Journal of Engineering Research and Applied Science
Available at www.journaleras.com
Volume 13 (1), June 2024, pp 2490-2499
ISSN 2147-3471 © 2024

Comparison of Traditional and Modern Topic Model Algorithms in Terms of Topic Determination in Official Documents

Zeynep Bozdogan[1,*], Resul Kara[2]

[1] Duzce University, Graduate School of Education, 81620, Duzce, Turkey
ORCID: https://orcid.org/0009-0007-6028-6951
[2] Duzce University, Faculty of Engineering, Computer Engineering, Duzce, Turkey
ORCID: https://orcid.org/0000-0001-8902-6837

Accepted 15 December 2023

Abstract
The rapid increase in the number of documents in the digital environment makes analysis and control of documents difficult. To overcome this complexity, separating and classifying digital documents according to certain criteria is becoming more and more important day by day. In order to classify documents effectively, various methods include innovative techniques such as machine learning, deep learning and topic modeling. In this study, LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) and NMF (Non-Negative Matrix Factorization) algorithms, which are widely used in the literature, are used in official documents. Performance comparison was made in terms of topic determination. It was observed that the NMF algorithm gave the most successful results with -5.217 in terms of c_umass metric and 88.1% in terms of correct classification.

Keywords*: LDA, LSA, NMF, Topic Modeling, Standard File Plan Codes with Retention Period*

1. Introduction
Text mining, a sub-branch of data mining, is a method that allows extracting information from texts. By applying this method, natural language processing methods and data mining techniques are used to discover information from texts in accordance with the structure of the text language on texts written in unstructured natural language. Text mining is used for various purposes such as text summarization, finding the subject of the text, text classification, text clustering, keyword extraction, sentiment analysis. Text Mining is the process of obtaining previously unknown, potentially useful, structured and organized data from unstructured and disorganized masses of electronic text. With the information obtained, relationships, hypotheses and trends that are not evident in the analyzed text sources are detected. Although text mining is considered a part of data mining, it is different from traditional data mining. The main difference is that text mining extracts patterns from natural language texts rather than event-based databases. In the simplest sense, text mining studies are data mining studies that accept texts as data sources and aim to obtain structured data from texts. For example; text classification (taxonomy), clustering, entity extraction, detailed taxonomy generation, sentiment analysis, document summarization, entity relationship modeling and topic modeling [1]. Topic Modeling (TM) is one of the subfields of text mining. Its main purpose is to reveal hidden or open issues in documents. It has an important place among the subfields of text mining, especially with its increasing importance and studies in recent years.

Within the scope of this study, a new data set created from non-confidential (confidential, top secret, personal, service specific) documents within a public institution electronic document management system was examined. This data set includes a total of 2100 Turkish official correspondence documents from three separate categories, including 700 documents belonging to 051, 105, 106 Standard File Plan (SSDP) Codes.

The content texts of the documents are currently determined by the person who created the document, which SSDP code will be assigned. The main purpose

*Corresponding author: zeynepbozdogan@duzce.edu.tr

of this study is to use topic model algorithms to automatically detect SSDP codes according to the contents of documents. In the study, LDA, LSA and NMF topic model algorithms were applied for classification operations performed with Python codes.

The classification performance results of the topic model algorithms applied on the data set developed within the scope of the study were compared and analyzed. It has been observed that the most effective results in terms of classification processes are obtained using the NMF algorithm.

In the continuation of the study, Chapter 2 includes studies in the literature on the subject, Chapter 3 includes information on the materials and methods selected and applied, Chapter 4 includes the findings and discussions of the study, and Chapter 5 includes results and recommendations.

2. Related studies

Topic model algorithms have been frequently used in classification processes to distinguish data in recent years. Topic modeling algorithms are unsupervised algorithms that enable given documents to be automatically classified into classes. Topic Model algorithms are text classification [2], text summarization [3], [4], keyword discovery [5], [6] etc. can be used for other purposes.

Tolegen et al. In their study [7] We applied text pre-processing operations on the "20newsgroups" data set created from the news contents, and then LSA, LDA, PLSA, BigARTM algorithms were applied on the data set by giving the number of topics as 10, 20, 30, 40, 50 on these texts and the performance of the applied algorithms was evaluated. Coherence score values were obtained to measure. As the number of subjects increases, the coherence score (cv) values for the LDA algorithm are between 0.48-0.34, for pLSA 0.68-0.57, for LDA 0.56-0.66, for ARTM 0.57-0.69, for the cooc method mentioned in the study 0.46-0.42, and for the tifidf mentioned in the study 0.64-0.64. It was observed that decreasing values were obtained between 0.56 and increasing values were obtained between 0.37 and 0.57 for the w2v method mentioned in the study.

