Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Reddit Platformu Üzerinden Bilimle İlgili Gönderilerden İlişkili Konu Modelleme Analizi ile Bilim Dünyasının Haritasının Çıkarılması

Merve YALÇIN [a], Semanur GÜRSOY [a], Özcan ÖZYURT [a,*]

[a] *Yazılım Mühendisliği Bölümü, OF Teknoloji Fakültesi, Karadeniz Teknik Üniversitesi, Trabzon, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: oozyurt@ktu.edu.tr*
DOI: 10.29130/dubited.1370614

Öz

Günümüz dünyasında ulaşılan teknolojinin ana kaynağı bilimdir. Bilim ve teknoloji alanlarındaki çalışmaların devam etmesiyle bilim dünyası her geçen gün yeniden şekillenmektedir. Bununla birlikte, teknolojinin gelişmesi teknolojik platformlardaki geleneksel yöntemlerle işlenemeyen veri miktarının her geçen gün artmasına sebep olmaktadır. Anlamlandırılmamış verinin işlenerek anlamlı hale getirilmesi şirketler, kurum ve kuruluşlar için büyük verinin yüksek oranda fayda sağlayan araçlar haline dönüştürülmesine olanak sağlayacaktır. Verinin işlenerek anlamlı hale getirilmesinde en etkili veri madenciliği tekniklerinden biri konu modellemedir. Bu çalışmada konu modelleme tekniklerinden olan ilişkili konu modelleme (İKM) kullanılarak Reddit platformu üzerinde bilimle ilgili paylaşımların anlamsal içerik analizi yapılmıştır. 2022 yılının ilk dokuz ayına ait Reddit paylaşımlarındaki gizli anlamlar ve bu anlamlar arasındaki korelasyon ortaya koyulmuş ve ilgili sonuçlar paylaşılmıştır. Elde edilen sonuçların bilime ilgili insanlara ve bilim insanlarına araştırmaları için fikir kaynağı olacağı düşünülmektedir.

Anahtar Kelimeler: *Veri Madenciliği, İlişkili Konu Modelleme, Reddit, Bilim*

Mapping the Science World with Correlated Topic Modeling Analysis from Science-Related Posts on the Reddit Platform

Abstract

The main source of technology in today's world is science. The world of science is being reshaped every day with the continuation of studies in the fields of science and technology. However, the development of technology causes the amount of data that cannot be processed with traditional methods on technological platforms to increase day by day. Making meaningful data meaningful by processing unmeaningful data will enable companies, institutions, and organizations to transform big data into highly beneficial tools. One of the most effective data mining techniques for processing and making sense of data is topic modeling. In this study, semantic content analysis of science-related posts on the Reddit platform was conducted using corelated topic modeling (CTM), one of the topic modeling techniques. In the first nine months of 2022, the hidden meanings in Reddit posts and the correlation between these meanings were revealed, and the relevant results were shared. It is thought that the results obtained will be a source of ideas for people interested in science and scientists for their research.
Keywords: *Data Mining, Correlated Topic Modeling, Reddit, Science*

Geliş: 03/10/2023, Düzeltme: 01/01/2024, Kabul: 10/01/2024

I. INTRODUCTION

With the development of technology and science, the use of technological devices such as cell phones, tablets, and computers has increased considerably. The use of technological devices on an individual basis increases the amount of data day by day. The total amount of data produced since the invention of the Internet, or even more, can now be produced in a single day. Social media has an important role in the increasing amount of data and the development of scientific understanding [1]. Especially Reddit, which is known as the "front face of the internet", has been the source of countless scientific studies since it is open-source compared to other social media platforms [2]. Due to its large user base and posts on every conceivable topic, the accumulated uninterpreted data constitutes a great resource for text mining [3].

With data mining, large-scale irregular data is processed and made meaningful for a purpose. This process allows for revealing relationships and patterns between data and making predictions about the future. By applying text mining to data in the social environment, the analysis process of unstructured textual data can be automated [4]. One of the most efficient approaches to handling and extracting insights from data is through the utilization of topic modeling, a technique prominent in data mining. Topic modeling, an unsupervised machine learning method, employs text mining strategies to arrange, delve into, and scrutinize extensive volumes of data [5], [6]. This technique facilitates rapid and straightforward analysis of data without necessitating training [6], [7]. Its applications are diverse, spanning various domains like topic modeling, the medical sector, scientific investigations, sentiment assessment, recommendation systems, concise summarization of text, and enhancing search queries for search engines [7]. An illustrative instance of topic modeling is Latent Dirichlet Allocation (LDA), which categorizes text within a document based on a specific subject [8]. LDA utilizes a probability model and Dirichlet distributions to uncover concealed meanings within documents [4].

An improved and more sophisticated iteration of LDA, Correlated Topic Modeling (CTM) is a modeling approach that reveals the connection between latent topics in documents. While the relationship between topics is ignored in the LDA algorithm, CTM reveals correlations between topics by replacing topic distributions with logistic normal distributions. As a result, CTM offers a more original model by connecting a hidden meaning to another meaning [4], [9]. Moreover, scientific studies have shown that CTM gives better results than LDA [10]. For example, Blei et al. applied CTM to a dataset of 57 million words of articles in the journal Science from 1990-1999 and obtained better results compared to LDA [9]. In addition, CTM provides a familiar way to visualize and make sense of unstructured documents [10].

