





RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Performance Comparison of Text Weighting Schemas on NMF-Based Topic Analysis

Metin Ağırlıklandırma Şemalarının NMF-Tabanlı Konu Analizi Alanında Başarımlarının Karşılaştırılması

Tolga Berber ^{*}, Melek Eriş Büyükkaya 

Karadeniz Teknik Üniversitesi, Fen Fakültesi, Bilgisayar Bilimleri Bölümü, Trabzon, TÜRKİYE
Corresponding Author / Sorumlu Yazar*: tolga.berber@fen.ktu.edu.tr

Abstract

Nowadays, it is feasible to analyze text data that is being generated at an exponential rate by transforming it into a sparse matrix of big size using a certain weighting method. A comprehensive text weighting approach consists of three fundamental components: Term Frequency, Document Frequency, and Vector Normalization. The multiplication of these three components yields numerical values that indicate the significance of a word for a text. Nevertheless, the unprocessed state of these values is unsuitable for the semantic analysis of textual material. There are multiple techniques available for this objective, and Topic Analysis, which seeks to identify subjects discussed in extensive text collections, is one of these techniques. The Non-Negative Matrix Factorization (NMF) approach is commonly employed in topic analysis. It involves transforming an input matrix into the product of two or more matrices, using both random and deterministic beginning values. This study involved conducting tests on a dataset of 20,000 articles sourced from Wikipedia, the online encyclopedia, with the aim of investigating the impact of text weighting methods and initial value approaches commonly employed in the literature on the NMF method. The number of clusters to be used in the studies was determined using an analytical procedure, which employed an upper limit. The results indicate that the "lnc" and "nnc" weighting schemes yielded the highest performance in NMF. These findings demonstrate that employing the "lnc" or "nnc" weighting scheme will lead to more favorable outcomes in the domain of topic analysis.

Keywords: Topic Analysis, Text-Weighting Schemas, Non-negative Matrix Factorization, Performance Comparison

Öz

Günümüzde üstel bir şekilde üretilen metin verisinin analiz edilebilmesi, bu verinin belirli bir ağırlıklandırma yaklaşımı ile büyük boyutlu seyrek bir matrise çevrilmesi ile mümkün olmaktadır. İdeal bir metin ağırlıklandırma yaklaşımının 3 temel bileşeni bulunmakta olup; bunlar Terim Frekansı, Doküman Frekansı ve Vektör Normalizasyonu bileşenleridir. Bu üç bileşenin çarpımı ile bir kelimenin bir metin için önemini ifade eden sayısal değerler elde edilir. Ancak bu değerlerin ham hali metin verisinin anlamsal olarak analiz edilmesi için uygun değildir. Bu amaçla çeşitli yöntemler bulunmakta olup, büyük metin koleksiyonlarında bahsedilen konuları bulmayı amaçlayan Konu Analizi bu yöntemlerden bir tanesidir. Bu amaçla konu analizinde bir girdi matrisini hem rastgele hem de deterministik başlangıç değeri ile iki veya daha fazla matrisin çarpımına dönüştürmeyi hedefleyen Negatif Olmayan Matris Ayrışımı (NMF) yönteminden sıklıkla faydalanılır. Bu çalışmada, literatürde kullanılan metin ağırlıklandırma yöntemlerinin ve başlangıç değer yaklaşımlarının NMF yöntemi üzerinde etkilerinin bulunması amacıyla, Vikipedi özgür internet ansiklopedisinden elde edilen 20.000 makale üzerinde denemeler yapılmıştır. Denemelerde kullanılacak küme sayısının elde edilmesi için analitik bir yöntem yardımıyla bir üst sınır kullanılmıştır. Elde edilen sonuçlara göre, NMF üzerinde en iyi başarıma "lnc" ve "nnc" ağırlıklandırma şemalarıyla ulaşılmıştır. Buda konu analizi alanında "lnc" veya "nnc" ağırlıklandırma şemalarının kullanılmasıyla daha başarılı sonuçlar elde edileceğini göstermiştir.

Anahtar Kelimeler: Konu Analizi, Metin Ağırlıklandırma Şemaları, Negatif Olmayan Matris Ayrışımı, Başarım Karşılaştırması

1. Introduction

Currently, owing to the increasing prevalence of Internet technologies and devices that are able to support them, the most commonly generated form of data is textual data. According to a survey conducted in 2022, around five billion individuals globally and around 63 million individuals in Turkey had at least one account on a social media platform. These individuals dedicate an average of 2 hours and 31 minutes per day to generate textual material on social media [1]. The content generated exclusively on social media platforms has generated a significant increase in

data, even on a daily basis. Hence, it is challenging to extract important information from datasets with exceedingly large volumes. In other words, it is feasible to analyze and interpret extensive data by substantially reducing the volume. This allows the implementation of many applications. Various applications, including market research, product feedback, the study of social perception, and prompt handling of complaints and requests, can be effortlessly achieved.

Topic analysis is a technique devised to extract concise and significant information from vast amounts of textual data [2], [3].

Table 1. Sample view from dataset.