In another study [8], Haghighi et al. Using 896,867 cleaned tweets about low back pain between January 1, 2014 and December 31, 2018, they applied text pre-processing operations on the data set they obtained and performed LDA, Dirichlet multinomial mixture (DMM), GPU-DMM, Biterm Topic Model (BTM) on the obtained data. NMF topic model algorithms were applied and coherence score values were used to measure the performance results of the algorithms. The best coherence score value was obtained as 0.562 when the data in the data set was divided into 60 topics using the LDA algorithm.

Aydın and Hallaç in their studies [9] Text pre-processing processes were applied on the data set created from current Turkish news, which includes 4 categories: economy, sports, politics and culture, and 100,000 samples for each category, 400,000 in total. Afterwards, 400 new news items that were not previously included in the training data set were given and tested for testing purposes as training data. The classification results applied on the data set with the LDA algorithm were evaluated with two different performance methods proposed in the study. The first and second methods achieved 94.2% and 90.9% accuracy, respectively.

Güven et al. [10] used a total of 4200 news data sets from 7 classes, 600 from each class, in their study. Text pre-processing operations were applied on the news contents in the data set and then LDA, LSA, NMF, n-GDA topic model algorithms were applied on the pre-processed texts and the results were compared. The applied LDA algorithm was found to reach the highest accuracy with 82.1% and 71.5% for five and seven classes, respectively. When only the number of classes was given as 3, the NMF algorithm was more successful with 95%. In all methods, it was observed that the least successful method was classical LDA, and the most successful was LSA. The n-GDA method mentioned in the study was modeled in 2 stages in the study, and the 2-GDA method provided an accuracy increase of approximately 4% to 9% for all classes compared to the classical GDA method.

In their study by Kaya and Gülbandıldı [1], Robert et al. LDA and Correlated Topic Models (CTM), Structural Topic Model (STM) algorithms were applied on the "Poliblogs5k" data set, which is the data set used in the study [11], using the methods and variables used in [11]. For all algorithms, the number of topics was used as 20 and the number of words was 10. To evaluate the results of the algorithms, comparisons were made with coherence score, perplexity and running times. The highest coherence score value was found to be 0.509 with the Structural Topic Model (STM) type 3 method specified in the study.

Güven et al. In their study [12], Tweets on the data set created from a total of 4000 data samples of 800

tweets obtained from Twitter for each emotion of 5 different emotions, including "angry, fear, happy, sad and surprised", were manually labeled and 80% of the data set was determined. 20% was used for training and 20% for testing purposes. Two different data sets with 3 and 5 classes were used. After pre-processing processes were applied on the texts, the performance results of the n-stage LDA method proposed in the study were compared with the LDA, LSA and Probabilistic-Latent Semantic Analysis (P-LSA) algorithms. As a result of the operations applied on data sets with 3 and 5 subjects, the classification success of the 2-stage proposed n-stage LDA method increased by 10-15% compared to the classical LDA method, and the 3-stage one increased by 11-21% compared to the classical LDA method. A lower performance result was obtained than the performance result of the method. In terms of running time, the LSA method was determined to be the fastest working method. It was observed that the 3-stage LDA method worked in a time close to the LSA method, and the p-LSA method was the slowest.

Altintas et al. In their study [13], text pre-processing was applied on the data set consisting of 109,243 user comments obtained from posts about cancer disease on the social sharing platform "reddit", and LDA, LSA and NMF topic model algorithms were applied on the pre-processed comments. When LDA, LSA and NMF topic model algorithms were applied by giving the number of topics as 5 in the algorithms, the most successful coherence score results were generally obtained with the LDA method. With different parameter values given to the LDA method, consistency score results between 0.498 and 0.538 were obtained when the number of subjects was 5. Since the most successful coherence score results were obtained with the LDA method, operations such as finding keywords on the comments and word cloud analysis were applied using the LDA method in the continuation of the study.

Preetham MC et al. In their study [14], numerical word vectors of the texts were created after text pre-processing operations were applied on the data set created with 1740 scientific research articles obtained from the New York City University website. In the study, it was aimed to classify documents with LDA and NMF algorithms and the performances of both algorithms were compared. The average coherence score value for the LDA algorithm was found to be 0.5282. The average coherence score value for the NMF algorithm was found to be 0.4937, the optimum number of topics was determined to be 9. With the LDA algorithm, the highest coherence score value was obtained as 0.55821 when the number of topics was given as 22. With the NMF algorithm, the highest coherence score value was obtained as 0.52012 when the number of subjects was given as 9. Generally, more successful results were obtained with LDA than with NMF.

