benzer şekilde tüm dönem boyunca diğer konulara nazaran daha yüksek bir eğilime sahip olmuştur. YÜG ekiplerinin başarısını artırmak için yapılacak çalışmalar nedeniyle gelecek dönemlerde de yüksek bir eğilime sahip olmaya devam edeceği değerlendirilmektedir.
- Bilgi ve İletişim Teknolojileri (BİT) (2) ile ilgili bir konunun üst düzey konular arasında tespit edilmesi, dünya ekonomisinin önemli bir bileşeni olmaya devam eden BİT sektöründeki büyümeye paraleldir. Konu artan bir eğilim göstermektedir ve bu eğilimin gelecekte de BİT sektöründeki artışa paralel olarak devam edeceği değerlendirilmektedir.
- Ürün Geliştirme (3) YÜG süreci ile ilgilidir. Ele alınan konulardan bazıları Modülerlik, Aşama Kapısı, Çevik Aşama Kapısı ve Yalın İlkeler'dir. 2000'li yıllarda çokça çalışılan bir konu olmasına rağmen zamanla bilinirliğini yitirmiştir.
- Daha önceki bazı çalışmalarda Girişimcilik konusundan bahsedilmiş olsa da hiçbirinde Akademik Girişimcilikten (4) bahsedilmemiştir. Bu çalışmada, Akademik Girişimcilik (4) ve Bilimsel Yayınlar (10)'dan oluşan Üniversitelerin Rolü kümesi, TM'nin en fazla çalışılan üç konusundan birisi olarak belirlenmiştir. Bu sonuçtan, araştırmacıların bir şirketle birlikte çalışmaktan daha ziyade kendi şirketlerini kurmaya ilgi duydukları sonucu çıkarılabilir. Bu nedenle hem araştırmacılar hem de politika yapıcılar

tarafından bu konunun dikkatle ele alınması ve hem üniversitelere hem de üniversite-sanayi ilişkilerine etkilerinin detaylı bir şekilde araştırılması gerektiği düşünülmektedir.

- Sürdürülebilirliğin (5) TM'nin en çok çalışılan konularından birisi olarak belirlenmesi, TM alanında bu konuya verilen önemin bir göstergesidir. Hem üst hem de alt düzey konularda (53 ve 55) hafif artan bir eğilim göstermiştir ve bu eğilimin gelecekte de devam edeceği değerlendirilmektedir.
- Öngörü ve Kestirim (6), TM alanında en çok kullanılan yöntemler olarak belirlenmiştir. Bu sonuç, TM alanının en önemli amaçlarından birisinin yeni teknolojilerin ve eğilimlerin mümkün olduğunca erken tanımlanması olduğu gerçeğiyle uyumludur. Geçmişte sahip olduğu yatay eğilim, bu yöntemlerin kullanımının veya bu yöntemler ile ilgili çalışmaların gelecekte de aynı şekilde devam edeceğinin bir göstergesidir.
- Eniyileme (7) ve Bölgesel Politikalar (8) TM derleminden çıkarılan iki üst düzey konudur. Eniyileme (7), yayılma modelleri, ikame modelleri, büyüme modelleri, sermaye yatırım analizi, üretim kontrolü ve çizelgeleme ile ilgiliyken, Bölgesel Politikalar (8), bölgesel inovasyon sistemleri, teknoloji transferi, yabancı inovasyon, yabancı Ar-Ge ve bölgesel özellikler ile ilgilidir. Her ikisi de incelenen dönemde azalan eğilimler göstermişlerdir. Eğimlerinden, konuların eğilimlerinin gelecekte de azalmaya devam edecekleri sonucuna ulaşılabileceği değerlendirilmektedir.
- Politika Değerlendirme (9) TM alanında geçmişte yapılmış çalışmaların hiçbirinde tespit edilmemiştir. Bu çalışmada Yönetişim (11) başlıklı konu ile birlikte TM'nin en çok çalışılan konularından birisi olarak belirlenmiştir. Ele alınan konulardan bazıları: a) sübvansiyonlar, vergi teşvikleri ve düşük faizli krediler gibi kamu Ar-Ge ve yenilik politikaları, b) kamu ve özel harcamalar arasında dışlama etkisi ve c) Ar-Ge, yenilik ve üretkenlik arasındaki bağlantıdır. Politika Değerlendirme (9) söz konusu dönemde artan bir eğilim göstermiş ve Yönetişim'e (11) yakın bir seviyeye ulaşmıştır. Bu durum, araştırmacıların, politikaların etkililiğini ve verimliliğini değerlendirme konusundaki artan ilgisini göstermektedir. Teknoloji politikaları ve bunların

değerlendirilmesinin, TM'nin en önemli konularından olmaya devam edecekleri düşünülmektedir.