Article Id	Url	Title	Text
12	https://en.wikipedia.org/wiki/Anarchism	Anarchism	Anarchism is a political philosophy and movement that is sceptical of authority and rejects all involuntary, coercive forms of hierarchy. Anarchism calls for the abolition of the state, which it holds to be unnecessary, undesirable, and harmful. As a historically...
7676	https://en.wikipedia.org/wiki/Creaky%20voice	Creaky voice	In linguistics, creaky voice (sometimes called laryngealisation, pulse phonation, vocal fry, or glottal fry) is a special kind of phonation in which the arytenoid cartilages in the larynx are drawn together; as a result, the vocal folds are compressed rather tightly, becoming relatively slack and compact. They normally...
30266	https://en.wikipedia.org/wiki/Tonyukuk	Tonyukuk	Tonyukuk (, , , born c. 646, died c. 726) was the бага-tarkhan (supreme commander) and adviser of four successive Göktürk khagans – Elteriš Qayan, Qapyan Qayan, İnäl Qayan and Bilgä Qayan. He conducted victorious...

Initially, probabilistic models were commonly employed, although more deterministic approaches are currently being integrated into the methodology. This enables a rapid and effortless overview of textual information. Numerous studies have been published on this topic [4], [5].

Non-negative matrix factorization (NMF) is a method designed to decompose a matrix into a product of two smaller matrices. This facilitates the analysis of the acquired matrices with reduced dimensions. Given this framework, NMF is an iterative method for topic analysis [5]. Despite the deterministic structure of the NMF, the initial values are typically chosen randomly. This reduced the reliability of the method. An approach called non-negative double-singular (NDSVD) was proposed as a solution to this problem [6]. This approach utilizes singular value decomposition (SVD). For relevant studies, see, for example, [7], [8], [9].

The purpose of utilizing NMF in the domain of topic analysis is to reduce the dimensions of the matrix arising from the process of converting text data into a numeric format. Although NMF is a reliable and efficient technique, its effectiveness is closely tied to the semantic content of the matrix obtained from textual input. Choosing an appropriate strategy will enhance the quality and precision of research, given the various focusing approaches available during the digitization phase of text data [10], [11], [12].

The main contribution of this study is to provide an objective assessment of several text weighting schemas using the NMF topic Modelling approach. Because text weighting is one of the most important factors that directly influences the results of topic modelling, choosing an optimal weighting method will help researchers establish a good starting point. Therefore, a series of experiments were conducted to measure the achievements in the topic analysis of the text weighting approaches defined in the literature, as part of this study. The results were compared in terms of coherence by using the NMF method. The subsequent sections of this paper are structured as follows. second section of the paper presents the dataset and methodologies employed. The third section of the manuscript discusses the experimental findings. In conclusion, the results acquired in the fourth section are consolidated, and suggestions for further research are presented.

2. Related Works

The use of NMF has been an active research area because of its analytical nature. In particular, it has been employed in several academic disciplines such as environmental sciences, medical informatics, and text mining [13]. For example, the NMF method has been used for genetic information extraction [14], [15], [16], [17], [18], to make inferences about textual medical data [19], [20], [21], [22], and pollution discrimination [23], [24], [25].

From a topic analysis perspective, several variants of NMF have been heavily employed in these studies. For instance, NMF and its variants have been used to extract useful insights from tourism sector reviews [26], [27]; health-related textual information [20], [21], [22], [28], [29]; low-source language texts [30], [31]; and accident or disaster analysis from textual data [32], [33]. However, most studies have integrated library-provided functions to preprocess text-weighting schemas. Because weighting is a key factor in topic analysis, using the most accurate text-weighting method is important. Hence, this study aims to close this gap in the literature by providing a comprehensive evaluation of well-known text weighting schemas.

3. Materials and Methods

3.1. Dataset

Studies in the field of topic analysis take advantage of the large amounts of text. The study utilized a version of the Wikipedia Internet Encyclopedia that was generated by the Huggingface group and dated March 1, 2022.

Wikipedia backups its database in XML format for a period of six months and opens it to researchers. However, these data are very difficult to process because they are compressed to 19 GB (open to 86 GB). Huggingface periodically retrieves textual data from Wikipedia and provides it in a JSON-like format (the whole export file is an invalid JSON, but each line is JSON). The study utilized a dataset consisting of 6.458.670 English-language articles that were processed and transformed into JSON format by Huggingface on March 1, 2022 (downloaded from <https://huggingface.co/datasets/wikipedia/tree/main/data/20220301.en>). We randomly selected 20.000 documents randomly from the Wikipedia Dataset to reduce computation requirements. The random 3 samples of the selected dataset are listed in Table 1.

3.2. Text Preprocessing

To analyze the obtained Wikipedia articles, all words must be represented numerically; hence, pre-processing steps were applied to the obtained data. For the normalization of the articles obtained in the first stage, all characters were translated into lower-case letters. A total of 416.450.007 characters were processed during this phase. Subsequently, characters were eliminated using unicode (universal-coded characters) categories. The KC Normal Form (NFKC), previously proposed by the Unicode consortium, was used at this stage. Then, the characters other than the Unicode category small letter (Ll) or large letter (Lu) and space (Zs) were excluded from the text data, and a total of 392.586.768 characters remained as a result. The resulting documents were split by space character, and a total of 64.535.969 text parts were obtained. Suffixes were removed from the text using the Snowball [34] stemmer, and English stop words were removed from the text. A total of 38.972.617 words

were obtained as a result of this phase. There are 516.119 words in our final dictionary. The resulting dictionary digitized 20,000 Wikipedia documents, and the final dimension of the document-term matrix was 20.000×516.119 .