In CTM, multiple topics are displayed in different proportions in each document. In this way, heterogeneity can be captured in grouped data containing many latent topics. A graphical representation of the terminology and notation used to describe the data, latent variables, and parameters in CTM is shown in Figure 1 [11].

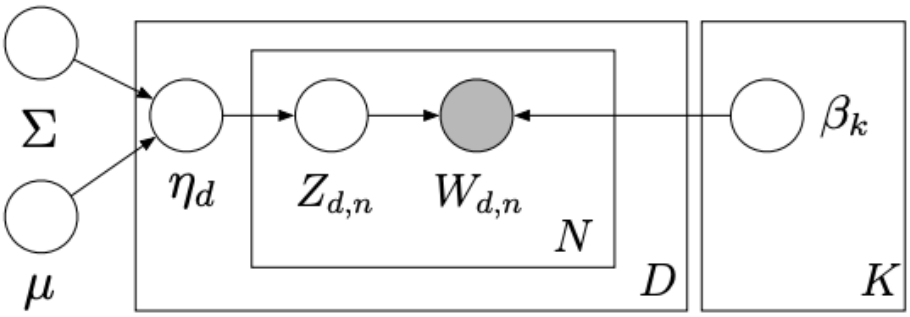

Figure 1. *Graphical representation of the CTM model*

The generative model corresponding to the graphical representation of the CTM in Figure 1 is as follows:

1. For each subject k, $k \epsilon K$
    a. Determine the K multinomial distribution of words within topics: $\beta k \sim N(\mu,\Sigma)$
2. For each document d, $d \epsilon D$
    a. Determine the K-dimensional vector, i.e. the topic distribution for each document: $\eta d \sim N(\mu,\Sigma)$
    b. For each word n, $n \epsilon \{1, \ldots, Nd\}$
        i. Assign a sample topic: $Zd,n \sim Mult(\theta)$
        ii. Choose a sample word: $Wd,n \sim Mult(\beta Zd,n)$

The graphical representation and the parameters specified in the generative model and their corresponding meanings are presented in Table 1.

Table 1. *Parameters used in CTM and their descriptions*

Parameter	**Description**
K	Number of latent meanings
D	All documents in the collection
k=1,2,…,K	Index of latent meanings
d=1,2,…,D	Index to an individual document in the collection
N	All words in documents
$W_{d,n}$	nth word in dth document
$Z_{d,n}$	The topic determined for the nth word in the dth document
β_k	Word distribution per topic
η_d	Topic distribution per document
μ	K-dimensional average
Σ	KxK covariance matrix
Θ	Document-specific topic rates

According to the CTM model, the only observable random variables are words organized into documents [9]. β1: Defines the K multinomial distribution for modeling the conditioned word on K topics [12]. Wd,n corresponds to the n-th word in the d-th document and is an item stored in a V-term dictionary. A topic β is therefore a distribution over the dictionary and is expressed as V-1. Since the model contains K topics, it is denoted as β1:K. Each word is assumed to be taken from one of the K topics, and topic assignment is performed Zd,n. Furthermore, each document is associated with a set of topic rates θ. Thus, θ represents a distribution over topic indexes and reflects the probability of retrieving words from each topic in the collection [9].

Many studies have been conducted in the field of CTM. Some of these studies are as follows: McDermott et al. (2022) analyzed 4771 articles on child-computer interaction published between 2021 and 2023 with CTM and presented a comprehensive empirical map of child-computer interaction. The results allow us to characterize child-computer interaction as a vibrant and diverse research environment that evolves dynamically over time, exhibiting increasing specialization and the emergence of different subfields, and moving from a technology-driven to a needs-driven agenda [13]. Blei and Lafferty (2007) analyzed the articles in Science magazine published between 1990 and 1999, consisting of 57 million words in total, by applying CTM. As a result of the analysis, it was revealed that CTM gave better results than LDA and supported more topics compared to LDA [9]. Xu et al. (2013) analyzed an image dataset using extended CTM and LDA methods to propagate topic correlations from image features to description words, and the analysis revealed that CTM performed better than LDA [14]. He et al. (2017) used an improved CTM method to reveal hidden meanings in over two million NYTimes news articles. In addition, a model was created that reveals topic correlations based on proximity between topic

vectors. With this model, it was shown that very large amounts of data and models can be processed with high performance [15]. Daenekindt and Huisman (2020) analyzed the abstracts of 16,978 articles on higher education between 1991 and 2018 with CTM. The analysis showed that scientific interest has changed over time. Some themes became more central over time, while other themes that may have been at the forefront and dominated the field became secondary over time. We also investigated which themes tended to cluster with each other. As a result, it was found that the clustering showed a steady state over time [16]. Dybowski and Adämmer (2018) combined CTM, a probabilistic topic model, with dictionary-based sentiment analysis to create a time series showing when and how the US president communicates tax policy news to the public. For this, 89,843 presidential documents were analyzed for tax policy content. Econometric analysis revealed that optimistic tax policy statements stimulate production, consumption, and investment even after the tax proposal is adopted [17]. Tu, Xia, and Wang (2014) used optical flow and IKM methods to recognize complex human action. First, the missing data on point trajectories were reconstructed. Then, using optical flow, a human silhouette was extracted based on width and height ratios. In the action classification phase, IKM was utilized. In the study on three different datasets, the proposed method was shown to be more effective than the compared methods [18]. Aznag et al. (2013) used probabilistic latent semantic analysis (PLSA), LDA, and a CTM to extract hidden meanings from web service descriptions. The analysis revealed that CTM outperformed PLSA and LDA by revealing hidden meanings. These hidden meanings can then be used as an efficient discovery and ranking mechanism for web services [19].