- İttifaklar (12), yeni teknolojiler edinmenin yaygın olarak kullanılan bir yöntemidir. Hem üst hem de alt düzey konularda (47) artan bir eğilim göstermiştir. Gelecekte, iş dünyasının bugünün dünyasından daha rekabetçi ve bilgi yoğun olacağı aşikardır. Bu nedenle, şirketlerin yeni bilgi ve teknolojiye duyacakları ihtiyaç nedeniyle ittifaklara daha fazla yönelecekleri değerlendirilmektedir. Bu ise konunun eğiliminin gelecekte de artmaya devam edeceği anlamına gelmektedir.

Tablo 16. Alt Düzey Konular İçin En Yüksek Olasılıklı Sözcükler ve Etiketler

Konu	En Yüksek Olasılıklı Sözcükler	Etiket	Yeni
0	forecasting forecast curve forecasts parameter logistic prediction cycles substitution estimation	Forecasting	
1	domestic developing_countries catch-up upgrading latecomer indigenous institutes provinces governments state-owned	Technological Upgrading	+
2	health medical healthcare care patients hospital clinical patient hospitals health_care	Healthcare	
3	business_model value_creation providers provider proposition delivery revenue co-creation offering offerings	Business Model	+
4	conflict boundary interdependence identity interpersonal collective cultures virtual champions cohesion	NPD Teams	
5	emissions scenario scenarios carbon climate emission climate_change consumption coal fuel	Climate Change	+
6	cluster clusters proximity spatial geographical spillovers geographic clustering agglomeration geography	Clusters	

Tablo 16 *(devam ediyor)*

7	centers federal defense military laboratory laboratories mission agencies agency sbir	Defense-related R&D	+
8	crisis safety security resilience disaster emergency adaptive crises event threat	Crisis Management	+
9	century modern revolution book principle mind societies economists mass living	Technological Revolution	+
10	city cities smart urban smart_city citizens housing smart_cities buildings residents	Smart City	+
11	absorptive tacit competencies knowledge_transfer routines knowledge_management dynamic_capabilities organizational_learning new_knowledge codification	Knowledge Management	
12	disruptive incumbent incumbents entrants discontinuous dominant_design designs technological_change mature inertia	Disruptive Technologies	
13	social_media big_data company_A tourism media company_B sites news websites website	Social Media	+
14	delphi round consensus panel opinion statements opinions rounds forecasting delphi_method	Delphi Method	+
15	students women faculty gender career student female teaching male graduate	Academic Productivity	+
16	cross-functional project_management npd_process front-end front_end launch npd_projects project_team proficiency execution	NPD Projects	
17	option cash valuation real_options asset profitability intangible volatility market_value earnings	Project Valuation	+
18	fuzzy algorithm weights optimization alternatives optimal scheduling weight simulation criterion	Scheduling	+

Tablo 16 *(devam ediyor)*

19	designers virtual objects prototype prototypes designer prototyping designs language object	Design	
20	biotechnology pharmaceutical drug biotech drugs clinical trials discovery pharmaceuticals gene	Biotechnology	
21	communities crowdsourcing open_source innovators code developers contest contests crowd solvers	Community Support	+
22	academics patenting faculty knowledge_transfer commercialization academia ttos engagement university-industry scientist	University-Industry Relations	
23	licensing commercialization license invention inventions licenses faculty royalty disclosure licensed	Licensing	
24	intention usage acceptance attitude usefulness attitudes satisfaction intentions privacy ease_of_use	Technology Acceptance	
25	institutes basic_research centres public_research co-operation applied_research laboratories labs budget profile	Research and Development	
26	incubator incubators spin-offs spin-off incubation ntbfs start-ups science_parks parks start-up	Incubation	
27	platform ecosystem platforms intermediaries ecosystems intermediary digital developers providers business_ecosystem	Ecosystems	+
28	workers mobility jobs labor skill human_capital skilled workforce wage employee	Labor Market	+
29	nanotechnology nano physics cern nanotechnologies nanotech fusion emerging_technologies sensor properties	Emerging Technologies	

Tablo 16 *(devam ediyor)*

30	agents adopters simulation agent adopter simulations system_dynamics technology_adoption modelling fitness	Diffusion (Technology Adoption)	
31	programme programmes commission public_sector agencies governments priorities oecd priority council	Policy-Making	
32	vehicle vehicles transport fuel cars automotive hybrid hydrogen mobility road	Motor Vehicle Industry	+
33	brand launch advertising brands reputation segment segments selling signals pricing	Brand	+
34	protection appropriability imitation protect appropriation iprs secrecy trademarks litigation patenting	Intellectual Property Rights (IPRs)	
35	legitimacy identity discourse collective logics logic norms framing tensions politics	(Politics of) Technology Assessment	
36	spillovers export expenditure spillover domestic expenditures exports goods imports machinery	International Technology Diffusion	+
37	diversification survival acquisitions acquired mergers exit hazard acquiring relatedness merger	Mergers and Acquisitions	+
38	organisations organisational organisation behaviour analysed analyse characterised commercialisation specialised specialisation	Organisational Behaviour	
39	smes small_firms innovators large_firms firm_size innovate innovating kibs manufacturing_firms turnover	Firm Size	