3.3. Text-Weighting Methods

The raw state of text data is difficult to analyze owing to its various linguistic and contextual characteristics. Therefore, the use of digitized text data is preferred over raw data. The most preferred method for digitizing text data in this area is the Bag of Words (BoW). Using this method, text data are generally translated into large matrices that have a sparse structure and show the relationship between words and documents. The methodologies employed in this approach were determined using the SMART Information Retrieval System created by Cornell University. Within this method, the importance of every word in the documentation comprises of three distinct components. These components consist of the Term Frequency Component (TFC), which quantifies the importance of a word in a document; the document frequency component (DFC), which measures the relevance of a term across the entire text data; and the (VNC), which indicates the process of vector normalization. The final word weight is given by the following equation obtained from the product of these three components:

$$X_{t,d} = TFB_{t,d} \times DFB_t \times VNB \quad (1)$$

Table 2 lists the predefined weighting schemes for the three components used by the SMART information-delivery system.

Table 2. SMART Weighting Components.

TFC		DFC		VNC	
n	$tf_{t,d}$	n	1	n	1
l	$\begin{cases} 1 + \log(tf_{t,d}), & tf_{t,d} > 0 \\ 0, & tf_{t,d} \leq 0 \end{cases}$	t	$\log \frac{N_d}{df_t}$	c	$\frac{1}{\sqrt{\sum_{i=1}^n w_i^2}}$
a	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p	$\max\left\{0, \log \frac{N_d - df_t}{df_t}\right\}$	u	$\frac{1}{u}$
b	$\begin{cases} 1, & tf_{t,d} > 0 \\ 0, & tf_{t,d} \leq 0 \end{cases}$			b	$\frac{1}{L_c^\alpha}, \alpha < 1$
L	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_t(tf_{t,d}))}$				

where $tf_{t,d}$ indicates the frequency of the word t within the document d ; N_d is the total number of documents; df_t is the number of different documents that include the word t ; w_i is the element i . of a vector; u is a normalization factor; L_c^α is the average number of words in the text collection; and α is a standardization factor.

This study uses all combinations defined in the SMART notation. However, owing to additional parameters, we excluded “u” and “b” values of VNC.

3.4. Topic Analysis

Topic analysis approaches, which have recently been developed to find topics discussed in large-volume text data, often use reduced matrices through dimension reduction methods. The most preferred dimensional reduction methods for these approaches are the Latent Dirichlet Allocation (LDA) [35] and the NMF [36]. The LDA method is a generative probability model. In this model, two different multivariate probability distributions were used. One of these distributions models the likelihood that any word passes a topic, whereas the other models the

probability that any text belongs to a topic. The process of calculating LDA aims to obtain the probability values initially randomly selected, iteratively approaching the actual probability value. Therefore, there may be differences in the results obtained from the LDA calculation. However, the classic NMF method aims to convert a large matrix into the product of two smaller matrices. The structure of this method also involves an iterative calculation process, and the starting point is usually chosen randomly. Unlike LDA, it has been demonstrated that an effective and deterministic starting point can be determined using the classic NMF method, thereby increasing the consistency of the results obtained [6]. Therefore, the success of topic analysis was measured using only the NMF method.

3.5. Non-negative Matrix Factorization (NMF)

In the classical NMF method, the input matrix is divided into two different matrices using a specific approach. A matrix display separated by the classical NMF is

$$X = WH \quad (2)$$

where $X \in \mathbb{R}^{m \times n}$ and $x_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n$ indicate the input matrix, $W \in \mathbb{R}^{m \times r}$ and $w_{ik} \geq 0, 1 \leq i \leq m, 1 \leq k \leq r$ the mixture matrix (base), $H \in \mathbb{R}^{r \times n}$ and $h_{kj} \geq 0, 1 \leq k \leq r, 1 \leq j \leq n$ the coding matrix and the r parameter also the difference size [37].

The classical NMF method aims to find the $\tilde{W} \approx W$ and $\tilde{H} \approx H$ separation matrices of the \tilde{X} matrix, which is a prediction of the X matrix. In this case, the differential to be obtained for the matrix \tilde{X} is

$$\tilde{X} = \tilde{W}\tilde{H} \quad (3)$$

where $\tilde{W} \in \mathbb{R}^{m \times r}$ and $\tilde{H} \in \mathbb{R}^{r \times n}$ are matrices.

NMF is an optimization method. This method uses the objective function of minimizing a certain $D(\cdot, \cdot)$ distance measurement between the matrix \tilde{X} and the input matrix X . This is shown in Equations (4) and (5):

$$\min_{\tilde{X} \in \mathbb{R}^{m \times n}} D(X, \tilde{X}) = \min_{\tilde{W} \in \mathbb{R}^{m \times r}, \tilde{H} \in \mathbb{R}^{r \times n}} D(X, \tilde{W}\tilde{H}), \quad (4)$$

$$D(X, \tilde{W}\tilde{H}) = \frac{1}{2} \|X - \tilde{W}\tilde{H}\|_F^2 = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n ((X)_{ij} - (\tilde{W}\tilde{H})_{ij})^2. \quad (5)$$

Here, $D: \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is an error function, and the Frobenius norm is indicated by the notation $\|\cdot\|_F^2$.