In this study, semantic content analysis of science-related posts on the Reddit platform was conducted with CTM as a new application area. In the first nine months of 2022, the hidden meanings in Reddit posts and the correlation between these meanings were revealed, and the relevant results were shared.

II. METHOD

A. DATA COLLECTION

We used files provided by Pushshift.io, which contains all Reddit data from 2005 to the present on a monthly basis. The posts were pulled through Pushshift.io. However, after Reddit withdrew support for Pushshift.io, other methods were explored, and comments were pulled through academic torrents. Due to the large size of a monthly post file, the monthly files were split into 10,000-line files by using the'split' command on Git Bash. Then, a Python script was written to extract the posts under the title of science. This process was repeated for each month, and the same method was applied for the comment files. The post data obtained for each month, together with the comments on that post, were combined to contain a maximum of 10,000 words. The data set with 21400 data points, consisting of 8,456,995 words, was saved in csv format.

B. REALIZATION OF DATA PRE-PROCESSING STAGES

The success of data mining is based on cleaning up the data. These processes are also known as data preprocessing. In topic modeling, the data set needs to go through certain steps in order not to adversely affect the analysis. The nltk and tomotopy libraries of the Python language were used to perform these steps.

B. 1. Deletion of Irrelevant Content

Since topic modeling algorithms are case sensitive, all uppercase letters in the posts were converted to lowercase letters. In addition, punctuation marks and numeric values that do not make sense for the algorithm were removed from the posts. These operations were performed with Python's own libraries.

B. 2. Removal of Stop Words

Stop words, words that occur frequently in the text but do not affect the semantic expression, are removed from the posts. This simplification speeds up the algorithm by making it easier to find meaningful patterns.

B. 3. Identifying Word Roots

At this stage, the words that make up the posts are reduced to their basic form. The stemming process used in text normalization reduces the word to its root and standardizes the words that make up the data set. In this way, it positively affects the algorithm's performance by reducing the number of unique words.

C. CTM MODEL CREATION SETTINGS

The Tomotopy library in Python was used to create the CTM models. Many parameters are used when creating the CTM. K is the number of topics, taking values between 1 and 32,767. This value was chosen according to the consistency test. min_df is the minimum document frequency of words. Words with a document frequency less than min_df are removed from the model. The default value is 0, meaning that no words are excluded. min_df was used in the model with a value of 5. rm_top is the number of very common words to exclude. To exclude these words from the model, it is necessary to set this value to 1 or more. The default value is 0, which means that no very common words are removed. rm_top was set to 40. num_beta_sample is the number of times beta parameters are sampled; the default value is 10. CTModel samples the beta parameter num_beta_sample for each document. The more betas sampled, the more accurate the distribution, but the longer it takes to learn. If there are few documents in the model, keeping this value large helps to get better results. Given the sufficient number of documents, we used a num_beta_sample of 5.

C. 1. Coherence Values

In order to find the model with appropriate consistency values, models were created between 5 and 40 topics, and consistency values were compared. These consistency values are given in Figure 2. The k value was chosen as 34 to have better consistency values and to reveal the relationship between the topics more clearly.

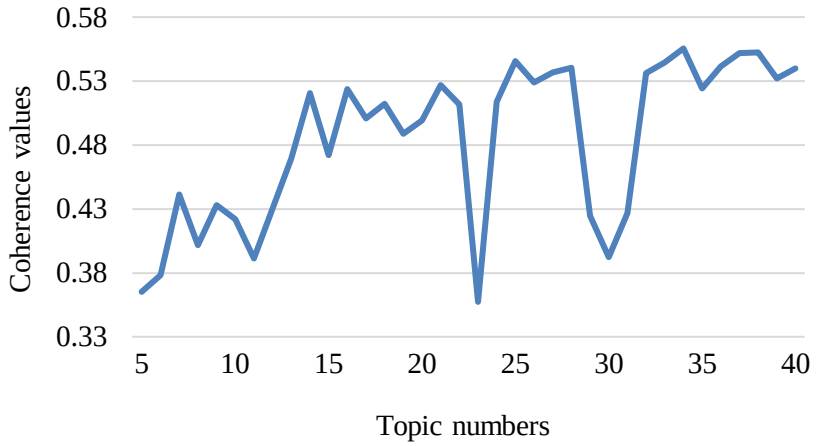

Figure 2. *CTM number of topics - consistency values*

III. FINDINGS

In this study, topic modeling of 21,400 science-related posts and a maximum of 10,000 words from these posts for the first nine months of 2022 on the Reddit platform was performed. The data obtained was subjected to pre-processing stages, then the topics were identified by applying the CTM model, and the relationships between topics were revealed. Tags were assigned for each topic, and the results were visualized.

The inter topic distance map is based on the visualization of topics in a two-dimensional space. The area of the circles representing the topics in the visual is directly proportional to the number of words belonging to the topic. In addition, topics that are closer to each other have more words in common. The distance map between topics is as shown in Figure 3. For example, topic 8 (renewable energy in agriculture) has common words with topic 12 (vehicle-borne air pollution) and topic 22 (social causes of death).