There are various studies in the literature on the automatic assignment of SSDP codes of documents for classification processes. In the study conducted by Binici [15], a data set consisting of 169 documents was created, 112 of the created data set were used for training and 57 for testing purposes. By applying Support Vector Machine (SVM), one of the machine learning algorithms, on the data set, 87.72% performance was achieved as a result of the classification operations. In our previous study on text mining [16], official correspondence documents were classified by applying Logistic Regression, Neural Network, SVM, Tree, Random Forest, Naive Bayes, kNN algorithms on the data set in the study to automatically classify SSDP codes.

3. Materials and methods

3.1. Dataset and preprocessing

In this section, the data set features used in the study and the text preprocessing steps applied on the data set are mentioned.

3.1.1. Data set

In the study, a data set obtained from a total of 2100 documents with anonymized content within the scope of KVKK, which does not have any confidentiality, including 700 for each code with 051, 105, 106 SSDP Codes, was used as the data set in a corporate Electronic Document Management System (EBYS). The created data set consists of three main columns;

- Document Content text: Documents that do not contain personal information or confidentiality
- Current SSDP Code of the Document: It represents the SSDP Code that the people who created the document wrote into the document as SSDP code information when creating the document.
- Edited SSDP Code of the Document: It represents a new column in which it is manually checked whether the correct SSDP code has been written in the document by the people who created the document, according to the content text of the document, and the incorrectly written SSDP code is corrected and added.

The standard file plan is a structure consisting of certain code numbers, subject headings, sub-headings, retention periods and storage codes specified for each code, under the same heading, for documents within the scope of related subjects. It was created to provide

a common standard filing and archiving system in public institutions and organizations [17], [18]. SSDP codes and names of the documents used in the data set are shown in Table 1

Table 1. SSDP codes and names in the data set created in the study

SSDP Master code	First Name
051	Scientific and Cultural Meetings
105	Course Schedules
106	Exam Schedules

3.1.2. Pre-Processes applied on the data set

Text preprocessing operations were applied sequentially on the data sets in the following order. The reason why the stopwords removal process was applied again in the 9th step was that after the root of the words remaining from the previous stopwords removal step was obtained, the stopwords could be found again in the word roots.

1. Converting all characters to lowercase: It is the process of converting all characters in the text to lowercase.
2. Removing punctuation and special characters: Removing punctuation and special characters from the text. This step removes punctuation marks and special characters such as commas, periods, and exclamation points in the text.
3. Cleaning numbers and figures: It is the process of removing the numbers and figures in the text from the texts.
4. Cleaning extra spaces: Removes more than one space character from the text .
4. Cleaning HTML tags: It is the process of cleaning the information of HTML tags in the text.
5. Url information cleaning: It is the process of cleaning the URL (Uniform Resource Locator) information in the text.
6. Removing Turkish stopwords: Stopwords are words commonly used in the language that do not generally carry meaning, and the stopwords list published by Aksoy and Öztürk was used as the stopword word list in the study [19].
7. Removing the words in the newly created list of ineffective words: The words in the list of words that are not distinctive in determining the SSDP code in the expressions that appear routinely and repeatedly in the documents in the data set have been removed from the text.
8. Obtaining the roots of words: The roots of the words were obtained using the Zemberek library [20] , [21] .
9. Removing Turkish stopwords: Stopwords are words that are commonly used in the language and are generally considered unimportant in text analysis processes.
10. Removing the words in the newly created list of ineffective words: The words in the list of words created from words that do not have any distinctiveness in determining the SSDP code in the expressions that appear routinely and repeatedly in the documents in the data set have been removed from the text.

The list specified in articles 7 and 10 is composed of words that are not distinctive in determining the SSDP code in the expressions that appear routinely and repeatedly in the documents in the data set. For example, "I request for your information", "I request for your information what is necessary", "I request for your information what is necessary" in the official documents. Expressions such as are standard expression patterns that should be included in official documents in accordance with the status of the document. They were removed from the document content texts as it was thought that they would not have any distinguishing effect in distinguishing the SSDP code of the document.

In order to apply topic model algorithms on the data sets used, the texts must be converted to digital format. For this, Term Frequency-Inverse Document Frequency (TF-IDF) digitization method was used.