The r separation dimension is an important parameter in the NMF method. However, the failure to calculate this parameter in advance and to experimentally determine it prolongs the completion time of the work. For example, in most NMF applications, calculation of the appropriate r value is typically performed experimentally and within a wide range of values (approximately 140). In this context, the brute-force approach is used to select the one with the best success value among all r values in the value range. This experimental approach is expressed as

$$\operatorname{argmax}_{r_{\min} < r < r_{\max}} P(X, \tilde{W}\tilde{H}) \quad (6)$$

where r_{\min}, r_{\max} represents a value in the range $r \in \mathbb{N}^+$, and large values of the $P: \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ function indicate high achievement. Researchers can also use natural numbers that have a specific pattern because they can use all natural numbers in this value range. However, r_{\min}, r_{\max} values do not have an analytical approach to determine, and these values are determined based on the experience of researchers. This is a factor that negatively affects the coherence of work with NMF.

As a general rule, the upper limit of the r parameter is expected to comply with

$$r \ll \min(m, n) \quad (7)$$

rule. Here, the expression \ll indicates a radically small r parameter value, whose upper limit is $\min(m, n)$, must be used.

Because the NMF method is a low-rank matrix approximation, the element counts of the \tilde{W} and \tilde{H} separation matrices obtained by this method are expected to be less than the element count of the X input matrix. Therefore, given the number of elements ($mr + nr$) of the separated matrices to be obtained with classical NMF, the following inequality must be satisfied:

$$mr + nr < mn. \quad (8)$$

Thus, the upper limit of the separation size is

$$r < \frac{mn}{m+n}. \quad (9)$$

If the input matrix is sparse, the upper limit is

$$r < \frac{nnz(X)}{m+n}. \quad (10)$$

Here $nnz: \mathbb{R}^{m \times n} \rightarrow \mathbb{N}^+$ denotes the number of elements in a matrix that are different from zero. These two inequalities produce values that are more suitable for estimating the upper boundary of parameter r than inequality (7). In this study, the highest r value that yields the inequality (10) was used [38].

The proposed calculation approaches for the classic NMF method iteratively minimize the objective function in Eq. (5). The most preferred of these approaches is the Frobenius rule's Multiplicative Update Rules (MUR) [36]. In addition to this approach, the steepest gradient descent and Newton-type numerical methods have been proposed [39], [40]. In our study, MUR was preferred because of its widespread use, and this method is shown in Equations (11)–(12).

$$G_w(\tilde{W}^t, \tilde{H}^t) = \tilde{W}^t \otimes (X \tilde{H}^t \oslash \tilde{W}^t \tilde{H}^t \tilde{H}^t), \quad (11)$$

$$G_h(\tilde{W}^{t+1}, \tilde{H}^t) = \tilde{H}^t \otimes (\tilde{W}^{t+1} X \oslash \tilde{W}^{t+1} \tilde{W}^{t+1} \tilde{H}^t). \quad (12)$$

Here, \otimes is the Hadamard product; \oslash is the Hadamard division; and, G_w and G_h are the functions that update the \tilde{W} and \tilde{H} matrices, respectively.

In all these calculation methods, the starting values of the \tilde{W} and \tilde{H} matrices were randomly selected from the uniform distribution. This negatively affects the consistency of the \tilde{W} and \tilde{H} matrices obtained with NMF applications [41], [42]. In addition, these randomly selected starting values also adversely affect the calculation time of the NMF method [43].

3.6. Non-Negative Double Singular Value Decomposition (NDSVD)

For the classic NMF method to produce more consistent and faster results, a deterministic starting value approach for \tilde{W} and \tilde{H} , called the NNDSVD, has been suggested and successful results have been obtained [6]. Algorithm 1 presents the algorithm for the NNDSVD method.

Algorithm 1. NNDSVD Algorithm

Input: $X \in \mathbb{R}_+^{m \times n}$, $r < \min(m, n) \in \mathbb{N}^+$

Output: $\tilde{W}^0 \in \mathbb{R}_+^{m \times r}$, $\tilde{H}^0 \in \mathbb{R}_+^{r \times n}$

- 1 Calculate the single value of the largest r of $X: U, S, V = svd(X, r)$
- 2 Specify as $\tilde{W}_{:,1} = \sqrt{s_{1,1}} \times U_{:,1}$ and $\tilde{H}_{1,:} = \sqrt{s_{1,1}} \times V_{1,:}$
- 3 For $\forall j \in [2, r]$
- 3.1 $x = U_{:,j}$ ve $y = V_{:,j}$
- 3.2 $x_+ = \begin{cases} x_i, & x_i > 0 \\ 0, & \text{otherwise} \end{cases}$, $x_- = \begin{cases} -x_i, & x_i < 0 \\ 0, & \text{otherwise} \end{cases}$, $y_+ = \begin{cases} y_i, & y_i > 0 \\ 0, & \text{otherwise} \end{cases}$ ve $y_- = \begin{cases} -y_i, & y_i < 0 \\ 0, & \text{otherwise} \end{cases}$
- 3.3 $\mu_+ = \|x_+\| \|y_+\|$ ve $\mu_- = \|x_-\| \|y_-\|$
- 3.4 $u = \begin{cases} \frac{x_+}{\|x_+\|}, & \mu_+ > \mu_- \\ \frac{x_-}{\|x_-\|}, & \mu_- \leq \mu_+ \end{cases}$, $v = \begin{cases} \frac{y_+}{\|y_+\|}, & \mu_+ > \mu_- \\ \frac{y_-}{\|y_-\|}, & \mu_- \leq \mu_+ \end{cases}$ and $\sigma = \max(\mu_+, \mu_-)$
- 3.5 $\tilde{W}_{:,j} = \sqrt{s_{j,j}} \times \sigma \times u$ and $\tilde{H}_{j,:} = \sqrt{s_{j,j}} \times \sigma \times v'$

With this algorithm, it has been demonstrated that the \tilde{W} and \tilde{H} matrices have non-negative values and decrease the value of the goal function without becoming stuck in local minima [6]. The classical NMF method used in this approach has been confirmed by several studies in which calculations are closer to conclusions in a shorter time and produce consistent results [7], [8], [9].