Tablo 16 *(devam ediyor)*

40	convergence configuration configurations causal quadrant negotiation converging configurational archetypes convergent	Technology Convergence	+
41	entrepreneurial entrepreneurs ventures entrepreneur venture venturing founders new_venture start-up start-ups	Entrepreneurship	
42	exploitation openness exploratory ambidexterity breadth focal exploitative slack depth moderating	Organisational Ambidexterity	+
43	ownership family board executives owners directors ceos shareholders corporate_governance owner	Corporate Governance	+
44	citations citation cited disciplines interdisciplinary bibliometric library disciplinary discipline sciences	Bibliometrics	
45	e-business alignment e-commerce agility lean deployment certification delivery automation adopting	Strategic Alignment	+
46	game digital media music video players games film movie player	Digital Media	+
47	alliance partner collaborations partnerships cooperative agreements partnership relational interorganizational consortia	Alliances	
48	distance class novelty classes similarity cloud computing pairs distant pair	Classification (Supervised Learning)	+
49	centrality nodes density node connections embeddedness network_analysis holes actor network_structure	Network Analysis	
50	electricity wind solar renewable_energy grid biomass heat renewable deployment nuclear	Renewable Energy	+

Tablo 16 *(devam ediyor)*

51	banks bank banking rural icts credit income poverty inclusive developing_countries	Subsistence Economies	+
52	evolutionary mode sectoral trajectories technological_change modes trajectory paradigm systemic accumulation	Technological Change	
53	green sustainability regulatory regulation iste recycling pollution regulations eco-innovation sustainable_development	Sustainability	
54	supplier outsourcing buyer contract contracts client buyers sourcing purchasing procurement	Buyer-Supplier Relationships	
55	regime transitions niche sustainability regimes niches socio-technical water landscape societal	Sustainability Transitions	+
56	subsidiaries subsidiary internationalization home mncs overseas abroad mnes multinational domestic	Multinational Enterprises (MNEs)	+
57	patenting inventors citations invention inventions inventor granted citation number_of_patents inventive	Patents (IPRs)	
58	percent corporation business_units spending metrics breakthrough revenue executives executive business_unit	Portfolio Management	
59	innovativeness constructs item scales moderating antecedents correlations latent loadings indirect	Organizational Innovativeness	
60	investors financing venture_capital finance equity crowdfunding startups venture fund investor	Startup Financing	
61	oecd capita income nations developing_countries world_bank country_A human_capital expenditure country_B	National Systems	

Tablo 16 *(devam ediyor)*

62	consumers preferences preference purchase attribute subjects experiment resistance consumption hedonic	Consumer Evaluation	
63	plant aircraft plants machine steel engine machines maintenance mechanical metal	Metallurgical Industry	+
64	roadmap roadmapping roadmaps intelligence maturity technology_management technology_development technology_roadmapping layer technology_roadmap	Roadmapping	
65	subsidies subsidy matching treatment additionality grants treated credits credit control_group	Subsidies	+
66	food agricultural water farmers agriculture land crop forest farm soil	Food Industry	
67	creativity cognitive personality emotional cognition style psychology memory intuition traits	Creativity	
68	architecture modularity modular modules module interfaces interface subsystems architectural architectures	Product Architecture	
69	dummy estimation regression dummies regressions specification explanatory control_variables standard_errors heterogeneity	Estimation Methods	+
70	score scores metrics ranking outputs rank rating rankings proposals evaluations	Decision-Making	
71	semiconductor electronics computers chip optical memory devices manufacturers hardware company_C	Semiconductor Industry	+
72	interview interviewees meetings interviewed in-depth documents coding interviewee contact decided	Data Collection Methods	+
73	mining topics text documents clusters keyword cluster clustering frequency discovery	Text Mining	+

Tablo 16 *(devam ediyor)*

74	stakeholders stakeholder engagement societal outputs triple_helix health dialogue library practitioners	Stakeholder Management	+
75	satisfaction employee rewards climate autonomy leader career intrinsic professionals reward	Career System	+
76	mobile standardization telecommunications broadband standardisation phone wireless providers switching mobile_phone	Telecommunications Industry	+
77	supply_chain manufacturer chain manufacturers inventory rfid logistics retailer retail supply_chains	Supply Chain Management	
78	optimal utility equilibrium proposition profits incentive marginal prices revenue game	Pricing	
79	foresight scenario scenarios futures exercise uncertainties workshop workshops backcasting visions	Foresight	