3.2. Topic model algorithms

In the study, LDA, LSA, NMF algorithms, which are topic model algorithms, were used to classify the texts in the data sets. This part of the study includes information about LDA, LSA, NMF algorithms.

3.2.1. LDA algorithm

LDA algorithm is one of the most widely used topic model algorithms in the literature and Blei et al. It was developed in 2003 by [22]. This statistical model is an unsupervised topic model algorithm that aims to separate documents into different topic groups based on the weights of words within the documents [23], [24], [25], [26]. This algorithm is an unsupervised topic model algorithm that divides documents into topics depending on the weight of the words that make up the documents. The fact that the algorithm is unsupervised means that manual inferences can be made in the process of determining which topic groups

produced represent which topic groups represent [24], [27]

LDA enables parsing documents into a certain number of subjects or topics. These topics are represented by a model in which the topics within each document are distributed in a certain proportion, and each topic is distributed in a certain proportion of words. These distributions allow for a clearer understanding of the meaning contained in documents. The basic logic of LDA is based on the assumption that the topics contained in the documents and the words in each topic have a certain structure. In this way, each document consists of a specific topic group, and the semantic structure of the documents is revealed by determining the relationships between these topic groups [25] , [26], [28] .

3.2.2. LSA algorithm

Deerwester et al. It was first proposed by in 1990. It is also known as LSI and was developed to analyze hidden semantic relationships in documents [29]. LSA is a topic model algorithm designed to make sense of the semantic structures contained in documents. This algorithm discovers semantic connections between words contained in a document and represents these connections through basic concepts, namely "topics". Essentially, LSA performs a mathematical abstraction process that represents the content of documents in a more meaningful way and makes it possible to group documents with similar meaning. In this way, it becomes possible to reveal hidden meaning relationships in large document collections and perform semantic analysis. The working logic of the LSA algorithm is to convert documents into a document-word matrix, analyze the frequency of words in documents with the TF-IDF method, and use the Singular Value Decomposition (SVD) method. It is based on applying noise reduction and dimension reduction on the matrix and thus revealing important information representing the documents [30], [31].

3.2.3. NMF algorithm

The NMF algorithm is a concept introduced by Paatero and Tapper in 1994 and refers to an algebraic model that transforms the matrices created by documents into lower-dimensional matrices containing non-negative values [32]. This algorithm works based on non-negative factorization of matrices. While non-negative factorization refers to the factorization of a matrix, NMF imposes a restriction that each element of this decomposition has non-negative values. This constraint enables the use of positive elements to ensure the representation of documents, allowing a more meaningful extraction of the topics and features they contain. The main purpose of NMF is to reveal hidden structures in document collections by separating the data matrix into its non-negative factors. This algorithm is an effective tool in determining the topics and features of documents and finds a wide range of applications, especially in areas such as text mining, knowledge discovery and content analysis. Additionally, the non-negative constraints of NMF make the information extraction processes more meaningful and useful, ensuring that the resulting factors are more interpretable and more suitable for real-world applications. In this context, the NMF algorithm has an important place in information extraction tasks such as document classification and topic modeling [33].

3.3. Evaluation methods

The most common method that can be used to evaluate the performance results of Topic Model algorithms is to manually infer the topics that the keywords in the topics can represent and, based on this inference, to check whether the correct topics are assigned to the documents. Apart from this, there are some evaluation metrics in the literature. These are coherence score [34] and perplexity [35] values.

In the study, the performance of topic model algorithms implemented using the u_mass sub-approach of coherence score, one of the evaluation methods mentioned above, was evaluated. In addition, the performance of the results was interpreted based on the keywords obtained manually for each topic.

3.3.1. Coherence score (theme coherence)

Coherence Score is an evaluation method used to measure topic consistency as a result of the application of topic model algorithms. There are different sub-approaches for the Coherence Score value: c_v, c_p, c_umass c_one-any, c_uci, c_npmi, c_a values [1], [36]. Among these approaches, the most frequently used in studies in the literature are c_v and c_umass values.

The formula of the coherence score c_umass metric value used to evaluate the results in the study is stated in Equation 1 [7], [37].

$$Coherence\ Score\ (umass) = \frac{2}{N.(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \tag{1}$$

Here " $P(w_i)$" refers to the number of documents containing the word w_i, " $P(w_i, w_j)$The expression " represents the number of documents containing both

w_i and w_j. The term ε is added to avoid log(0) evaluation. Closer to 0 of the coherence score c_umass metric value indicates higher topic coherence [38].