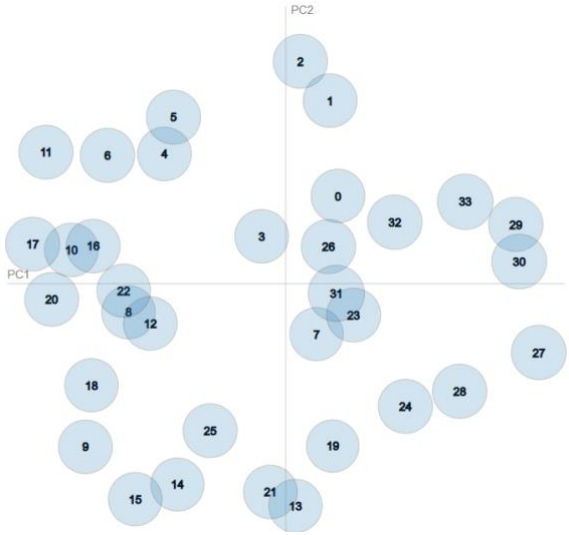

Figure 3. *Distance map between topics*

The bar scatter plot of the topic distributions is shown in Figure 4. Among 35 topics, scientific research (topic 31) has the highest distribution rate. This is followed by risk factors increasing with age (topic 27) and brain imaging studies (topic 33). The topics with the lowest distribution rate are eating habits and health (topic 19) and taxation and crime relationship (topic 14).

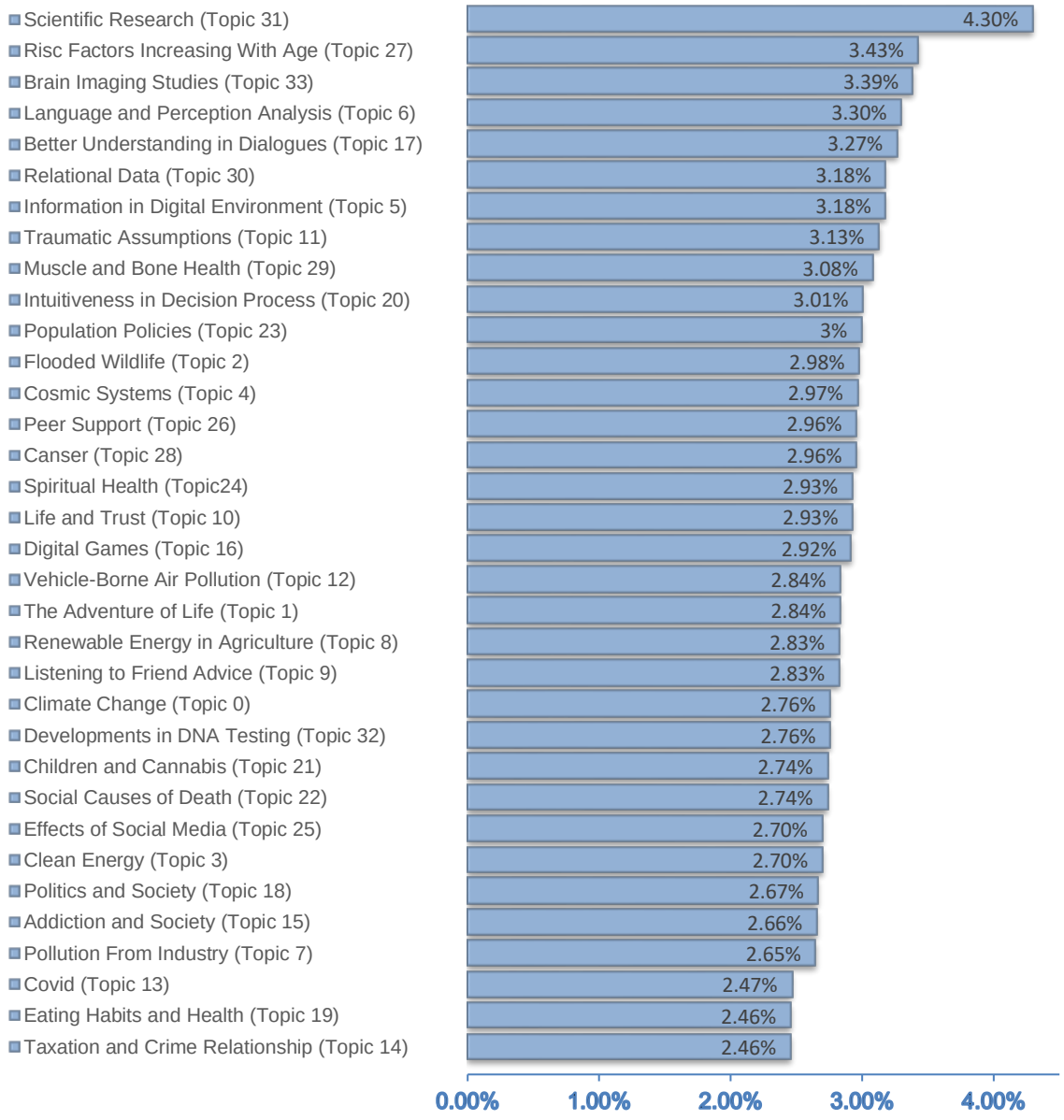

Figure 4. *Bar graph of topic distributions*

The visual showing the relationships between the topics is shown in Figure 5. There are relationships between almost all topics. However, when all the relationships are shown on the figure, the relationships above the threshold value of 0.46 are shown because it is difficult to understand the figure. The thick link between the topics shows strong relationships, while the thin link shows relatively weaker relationships.