Tabloda, *Konu, En Yüksek Olasılıklı Sözcükler* ve *Etiket* sütunları bir önceki tablo ile aynı anlamlara sahiptirler. Bu tabloya eklenen *Yeni* sütunu, tespit edilen konunun önceki çalışmalardan herhangi birinde var olup olmadığını göstermektedir. Alt düzey konularda dikkate değer veya açıklama gerektiren hususlar şunlardır:

- Yeni sütununa göre ilk kez tespit edilen konu sayısı 38'dir.
- En çok çalışılan sektörler şunlardır: Savunma Sanayi (7), Motorlu Araç Sanayi (32), Dijital Medya Sanayi (46), Metalurji Sanayi (63), Gıda Sanayi (66), Yarı İletken Sanayi (71) ve Telekomünikasyon Sanayi (76).
- En çok kullanılan yöntemler şunlardır: Bibliyometri (44), Sınıflandırma (48), Ağ Analizi (49), Tahmin (69) ve Metin Madenciliğidir (73). Delphi (14) sadece çalışmalarda kullanılan bir yöntem olarak değil, yöntemin kendisi ile ilgili yapılan çalışmalarda ele alınan bir konu olarak da tespit edilmiştir.
- Toplum Desteği (21) ve Sınıflandırma (48) çalışmanın şaşırtıcı konuları olarak belirlenmiştir. Öznel bir değerlendirme olarak, bu konuların TM alanında en çok çalışılan 80 konu arasında tespit edilmeleri beklenen bir durum değildir.

- İşgücü Piyasası (28) ve Kariyer Sistemi (75), Yayılım (30) ve Uluslararası Teknoloji Yayılımı (36) ve Sürdürülebilirlik (53) ve Sürdürülebilirlik Geçişleri (55) gibi birbiriyle yakından ilişkili konular tespit edilmiştir. Bu konuların da üst düzey konularda yapıldığı gibi kümelenmelerinin doğru olacağı değerlendirilmektedir.
- Ekosistemler (27) platform ekosistemleri, iş ekosistemleri ve inovasyon ekosistemleri gibi konuları kapsamaktadır ve son dönemde keskin bir artış eğilimi göstermiştir.
- Savunma ile ilgili Ar-Ge (7) etiketli konu, savunma sanayiinin Ar-Ge'nin ana itici güçlerinden biri olduğunu göstermektedir ancak Araştırma ve Geliştirme (25) eğilimindeki düşüşle uyumlu bir düşüş eğilimi göstermiştir.
- Konuların eğilimleri ile ilgili hususlar aşağıdaki gibidir:
    - Keskin bir şekilde artan eğilimlere sahip konular Akıllı Şehir (10), Ekosistemler (27) ve Örgütsel Çok Yönlülük (42)'dir.
    - Keskin bir düşüş eğilimi gösteren konu Stratejik Uyum (45)'dur.
    - Trendleri artan konular Üniversite-Sanayi İlişkileri (22), Girişimcilik (41), Sürdürülebilirlik (53), Sürdürülebilirlik Geçişleri (55), Tahmin Yöntemleri (69), Metin Madenciliği (73) ve Paydaş Yönetimi (74)'dir.
    - Eğilimleri azalan konular Savunma ile ilgili Ar-Ge (7), YÜG Projeleri (16), Proje Değerleme (17), Araştırma ve Geliştirme (25), Teknolojik Değişim (52), Portföy Yönetimi (58), Metalurji Endüstrisi (63) , Yarı İletken Endüstrisi (71) ve Veri Toplama Yöntemleri (72)'dir.

Son olarak, konu modellemeye ilgi duyanlar için bir not olmak üzere, alt düzey konulardaki her konunun en yüksek olasılıklara sahip olduğu 100 belgesinin türleri (toplam 8.000 belge) incelenmiş ve sadece üç özet bulunduğu tespit edilmiştir. Bu durum, tam metin makalelerden oluşan bir derleme özet eklenmesinin, derlemden çıkarılan konulara önemli bir katkısının olmadığını göstermektedir.

G. THESIS PERMISSION FORM / TEZ İZİN FORMU

(Please fill out this form on computer. Double click on the boxes to fill them)

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☒

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : TEKİN
Adı / Name : YAŞAR
Bölümü / Department : Bilim ve Teknoloji Politikası Çalışmaları / Science and Technology Policy Studies

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English): APPLICATION OF TEXT MINING TO TECHNOLOGY MANAGEMENT DOMAIN TO EXTRACT TOPICS AND TRENDS

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☒

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☒
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two years. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. / A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature ..... **Tarih** / Date .....
(Kütüphaneye teslim ettiğiniz tarih. Elle doldurulacaktır.)
(Library submission date. Please fill out by hand.)

Tezin son sayfasıdır. / This is the last page of the thesis/dissertation.