3.7. Computational Complexity

Owing to the nature of NMF, most algorithms developed to determine the exact factors are NP-Hard. These algorithms, which can be solved in polynomial time, impose certain restrictions on r . In this case the running time complexity of NMF algorithms $O((mn)^{cr^2})$ for a constant c , which is a grow exponentially for large values of r [38]. Hence, selecting the most suitable text-weighting scheme is important to reduce the computational requirements of the approach.

3.8. Topic Coherence

Using the results obtained from topic analysis, the maximum weight for each topic was determined using $M = 5, \dots, 20$ words. There are many methods in the literature that measure the coherence of topics obtained. In these methods, topic coherence is calculated based on a combination of words obtained. The most popular of these methods is called U_{mass} , and the equation is shown in Equation (13) [44].

$$U_{mass} = \sum_{0 < i < j \leq M} \mathcal{S}(w_i, w_j) = \sum_{0 < i < j \leq M} \log \frac{\mathcal{D}(w_i, w_j) + 1}{\mathcal{D}(w_i)} \quad (13)$$

Here, $\mathcal{D}(w_i)$ represents the total number of documents passed by the word w_i , and $\mathcal{D}(w_i, w_j)$ represents the total number of documents passed by the words w_i and w_j together. After calculating the U_{mass} value for all topics obtained, the average of all U_{mass} values are taken so that the overall evaluation of the subject analysis results can be made. This is expressed in Equation (14).

$$C_{U_{mass}} = \frac{1}{r} \sum_{i=1}^r U_{mass_i} \quad (14)$$

where $C_{U_{mass}}$ denotes the overall coherence value of the object analysis results, r denotes the number of objects, and U_{mass_i} denotes the U_{mass} value of i . the topic.

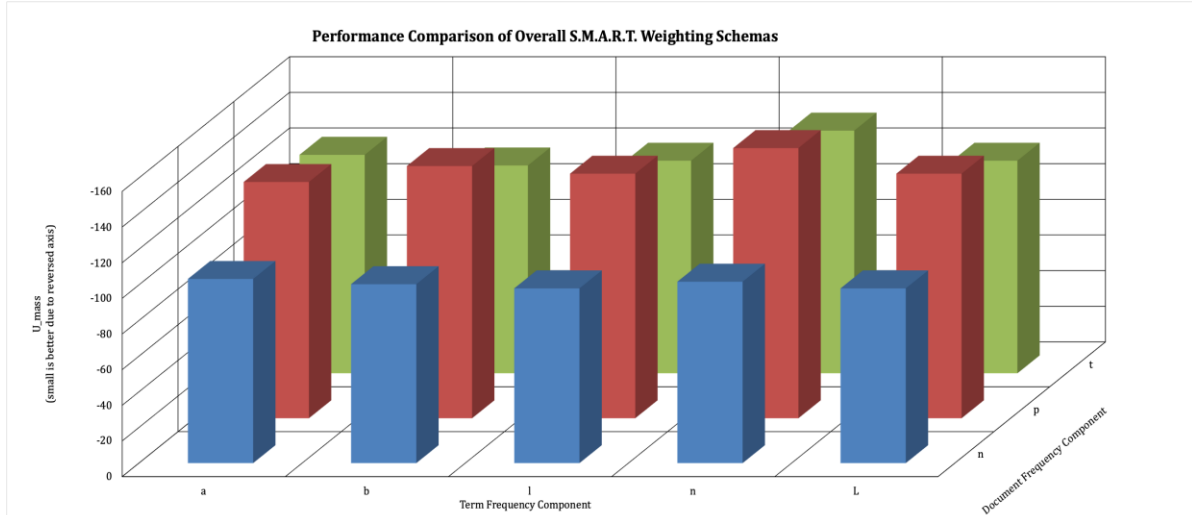


Figure 1. Overall Performances of SMART Component Values.

In the end, the obtaining of large $C_{U_{mass}}$ values indicates that the analysis results have a high consistency.

4. Experimental Results

We conducted a series of experiments on the Wikipedia dataset to determine the most effective combination of the SMART components. Although each SMART component has a different impact on generating document-term matrices, they generated document-term matrices with the same number of non-zero elements in our experiments. The overall properties of the obtained document term matrices are listed in Table 3.

Table 3. General Properties of Document-Term Matrix.

Total Number of Documents	6458670
Total Number of Selected Documents	20000
Total Number of Unique Tokens	516119
Total Number of Non-zero Entries in Each Document-Term Matrix	14579095
Sparsity of Each Document Term Matrix	%0.1412

We used these properties to determine the upper limit of topic count. Using equation (10), the upper limit of the topic count was calculated as 27 for our experimental setup.