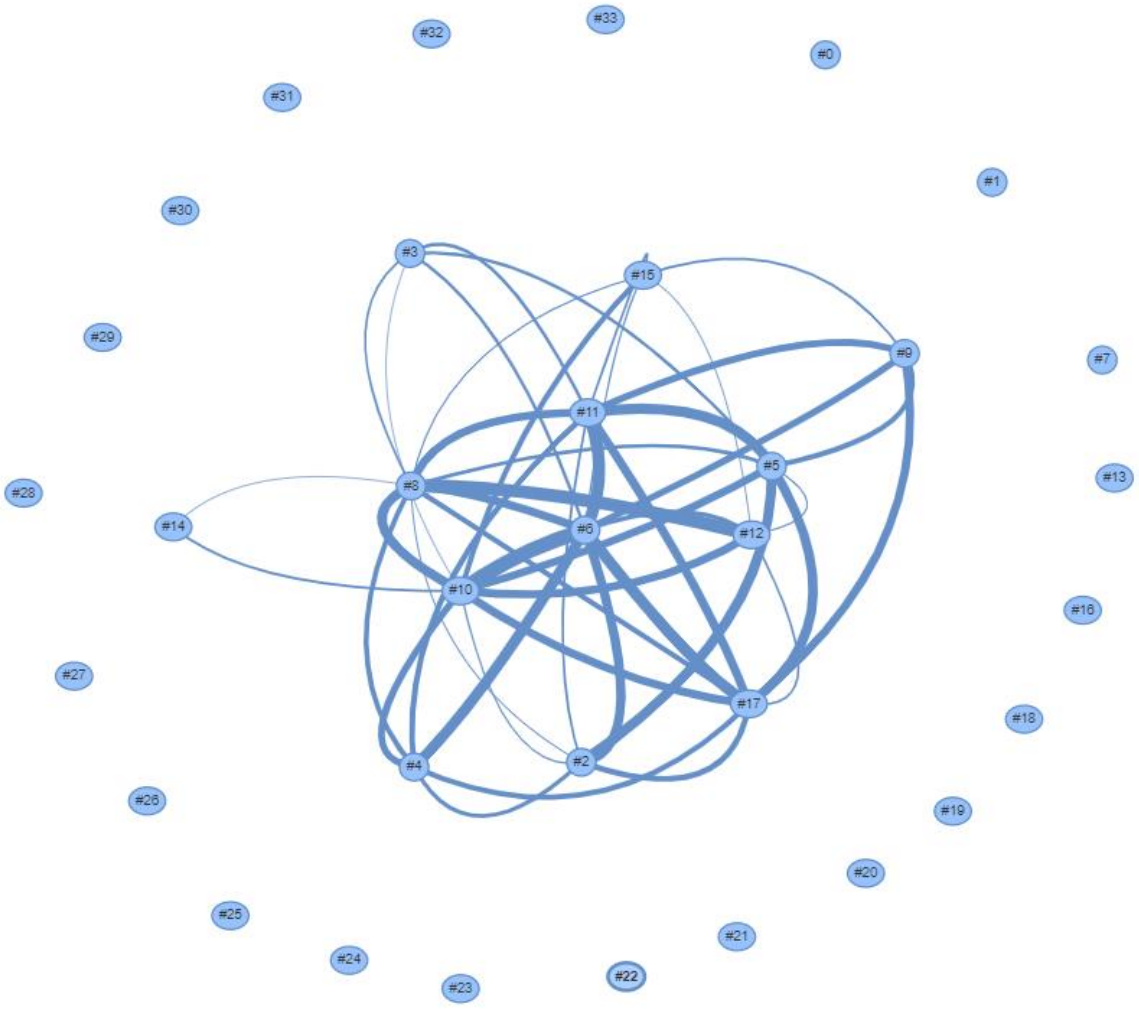

Figure 5. *Visualization of relationships between topics*

The relationships between the topics are also shown in the numerical values in Figure 6. In the figure, positive values indicate that there is a positive relationship between the topics, while negative values indicate that there is an opposite relationship between the topics. It is also seen that the relationship between the topic and itself is expressed as 1. Looking at Figure 5, the strong relationship between topic-6 (language and perception analysis) and topic-10 (life and trust) is represented in Figure 6 with a value of 0.51. Considering how much people understand each other with the language they use as a means of communication and the importance of language for the formation of trust in human relations, it will be understood that this relationship is meaningful.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
0	1.00	0.39	0.37	0.40	0.38	0.36	0.37	0.37	0.39	0.33	0.36	0.36	0.37	0.25	0.33	0.31	0.34	0.36	0.33	0.27	0.31	0.25	0.35	0.33	0.19	0.26	0.27	0.26	0.15	0.30	0.26	0.13	0.26	0.18
1	0.39	1.00	0.46	0.43	0.43	0.44	0.44	0.41	0.44	0.38	0.41	0.42	0.42	0.32	0.38	0.35	0.38	0.41	0.35	0.31	0.35	0.31	0.37	0.37	0.26	0.29	0.34	0.34	0.25	0.39	0.34	0.11	0.30	0.24
2	0.37	0.46	1.00	0.45	0.47	0.49	0.49	0.42	0.46	0.44	0.46	0.46	0.44	0.38	0.40	0.42	0.42	0.47	0.39	0.33	0.38	0.38	0.37	0.42	0.33	0.34	0.39	0.39	0.30	0.42	0.40	0.18	0.30	0.24
3	0.40	0.43	0.45	1.00	0.46	0.46	0.46	0.44	0.46	0.42	0.46	0.46	0.45	0.35	0.42	0.41	0.42	0.45	0.42	0.36	0.39	0.34	0.40	0.40	0.29	0.34	0.37	0.34	0.25	0.37	0.34	0.16	0.30	0.22
4	0.38	0.43	0.47	0.46	1.00	0.48	0.49	0.43	0.47	0.45	0.48	0.47	0.45	0.38	0.41	0.43	0.44	0.47	0.41	0.35	0.42	0.37	0.39	0.42	0.32	0.35	0.38	0.36	0.27	0.39	0.38	0.16	0.30	0.24
5	0.36	0.44	0.49	0.46	0.48	1.00	0.52	0.44	0.47	0.47	0.48	0.49	0.46	0.41	0.43	0.45	0.43	0.49	0.42	0.33	0.43	0.41	0.39	0.45	0.34	0.39	0.42	0.39	0.32	0.41	0.43	0.19	0.28	0.24
6	0.37	0.44	0.49	0.46	0.49	0.52	1.00	0.43	0.48	0.48	0.51	0.50	0.47	0.42	0.44	0.46	0.45	0.49	0.43	0.35	0.44	0.41	0.41	0.45	0.37	0.39	0.42	0.40	0.32	0.42	0.43	0.20	0.29	0.23
7	0.37	0.41	0.42	0.44	0.43	0.44	0.43	1.00	0.46	0.41	0.44	0.43	0.45	0.36	0.42	0.40	0.41	0.44	0.41	0.40	0.37	0.37	0.39	0.39	0.32	0.33	0.34	0.35	0.28	0.39	0.34	0.17	0.26	0.19