3.3.2. Perplexity

It is a metric that measures how successfully a topic model predicts the topics of new data. It may give misleading results because it cannot find the relationship between the topics or the words in the topic, and may not have a sufficient indicator value for evaluation when used alone [1], [39].

A lower perplexity value indicates that the model is more successful in explaining the analyzed documents [33].

The formula for the perplexity value is stated in Equation 2 [40].

$$PPL(LM, W) = exp\{-\frac{1}{k}\sum_{i=1}^{k} log\ log\ LM(w_i|w_{1:i-1})\ \} \quad (2)$$

Here, $log\ log\ LM(w_i|w_{1:i-1})$, i. is the log-likelihood value of the symbol based on previous symbols $.w_i|w_{1:i-1}$

4. Findings and discussions

The data obtained as a result of the application of LDA, LSA, NMF topic model algorithms on the data set created within the scope of the study are mentioned in this section.

When running each of the Topic Model algorithms, the value of the topic number parameter is given as the number of classes of data in the data sets. As a result of the work of the algorithms, it is aimed to divide the texts in the data set into as many classes as the given number of classes.

Table 2 , Table 3 and Table 4 include the keywords obtained for each topic information as a result of applying the LDA, LSA and NMF topic model algorithms, respectively, on the data set used in the study.

Table 2. The top 5 words detected in the topics as a result of the application of the LDA Topic Model algorithm to the data set.

LDA					
	word 0	Word 1	word 2	word 3	word 4
Topic 0	effective	workshop	university	conference	faculty
Topic 1	lesson	program	teaching	departmen t	branch
Topic 2	exam	program	section	year	period

Table 3. The top 5 words detected in the topics as a result of the application of the LSA Topic Model algorithm to the data set.

LSA					
	word 0	word 1	word 2	word 3	word 4
Topic 0	exam	program	period	year	lesson
Topic 1	exam	spring	Search	prepare	final
Topic 2	effective	conference	workshop	plan	exam

Table 4. The top 5 words detected in the topics as a result of the application of the NMF Topic Model algorithm to the data set.

NMF					
	word 0	word 1	word 2	word 3	word 4
Topic 0	exam	program	section	year	period
Topic 1	lesson	department	branch	teaching	program
Topic 2	effective	conference	faculty	plan	workshop

Table 2 , Table 3 and Table 4 are examined, it can be deduced which class each topic refers to from the top words representing the class. For example; When Table 4 is examined, it can be concluded that Topic 2, which is the topic containing the words "active", "workshop", "conference", "plan", "faculty", can refer to the class coded 051 (scientific and cultural meetings). It can be concluded that Topic 1, which is the topic containing the words "course", "program", "instruction", "department", "branch", can refer to the SSDP coded class 105 (course programs). It can be concluded that Topic 0, which is the topic containing the words "course", "program", "instruction", "department", "branch", can refer to the SSDP coded class 106 (exam programs).

In Table 5 , Table 6 and Table 7 , as a result of applying the LDA, LSA and NMF topic model algorithms, respectively, on the data set used in the study, how many of each topic class are calculated as a result of applying the topic model algorithms for each original class information (SSDP Code) given at the beginning. The number and rate of predictions are shown. For example, in Table 5 , as a result of the application of the LDA topic algorithm on the data set, the class of 569 documents out of a total of 700, whose SSDP code was initially given as "105" by the people who created the document, was classified as "Topic 1".

The "Original Class" column in Tables 5,6,7 contains the SSDP Code information written on the document by the users who created the document in the data set. The "Predicted Class" column contains the SSDP Code information predicted for the documents as a result of the application of the topic model algorithm on the data set. The "Number of Rows" column indicates the number of predictions made as a result of the topic model algorithms for each SSDP code, and the "Row Rate" column indicates the ratio.