Because the normalization components u and b require additional parameter estimations, we exclusively used the c -component. In addition, our experimental configuration excludes the n component of term normalization to avoid generating document vectors that are not normalized. In total, we conducted 15 experiments to assess the effectiveness of each combination of word frequency (n, l, a, b, L) and document frequency (n, t, p) components in Topic Analysis, using the Wikipedia dataset.

After the creation of document term matrices for each SMART component, we calculated the W and H matrices using NMF on the GPU code implemented using the CuPy library. We implemented the NNDSVD algorithm to determine the initial values of W and H matrices. We ran all our models on an NVIDIA Quadro P4000 GPU with 8 GB of VRAM. The peak memory usage in each experiment was approximately 300 MB.

The performance of each topic-analysis result was evaluated using U_{mass} coherence measure. We took the average of all topic coherence values to provide an overall comparison between the SMART components. The average U_{mass} values for each pair are presented in Table 4 and Figure 1.

Table 4. Coherence Values of Each Combination.

	n	p	t
a	-103.31	-132.53	-122.59
b	-100.37	-141.47	-116.60
l	-98.01	-137.18	-119.28
n	-101.77	-151.55	-136.15
L	-98.01	-137.18	-119.28

The findings demonstrate that the “n” document frequency component consistently attained the highest coherence scores, irrespective of the TFC while the “p” document frequency component obtained the worst coherence values. Nevertheless, we were unable to determine the particular TFC that achieved the highest coherence scores across all document frequency components. The TFCs “l” and “L” exhibit the highest coherence scores when combined with the document frequency component “n”. The “a” TFC has the highest coherence score with the “p” document frequency component, whereas the “b” TFC has the highest coherence score with the “t” document frequency component. The TFCs “l” and “L” achieved the same values for all experimental setups. In summary, the best configuration of the SMART components for the Wikipedia dataset is the “lnc” configuration.

Because topic analysis is not only a field of study for mathematics but also a field of textual data summarization, coherence scores are insufficient to present textual concepts. To overcome this insufficiency, many researchers have attempted to name these topics. The topic-naming process involves two stages: finding the most effective topic terms and generating a title for the found terms. The first stage is straightforward, but the second stage requires domain expertise and is generally subject to subjective assessment. In our experimental setup, we included a well-known generative artificial intelligence (AI) system to remove subjective naming assessment [45]. Hence, we generated titles objectively. The prompts used to generate the topic titles are given in Table 5.

Table 5. ChatGPT Title Generation Prompt.

Propose a single topic header for the following words:
<comma joined the most effective topic words>

We generated a total number of 405 topic titles using ChatGPT with a given prompt. Some titles produced were the same. For example, our prompt generated “*Japanese History and Culture*” topic titles for 12 of the topics of different SMART components.

Table 6. The First 40 Topic Titles and Their Statistical Properties.

Topic Titles	Frequency	Avg. Coherence Score
Japanese History and Culture	12	-22.23
International Organizations	6	-33.25
Key Terms and Concepts in American Football	6	-64.14
Mathematical Concepts and Terminology	5	-82.30
Various Professions	4	-77.49
Exploring Computer Hardware and Software Components	4	-116.63
Exploring Demographic Statistics and Trends: A Comprehensive Analysis	4	-117.45
Topics in Abstract Algebra	4	-131.25
Exploring Telecommunication Technologies and Services	4	-132.60
Roman Empire and its Key Figures	4	-145.04
Exploring Subatomic Particles and Fundamental Forces in Particle Physics	3	-78.21
International Organizations and Agencies	3	-91.57
Chemical Elements and Compounds	3	-95.82
Topics in Linguistics	3	-104.11
Mathematical Concepts and Structures	3	-114.38
Exploring the Fundamentals of Particle Physics and Quantum Mechanics	3	-116.99
Key Components and Processes in Cellular Biology	3	-119.15
Countries and Territories	3	-129.36
Medieval Empires and Kingdoms	3	-133.63
Logical Reasoning and Mathematical Concepts	3	-144.98
Roman History and Figures	3	-145.84
Medieval Kingdoms and Empires	3	-147.85
Exploring the Roman Empire: History, Calendar, and Cultural Influence	2	-43.63
Exploring Key Terms and Concepts in American Football	2	-53.00
Education and its Impact on Society	2	-56.18
Exploring the Elements: Actinides, Transuranium Elements, and their Properties	2	-61.14
Various Professions and Nationalities	2	-70.64
Economic Indicators and Policies	2	-74.01
Titles, Figures, and Institutions in Historical European Context	2	-78.07
Economic Development and Policies in the European Union: A Comprehensive Analysis	2	-78.99
Key Terms in American Football	2	-80.85
Government and Politics	2	-82.07
Geography and Climate Features	2	-87.18
Exploring the World of Polyhedra and Geometric Shapes	2	-87.41
Key Concepts in Economic Policy and Governance	2	-93.74
Exploring the World of Computing: Software, Hardware, Interfaces, and More	2	-93.82
Phonetics and Phonology: An Exploration of Speech Sounds and Language	2	-95.69
Exploring the World of Music and Entertainment	2	-96.52
Chemical Reactions and Compounds	2	-96.91
Exploring the World of Music: Albums, Instruments, Studios, and More	2	-97.15

The first 40 topics, sorted by coherence scores, are listed in Table 6. According to the results, the majority of the selected Wikipedia titles are related to historical figures and cultural heritage.