8	0.39	0.44	0.46	0.46	0.47	0.47	0.48	0.46	1.00	0.46	0.49	0.48	0.50	0.40	0.46	0.46	0.44	0.47	0.44	0.40	0.44	0.38	0.44	0.42	0.33	0.38	0.37	0.37	0.27	0.39	0.35	0.17	0.27	0.20
9	0.33	0.38	0.44	0.42	0.45	0.47	0.48	0.41	0.46	1.00	0.49	0.48	0.45	0.42	0.44	0.46	0.45	0.48	0.43	0.36	0.44	0.44	0.42	0.43	0.36	0.41	0.39	0.35	0.29	0.37	0.36	0.19	0.25	0.18
10	0.36	0.41	0.46	0.46	0.48	0.48	0.51	0.44	0.49	0.49	1.00	0.49	0.48	0.42	0.46	0.47	0.45	0.48	0.45	0.37	0.45	0.42	0.43	0.43	0.34	0.41	0.41	0.36	0.28	0.38	0.37	0.19	0.26	0.20
11	0.36	0.42	0.46	0.46	0.47	0.49	0.50	0.43	0.48	0.48	0.49	1.00	0.47	0.41	0.45	0.46	0.45	0.48	0.45	0.36	0.45	0.40	0.42	0.42	0.34	0.39	0.39	0.36	0.27	0.38	0.37	0.19	0.27	0.21
12	0.37	0.42	0.44	0.45	0.45	0.46	0.47	0.45	0.50	0.45	0.48	0.47	1.00	0.39	0.45	0.46	0.44	0.46	0.44	0.37	0.43	0.40	0.42	0.41	0.32	0.39	0.38	0.35	0.27	0.37	0.34	0.17	0.25	0.19
13	0.25	0.32	0.38	0.35	0.38	0.41	0.42	0.36	0.40	0.42	0.42	0.41	0.39	1.00	0.40	0.43	0.38	0.40	0.39	0.35	0.38	0.43	0.36	0.40	0.38	0.38	0.35	0.36	0.30	0.34	0.35	0.16	0.18	0.16
14	0.33	0.38	0.40	0.42	0.41	0.43	0.44	0.42	0.46	0.44	0.46	0.45	0.45	0.40	1.00	0.45	0.42	0.44	0.43	0.35	0.42	0.40	0.41	0.39	0.31	0.39	0.35	0.33	0.26	0.34	0.32	0.16	0.22	0.16
15	0.31	0.35	0.42	0.41	0.43	0.45	0.46	0.40	0.46	0.46	0.47	0.46	0.46	0.43	0.45	1.00	0.43	0.45	0.44	0.38	0.44	0.43	0.41	0.42	0.36	0.42	0.38	0.36	0.28	0.36	0.35	0.18	0.21	0.18
16	0.34	0.38	0.42	0.42	0.44	0.43	0.45	0.41	0.44	0.45	0.45	0.45	0.44	0.38	0.42	0.43	1.00	0.45	0.42	0.34	0.41	0.38	0.38	0.38	0.31	0.36	0.35	0.32	0.26	0.34	0.33	0.16	0.24	0.19
17	0.36	0.41	0.47	0.45	0.47	0.49	0.49	0.44	0.47	0.48	0.48	0.48	0.46	0.40	0.44	0.45	0.45	1.00	0.44	0.36	0.44	0.40	0.41	0.43	0.35	0.38	0.40	0.36	0.30	0.39	0.37	0.17	0.26	0.23
18	0.33	0.35	0.39	0.42	0.41	0.42	0.43	0.41	0.44	0.43	0.45	0.45	0.44	0.39	0.43	0.44	0.42	0.44	1.00	0.35	0.41	0.37	0.39	0.38	0.30	0.36	0.34	0.32	0.24	0.31	0.30	0.16	0.20	0.15
19	0.27	0.31	0.33	0.36	0.35	0.33	0.35	0.40	0.40	0.36	0.37	0.36	0.37	0.35	0.35	0.38	0.34	0.36	0.35	1.00	0.33	0.34	0.33	0.33	0.32	0.29	0.28	0.30	0.26	0.32	0.28	0.13	0.18	0.12
20	0.31	0.35	0.38	0.39	0.42	0.43	0.44	0.37	0.44	0.44	0.45	0.45	0.43	0.38	0.42	0.44	0.41	0.44	0.41	0.33	1.00	0.37	0.38	0.37	0.30	0.36	0.35	0.28	0.24	0.31	0.32	0.15	0.20	0.16