Table 5 , Table 6 and Table 7 are examined, it can be deduced which of the original class labels each predicted topic information corresponds to. For example; In Table 7 , out of a total of 700 documents whose SSDP code was given as "051" by the people who created the document, 690 (98.571%) were classified as "Topic 2" as a result of the NMF algorithm, 9 of them (1.286 %) were classified as "Topic 1" and 1 of them was classified as "Topic 1" as a result of the NMF algorithm. (0.143 %) is classified as “Topic 0”. Based on this, since the largest number of classification predictions were made as "Topic 2", it can be deduced that the class expressed as "Topic 2" as a result of the NMF algorithm corresponds to the "051" SSDP code class label. Since the highest number of "Topic 1" classification predictions were made for 700 documents with the SSDP code numbered "105", it can be deduced that the class expressed as "Topic1" as a result of the NMF algorithm corresponds to the "105" SSDP code class label. Since the highest number of "Topic 0" classification predictions were made for 700 documents with the SSDP code numbered "106", it can be deduced that the class expressed as "Topic 0" as a result of the NMF algorithm corresponds to the "106" SSDP code class label. In line with the same logic, in Table 5, since the SSDP code of 569 (81.286%) of a total of 700 documents whose SSDP code was given as "105" was estimated as "Topic 1" as a result of the application of the LDA algorithm, the SSDP Code numbered 105 of "Topic 1" was used as SSDP. Since the SSDP code of 563 (80.429%) of a total of 700 documents whose code was given as "106" was estimated as "Topic 2" as a result of the application of the LDA algorithm, the SSDP Code of "Topic 2" was determined as 105, and the SSDP code numbered 105 was used for a total of 700 documents whose SSDP code was given as "051". Since the SSDP code of 619 (88.429%) of the documents was estimated as "Topic 0" as a result of applying the LDA algorithm, it can be concluded that "Topic 0" corresponds to the SSDP Code numbered 051.

According to the results in Table 6, although the number of classes was given to the algorithm as 3 for the classification process of the LSA algorithm, it was seen that the documents were divided into only 2 classes as a result of the data classification process. It is thought that this may be due to the fact that the documents belonging to the codes "105" and "106" have semantically similar contents to each other.

Table 5 , Table 6 and Table 7 , it can be said that as a result of the application of LDA and NMF algorithms on the data set, the most successful classification process for both algorithms was achieved for the "051" code. It has been observed that the most successful results for all algorithms are generally achieved with the NMF algorithm.

According to the Coherence Score (c_umass) values in Table 8 , as a result of applying topic model algorithms on the data set, it was seen that the most successful topic classification result was the NMF algorithm.

Table 5. Numbers and ratios determined according to the original class and predicted class as a result of the application of the LDA Topic Model algorithm to the data set.

LDA			
Original Grade	Predicted Class	Number of Rows	Row Rate (%)
105	Topic 0	84	12
105	Topic 1	569	81,286
105	Topic 2	47	6,714
106	Topic 0	41	5,857
106	Topic 1	96	13,714
106	Topic 2	563	80,429
051	Topic 0	619	88,429
051	Topic 1	58	8,286
051	Topic 2	23	3,286

Table 6. Numbers and ratios determined according to the original class and predicted class as a result of the application of the LSA Topic Model algorithm to the data set.

LSA			
Original Grade	Predicted Class	Number of Rows	Row Rate (%)
105	Topic 0	698	99,714
105	Topic 2	2	0.286
106	Topic 0	697	99,571
106	Topic 2	3	0.429
051	Topic 0	115	16,429
051	Topic 2	585	83,571

Table 7. Numbers and ratios determined according to the original class and predicted class as a result of the application of the NMF Topic Model algorithm to the data set.

NMF			
Original Grade	Predicted Class	Number of Rows	Row Rate (%)
105	Topic 0	53	7,571
105	Topic 1	610	87,143
105	Topic 2	37	5,286
106	Topic 0	551	78,714
106	Topic 1	109	15,571
106	Topic 2	40	5,714
051	Topic 0	one	0.143
051	Topic 1	9	1,286
051	Topic 2	690	98,571

Table 8. Coherence score (c_umass) values obtained as a result of applying Topic Model algorithms on the data set.

Topic Model Algorithm	Coherence Score (c_umass)
LDA	-6.444
LSA	-10,908
NMF	-5.217

5. Conclusions and recommendations

The aim of the study is to classify official documents according to their SSDP code using topic model algorithms. LDA, LSA and NMF algorithms were applied on the data set created from 2100 documents obtained from a corporate electronic document management system, and the official documents in the data set were classified into classes.

As a result of the classification process of documents with topic model algorithms, it has been observed that the NMF algorithm provides the most successful classification according to the c_umass metric for the data set. According to the study, 1851 of 2100 documents were classified into correct topics and had a classification success of 88.1% .

As future works; By increasing the number of documents in the data set and the SSDP main codes and sub-codes of the documents, topic determination can be made with traditional topic model algorithms and modern topic model algorithms.

References

[1] A. KAYA and E. GÜLBANDILAR, "Comparison of Topic Modeling Methods," *Eskişehir Türk Dünyası Uygul. and Research Center. Informatics Journal.* , vol. 3, no. 2, pp. 46–53, May 2022, doi: 10.53608/estudambilisim.1097978.