Although our approach for generating topic titles produced unbiased titles, some of the generated titles were insufficient because of the large range of topic terms generated. For example, one of the topic of SMART configuration “Ltc” produced “disambigu, album, unincorpor, surnam, township, indiana, iowa, mathemat, kentucky, uss, softwar, nebraska, missouri, wisconsin, cuisin, ohio, ix, counti, band, entertain” terms. The generated topic titles for these terms produce 15 different topic titles, as listed in Table 7.

Table 7. Proposed Topic Titles for Example.

Proposed Topic Titles
1. Disambiguation and its importance in information retrieval
2. The significance and impact of albums in the music industry
3. Unincorporated areas and their role in local governance
4. The history and significance of surnames in genealogy and identity
5. Township governance and its role in local administration
6. Exploring the states of Indiana, Iowa, Nebraska, Missouri, and Wisconsin
7. The beauty and complexity of mathematics
8. Kentucky's rich history and cultural heritage
9. The role and impact of the United States Navy's USS (United States Ship) vessels
10. The evolution and importance of software in modern society
11. Exploring the diverse landscapes and attractions of Ohio
12. Understanding the Roman numeral system and its applications
13. County governance and its impact on local communities
14. The power of music and the role of bands in entertainment
15. Exploring the world of entertainment and its various forms

The main cause of this issue is the contextual diversity of the terms that describe the topic. Put simply, the topic encompasses a wide range of themes that are too diverse to be consolidated under a single title. We named this topic as *The Null Topic (TNT)* and used the coherence value of this topic as a threshold to discriminate high quality topics from low quality ones.

According to the TNT Coherence score, the distribution of high-quality topics by DFC (n, p, and t) and TFC (a, b, l, L, and n) of the SMART system is given in Table 8.

Table 8. Distribution of High Quality Topics

	n	p	t
a	25	22	24
b	25	20	24
l	26	21	23
L	26	21	23
n	27	18	24

As a result we could say the most of the high quality topic titles are generated from “n” DFC component. Moreover, all the topics generated by the “n” TFC component in combination with the “n” DFC component were high-quality topics.

5. Conclusions

We conducted a series of experiments to determine the best combination of SMART system components for NMF-based topic analysis tasks. Because the SMART system is still widely used in Topic Analysis tasks, researchers must select the most suitable SMART components to make inferences about textual data. In this work, we have found that the “lnc” and “nnc” SMART configurations work best in terms of Topic Coherence and Topic Quality, respectively. This provides a good starting point for any Topic Analysis task and decreases the time required.

Another contribution of this study is the integration of a generative AI tool into a Topic Analysis task to overcome the subjectivity of topic title generation. We designed a prompt to generate a single topic title for topic terms. This approach provides another opportunity to measure topic quality.

As a side product, we propose a new method to assess topic quality using an automatic coherence thresholding method. This method uses the coherence values of AI-generated titles with more than one alternative. Hence, our approach constructs a basis for the automated objective evaluation of Topic Analysis in terms of topic quality. However, this approach needs to be supported by further analysis.

Ethics committee approval and conflict of interest statement

This study did not require approval from the ethics committee. This article has no conflicts of interest with any individual or institution.

Author Contribution Statement

The first Author contributed to the formulation and implementation of the data analysis techniques, writing of the manuscript, and interpretation of the results. The second author contributed to the literature survey, theoretical foundations, and the writing and interpretation of the results.