21	0.25	0.31	0.38	0.34	0.37	0.41	0.41	0.37	0.38	0.44	0.42	0.40	0.40	0.43	0.40	0.43	0.38	0.40	0.37	0.34	0.37	1.00	0.35	0.39	0.37	0.36	0.34	0.35	0.31	0.34	0.34	0.14	0.19	0.17
22	0.35	0.37	0.37	0.40	0.39	0.39	0.41	0.39	0.44	0.42	0.43	0.42	0.42	0.36	0.41	0.41	0.38	0.41	0.39	0.33	0.38	0.35	1.00	0.34	0.25	0.35	0.31	0.28	0.18	0.30	0.26	0.14	0.20	0.14
23	0.33	0.37	0.42	0.40	0.42	0.45	0.45	0.39	0.42	0.43	0.43	0.42	0.41	0.40	0.39	0.42	0.38	0.43	0.38	0.33	0.37	0.39	0.34	1.00	0.37	0.36	0.37	0.39	0.30	0.38	0.38	0.14	0.23	0.20
24	0.19	0.26	0.33	0.29	0.32	0.34	0.37	0.32	0.33	0.36	0.34	0.34	0.32	0.38	0.31	0.36	0.31	0.35	0.30	0.32	0.30	0.37	0.25	0.37	1.00	0.30	0.31	0.36	0.34	0.33	0.33	0.12	0.18	0.17
25	0.26	0.29	0.34	0.34	0.35	0.39	0.39	0.33	0.38	0.41	0.41	0.39	0.39	0.38	0.39	0.42	0.36	0.38	0.36	0.29	0.36	0.36	0.35	0.36	0.30	1.00	0.34	0.30	0.15	0.29	0.30	0.14	0.16	0.14
26	0.27	0.34	0.39	0.37	0.38	0.42	0.42	0.34	0.37	0.39	0.41	0.39	0.38	0.35	0.35	0.38	0.35	0.40	0.34	0.28	0.35	0.34	0.31	0.37	0.31	0.34	1.00	0.33	0.26	0.32	0.33	0.14	0.22	0.21
27	0.26	0.34	0.39	0.34	0.36	0.39	0.40	0.35	0.37	0.35	0.36	0.36	0.35	0.36	0.33	0.36	0.32	0.36	0.32	0.30	0.28	0.35	0.28	0.39	0.36	0.30	0.33	1.00	0.30	0.36	0.34	0.10	0.20	0.18
28	0.15	0.25	0.30	0.25	0.27	0.32	0.32	0.28	0.27	0.29	0.28	0.27	0.27	0.30	0.26	0.28	0.26	0.30	0.24	0.26	0.24	0.31	0.18	0.30	0.34	0.15	0.26	0.30	1.00	0.33	0.30	0.06	0.20	0.23
29	0.30	0.39	0.42	0.37	0.39	0.41	0.42	0.39	0.39	0.37	0.38	0.38	0.37	0.34	0.34	0.36	0.34	0.39	0.31	0.32	0.31	0.34	0.30	0.38	0.33	0.29	0.32	0.36	0.33	1.00	0.38	0.04	0.27	0.26
30	0.26	0.34	0.40	0.34	0.38	0.43	0.43	0.34	0.35	0.36	0.37	0.37	0.34	0.35	0.32	0.35	0.33	0.37	0.30	0.28	0.32	0.34	0.26	0.38	0.33	0.30	0.33	0.34	0.30	0.38	1.00	-0.02	0.27	0.27
31	0.13	0.11	0.18	0.16	0.16	0.19	0.20	0.17	0.17	0.19	0.19	0.19	0.17	0.16	0.16	0.18	0.16	0.17	0.16	0.13	0.15	0.14	0.14	0.14	0.12	0.14	0.14	0.10	0.06	0.04	-0.02	1.00	-0.07	-0.18
32	0.26	0.30	0.30	0.30	0.30	0.28	0.29	0.26	0.27	0.25	0.26	0.27	0.25	0.18	0.22	0.21	0.24	0.26	0.20	0.18	0.20	0.19	0.20	0.23	0.18	0.16	0.22	0.20	0.20	0.27	0.27	-0.07	1.00	0.27
33	0.18	0.24	0.24	0.22	0.24	0.24	0.23	0.19	0.20	0.18	0.20	0.21	0.19	0.16	0.16	0.18	0.19	0.23	0.15	0.12	0.16	0.17	0.14	0.20	0.17	0.14	0.21	0.18	0.23	0.26	0.27	-0.18	0.27	1.00