[2] Z. Li, W. Shang, and M. Yan, "News text classification model based on topic model," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* , Jun. 2016, pp. 1–5, doi: 10.1109/ICIS.2016.7550929.

[3] N. Akhtar, MMS Beg, and H. Javed, "TextRank enhanced Topic Model for Query focused Text Summarization," in *2019 Twelfth International Conference on Contemporary Computing (IC3)* , Aug. 2019, pp. 1–6, doi: 10.1109/IC3.2019.8844939.

[4] DFO Onah, ELL Pang, and M. El-Haj, "A Data-driven Latent Semantic Analysis for Automatic Text Summarization using LDA Topic Modelling," in *2022 IEEE International Conference on Big Data (Big Data)* , Dec. 2022, pp. 2771–2780, doi: 10.1109/BigData55660.2022.10020259.

[5] X. Wang, L. Zhang, and D. Klabjan, "Keyword-based Topic Modeling and Keyword Selection," in *2021 IEEE International Conference on Big Data (Big Data)* , Dec. 2021, pp. 1148–1154, doi: 10.1109/BigData52589.2021.9671416.

[6] S. Xu, J. Guo, and X. Chen, "Extracting topic keywords from Sina Weibo text sets," in *2016 International Conference on Audio, Language and Image Processing (ICALIP)* , Jul. 2016, pp. 668–673, doi: 10.1109/ICALIP.2016.7846663.

[7] G. Tolegen, A. Toleu, R. Mussabayev, and A. Krassovitskiy, "A Clustering-based Approach for Topic Modeling via Word Network Analysis," in *2022 7th International Conference on Computer Science and Engineering (UBMK)* , Sep . 2022, pp. 192–197, doi: 10.1109/UBMK55850.2022.9919530.

[8] P. Delir Haghighi, F. Burstein, D. Urquhart, and F. Cicuttini, "Investigating Individuals' Perceptions Regarding the Context Around the Low Back Pain Experience: Topic Modeling Analysis of Twitter Data," *J. Med. Internet Pic.* , vol. 23, no. 12, p. e26093, Dec. 2021, doi: 10.2196/26093.

[9] G. AYDIN and İ. HALLAÇ, "Automatic Topic Detection in Turkish Texts," *Fırat University Engineering Science. Journal.* , vol. 33, no. 2, pp. 599–606, Sep. 2021, doi: 10.35234/fumbd.899917.

[10] ZA Guven, B. Diri, and T. Cakaloglu, "Comparison of Topic Modeling Methods for Type Detection of Turkish News," in *2019 4th International Conference on Computer Science and Engineering (UBMK)* , Sep. 2019, pp. 150–154, doi: 10.1109/UBMK.2019.8907050.

[11] ME Roberts, BM Stewart, and D. Tingley, "stm : An R Package for Structural Topic Models," *J. Stat. Softw.* , vol. 91, no. 2, 2019, doi: 10.18637/jss.v091.i02.

[12] ZA GÜVEN, B. DİRİ, and T. ÇAKALOĞLU, "Comparison of n-stage Latent Dirichlet Discrimination and other topic modeling methods for sentiment analysis," *Gazi University Engineering Architect. Faculty Journal.* , vol. 35, no. 4, pp. 2135–2146, Jul. 2020, doi: 10.17341/gazimmfd.556104.

[13] V. ALTINTAŞ, M. ALBAYRAK, and K. TOPAL, "Hidden topic modeling with Dirichlet separation for posts about cancer disease," *Gazi University Engineering Architect. Faculty Journal.* , vol. 36, no. 4, pp. 2183–2196, Sep. 2021, doi: 10.17341/gazimmfd.734730.

[14] SP MC, BR Reddy, DS Tharun Reddy, and D. Gupta, "Comparative Analysis of Research Papers Categorization Using LDA and NMF Approaches," in *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)* , Nov. 2022, pp. 1–7, doi: 10.1109/NKCon56289.2022.10127059.

[15] K. BİNİCİ, "A Study on Automatic Assignment of Standard File Plan Numbers to e-Documents with a Machine Learning Approach," *Information Management* , vol. 2, no. 2, pp. 116–126, Dec. 2019, doi: 10.33721/by.654464.

[16] Z. Bozdoğan and R. Kara, “Subject Detection in Official Correspondence Using Text Mining.”
[17] N. ÇİÇEK, “The Power of Function in File Classification Plans,” in *Information Management in a Changing World Symposium* , 2007, pp. 235–244.
[18] “TC Council of Higher Education, Higher Education Institutions and Higher Education Institutions Standard File Plan with Retention Period.” https://www.yok.gov.tr/Documents/Universiteler/Standart_Dosya_Plani.pdf (accessed Oct. 18, 2023).
[19] A. Aksoy and T. Öztürk, “Turkish Stop Words Turkish Filler Words.” https://github.com/ahmetax/trstop.
[20] AA Akın and MD Akın, “Zemberek, an open source NLP framework for Turkic Languages,” *Structure* , vol. 10, pp. 1–5, 2007.
[21] “Zemberek Library.” https://github.com/ahmetaa/zemberek-nlp.
[22] DM Blei, AY Ng, and MI Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Pic.* , vol. 3, no. null, pp. 993–1022, Mar. 2003.
[23] I. Vayansky and SAP Kumar, “A review of topic modeling methods,” *Inf. Syst.* , vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.
[24] E. Ekinci, “A Unique Topic Modeling Method Based on Semantic Similarities of Documents,” Kocaeli University, 2019.
[25] B. ÇULLU and A. OKURSOY, “Investigating the Service Quality of Cargo Companies with Text Mining,” *Anadolu Üniversitesi Sos. Science. Journal.* , vol. 23, no. 2, pp. 399–422, Jul. 2023, doi: 10.18037/ausbd.1205507.
[26] I. AlAgha, “Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives,” *J. Inf. Sci. THEORY PRACTICE.* , vol. 9, no. 1, 2021, doi: https://doi.org/10.1633/JISTaP.2021.9.1.3.
[27] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” in *Proceedings of the 22nd International Conference on Neural Information Processing Systems* , 2009, pp. 288–296.
[28] S. Ozturk *et al.* , “Turkish labeled text corpus,” in *2014 22nd Signal Processing and Communications Applications Conference (SIU)* , Apr. 2014, pp. 1395–1398, doi: 10.1109/SIU.2014.6830499.
[29] S. Deerwester, ST Dumais, GW Furnas, TK Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.* , vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
[30] Y. Li and B. Shen, “Research on sentiment analysis of microblogging based on LSA and TF-IDF,” in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)* , Dec. 2017, pp. 2584–2588, doi: 10.1109/CompComm.2017.8323002.
[31] J. Zeng *et al.* , “Statutes Recommendation Based on Text Similarity,” in *2017 14th Web Information Systems and Applications Conference (WISA)* , Nov. 2017, pp. 201–204, doi: 10.1109/WISA.2017.52.
[32] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics* , vol. 5, no. 2, pp. 111–126, Jun. 1994, doi: 10.1002/env.3170050203.
[33] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, “Topic modeling algorithms and applications: A survey,” *Inf. Syst.* , vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.
[34] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing Semantic Coherence in Topic Models,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* , 2011, pp. . 262–272, [Online]. Available: https://aclanthology.org/D11-1024.
[35] F. Jelinek, RL Mercer, LR Bahl, and J. K. Baker, “Perplexity—a measure of the difficulty of speech recognition tasks,” *J. Acoust. Soc. Am.* , vol. 62, no. S1, pp. S63–S63, Dec. 1977, doi: 10.1121/1.2016299.
[36] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* , Feb. 2015, pp. 399–408, doi: 10.1145/2684822.2685324.
[37] J.M. Campagnolo, D. Duarte, and G. Dal Bianco, “Topic Coherence Metrics: How Sensitive Are They?,” *J. Inf. Data Manag.* , vol. 13, no. 4, Oct. 2022, doi: 10.5753/jidm.2022.2181.
[38] P. Tijare and P. Jhansi Rani, “Exploring popular topic models,” *J. Phys. Conf. Ser.* , vol. 1706, no. 1, p. 012171, Dec. 2020, doi: 10.1088/1742-6596/1706/1/012171.
[39] P. Dasgupta, J. Amin, C. Paris, and C. R. MacIntyre, “News Coverage of Face Masks in Australia During the Early COVID-19 Pandemic: Topic Modeling Study,” *JMIR Infodemiology* , vol. 3, p. e43011, Aug. 2023, doi: 10.2196/43011.
[40] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and DP Radicioni, “Semantic coherence markers: The contribution of perplexity metrics,” *Artif. Intel. Med.* , vol. 134, p. 102393, Dec. 2022, doi: 10.1016/j.artmed.2022.102393.