Figure 6. *Values expressing the relationships between topics ("Topic-Topic relationship")*

IV. CONCLUSION

In this study, the CTM model was successfully applied to the Reddit data set, known as the "front face of the internet", related to science, consisting of 8,456,995 words, for the first nine months of 2022, and the hidden topics of the data set and the correlations between these topics were revealed. In the study, it was seen that Topic 31, which has the scientific research label, had the highest rate among the topics with 4.30%. In terms of relationships between topics, it was seen that Topic 6 and Topic 10 had the highest relationship with a value of 0.51. One of the most important problems encountered while carrying out the study was obtaining the data set. For this, firstly, the Reddit API provided by Pushshift.io was used, but due to the restriction imposed by Reddit, the necessary data could not be retrieved. Then the web scraping method was tried, and the same restriction was encountered. Finally, Pushshift.io's monthly files containing all Reddit data from 2005 to the present were used, and subreddits were extracted using this method. Reddit restrictions were encountered when withdrawing comments, and another method was investigated, and monthly comment files were extracted via Academic Torrents. By purifying irrelevant data from these files, science-related data was obtained. Considering this situation in future studies, data extraction can be carried out successfully via academic torrents, while the relevant steps can be implemented more easily by following the sharing on Github. In the study, IKM, one of the models that has not been studied much in the field of subject modeling, was used. It is thought that this study can be used as a resource in the field of subject modeling, especially for people who want to work on IKM and for those who are interested in science. In addition, this study makes it possible to predict the course of science-related research in the future by looking at the current state of science.

ACKNOWLEDGEMENT: The activities carried out in this study were supported by TUBITAK in 2023 as study number 1919B012220329 within the scope of the TUBITAK 2209-A University Students Domestic Research Projects Support Program.

V. REFERENCES

[1] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn and P. Io, "The Pushshift Reddit Dataset", in *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, Zenodo, 2020.

[2] F. Tekin ve A. Turan , "Çalışan kadınların sosyal medya kullanım karakteristikleri", *Sakarya Üniversitesi İşletme Enstitüsü Dergisi*, c. 2, sayı. 1, ss. 27-32, 2020.

[3] U. Yakar (2020). *Geniş İçeriği ile Dikkat Çeken Sosyal Platform Reddit Nedir, Ne İşe Yarar, Nasıl Kullanılır?* [Çevrimiçi]. Erişim: https://www.webtekno.com/reddit-nedir-ne-ise-yarar-kullanim-h120297.html, 2020.

[4] A. Kaya, and E. Gülbandılar, "Konu Modelleme Yöntemlerinin Karşılaştırılması", *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, c. 3, sayı. 2, ss. 46-53, 2022.

[5] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling", *Computer Science and Informnation Technology*, vol. 6, pp. 201–210, 2016.

[6] F. Pascual. (2019, September 26). *Topic Modeling: An Introduction* [Online]. Available: https://monkeylearn.com/blog/introduction-to-topic-modeling/#what-is

[7] Y. Peddireddi. (2021, May 1). *Topic Modelling in Natural Language Processing* [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/topic-modelling-in-natural-language-processing/

[8] S. Li. (2018, May 31). *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python* [Online]. Available: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

[9] D. M. Blei and J. D. Lafferty, "A Correlated Topic Model of Science", *The Annals of Applied Statics*, vol. 1(1), pp. 17-35, 2007.

[10] D. M. Blei and J. D. Lafferty, "Correlated Topic Models" , *Advances in Neural Information Processing Systems*, vol. 18, 2005.

[11] M. K. Oo and M. A. Khine, "Correlated Topic Modeling for Big Data with MapReduce", in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE).*, Nara, Japan, 2018, pp. 408-409.

[12] K. Salomatin, Y. Yang and A. Lad, "Multi-field Correlated Topic Modeling", in *Proceedings of the SIAM International Conference on Data Mining, SDM,*. Sparks, Nevada, USA, 2009, pp. 628-637.

[13] T. McDermott, J. Robson, N. Winters and L. E. Malmberg, "Mapping the Changing Landscape of Child-Computer Interaction Research Through Correlated Topic Modelling", *in Proceedings of Interaction Design and Children, IDC,* Braga, Portugal, 2022, pp. 82–97.

[14] X. Xu, A. Shimada and R. I. Taniguchi, “Correlated Topic Model For Image Annotation”, *in FCV 2013 - Proceedings of the 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, 2013, pp. 201–208.

[15] J. He, Z. Hu, T. Berg-Kirkpatrick, Y. Huang and E. P. Xing, “Efficient Correlated Topic Modeling With Topic Embedding”, *in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, vol. Part F129685, pp. 225–233.

[16] S. Daenekindt and J. Huisman, “Mapping the Scattered Field of Research on Higher Education”, *High Educ*, vol. 80, pp. 571–587, 2020.

[17] T. P. Dybowski and P. Adämmer, “The Economic Effects of U.S. Presedental Tax Commucation: Evidence From A Correlated Topic Model”, *European Journal of Political Economy*, vol. 55, pp. 511-525, 2018.

[18] H. Tu, L. Xia & Z. Wang, “The Complex Action Recognition via Correlated Topic Model”, *Scientific World Journal*, vol. 2014, 2014.

[19] M. Aznag, M. Quafafou and Z. Jarir, “Correlated Topic Model For Web Services Ranking”, *International Journal of Advanced Computer Science and Applications*, vol. 4, pp. 283-291, 2013.