References

- [1] “Reports & Content — Kepios.” Accessed: Aug. 25, 2023. [Online]. Available: <https://kepios.com/reports>.
- [2] Vayansky, I., Kumar, S.A.P., 2020. A review of topic modeling methods. *Information Systems*, Vol. 94, p. 101582. DOI: 10.1016/J.IS.2020.101582.
- [3] Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, Vol. 55, No. 4, pp. 77–84. DOI: 10.1145/2133806.2133826.
- [4] Schachtner, R., 2010. Extensions of Non-negative Matrix Factorization and Their Application to the Analysis of Wafer Test Data. PhD Thesis, Universität Regensburg, Regensburg.
- [5] Shen, J., Israël, G.W., 1989. A receptor model using a specific non-negative transformation technique for ambient aerosol. *Atmospheric Environment*, Vol. 23, No. 10, pp. 2289–2298. DOI: 10.1016/0004-6981(89)90190-X.
- [6] Boutsidis, C., Gallopoulos, E., 2008. SVD-based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, Vol. 41, No. 4, pp. 1350–1362. DOI: 10.1016/J.PATCOG.2007.09.010.
- [7] Yamashita, A., Nagata, T., Yagyu, S., Asahi, T., Chikyow, T., 2022. Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis. *Manufacturing Letters*, Vol. 2, No. 1, pp. 23–37. DOI: 10.1080/27660400.2022.2029222.
- [8] Wang, Z., Yu, Y., 2022. Revealing the spatial and temporal distribution of different chemical states of lithium by EELS analysis using non-negative matrix factorization. *Micron*, Vol. 154, p. 103213. DOI: 10.1016/J.MICRON.2022.103213.
- [9] Lu, H., Zhao, Q., Sang, X., Lu, J., 2020. Community Detection in Complex Networks Using Nonnegative Matrix Factorization and Density-Based Clustering Algorithm. *Neural Processing Letters*, Vol. 51, No. 2, pp. 1731–1748. DOI: 10.1007/S11063-019-10170-1.
- [10] Wang, J., Zhang, X.L., 2023. Deep NMF topic modeling. *Neurocomputing*, Vol. 515, pp. 157–173. DOI: 10.1016/J.NEUCOM.2022.10.002.
- [11] Habbat, N., Anoun, H., Hassouni, L., 2021. Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets. *Lecture Notes in Networks and Systems*, Vol. 183, pp. 147–161. DOI: 10.1007/978-3-030-66840-2_12.
- [12] Egger, R., Yu, J., 2022. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, Vol. 7, p. 886498. DOI: 10.3389/FSOC.2022.886498.
- [13] Guo, Y.-T., Li, Q.-Q., Liang, C.-S., 2024. The rise of nonnegative matrix factorization: Algorithms and applications. *Information Systems*, Vol. 123, p. 102379. DOI: 10.1016/j.is.2024.102379.
- [14] Takasawa, K., et al., 2024. Advances in cancer DNA methylation analysis with methPLIER: Use of non-negative matrix factorization and knowledge-based constraints to enhance biological interpretability. *Experimental & Molecular Medicine*. DOI: 10.1038/s12276-024-01173-7.
- [15] Chen, D., et al., 2024. Comprehensive analyses of solute carrier family members identify SLC12A2 as a novel therapy target for colorectal cancer. *Scientific Reports*, Vol. 14, No. 1, p. 4459. DOI: 10.1038/s41598-024-55048-y.
- [16] Dey, A., Das Sharma, K., Bhattacharjee, P., Chatterjee, A., 2024. Identification of disease-related biomarkers in time-varying ‘Omic data: A non-negative matrix factorization aided multi-level self-organizing map based approach. *Biomedical Signal Processing and Control*, Vol. 90, p. 105860. DOI: 10.1016/j.bspc.2023.105860.
- [17] Shi, Y., Jin, Z., Deng, J., Zeng, W., Zhou, L., 2024. A novel high-dimensional kernel joint non-negative matrix factorization with multimodal information for lung cancer study. *IEEE Journal of Biomedical and Health Informatics*, Vol. 28, No. 2, pp. 976–987. DOI: 10.1109/JBHI.2023.3335950.
- [18] Ramamoorthy, T., Kulothungan, V., Mappillairaju, B., 2024. Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Frontiers in Artificial Intelligence*, Vol. 7, p. 1329185. DOI: 10.3389/frai.2024.1329185.
- [19] Subbarayudu, Y., Sureshbabu, A., 2024. The detection of community health surveillance using distributed semantic-assisted non-negative matrix factorization on topic modeling through sentiment analysis. *Multimedia Tools and Applications*. DOI: 10.1007/s11042-024-18321-w.
- [20] Choi, D., et al., 2023. WellFactor: Patient Profiling using Integrative Embedding of Healthcare Data. In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*, IEEE, pp. 616–625. DOI: 10.1109/BigData59044.2023.10386138.
- [21] Ahammad, T., 2024. Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. *Natural Language Processing Journal*, Vol. 6, p. 100053. DOI: 10.1016/j.nlp.2024.100053.
- [22] Zong, L., Yang, Y., Xia, H., Yuan, J., Guo, M., 2023. Elucidating the Impacts of Various Atmospheric Ventilation Conditions on Local and Transboundary Ozone Pollution Patterns: A Case Study of Beijing, China. *Journal of Geophysical Research: Atmospheres*, Vol. 128, No. 20. DOI: 10.1029/2023JD039141.
- [23] Knobel, P., et al., 2023. Socioeconomic and racial disparities in source-apportioned PM_{2.5} levels across urban areas in the contiguous US, 2010. *Atmospheric Environment*, Vol. 303, p. 119753. DOI: 10.1016/j.atmosenv.2023.119753.
- [24] Westervelt, D.M., et al., 2024. Low-Cost Investigation into Sources of PM_{2.5} in Kinshasa, Democratic Republic of the Congo. *ACS ES&T Air*, Vol. 1, No. 1, pp. 43–51. DOI: 10.1021/acsestair.3c00024.

- [25] Karamouzi, E., Pontiki, M., Krasonikolakis, Y., 2024. Historical Portrayal of Greek Tourism through Topic Modeling on International Newspapers. In Proceedings, pp. 121–132. Accessed: Mar. 30, 2024. [Online]. Available: <https://aclanthology.org/2024.latechclfl-1.13>.
- [26] Athurugiriya, A.A.A.G., Sumathipala, P., Hemachandra, K.M.T.A., 2023. Development of an Enhanced Quality Score Calculation Method for Accurate Assessment of Hotel Quality. In Proceedings of the 2023 5th International Conference on Advancements in Computing (ICAC), IEEE, pp. 804–809. DOI: 10.1109/ICAC60630.2023.10417334.
- [27] Mahmoudi, L., Hossein Shari, M., Bagheri, R., 2024. Exploring Healthcare Research Patterns in Developed and Developing Countries: A Topic Modeling Perspectives. DOI: 10.21203/RS.3.RS-3865906/V1.
- [28] Hornback, A., et al., 2023. Latent Topic Extraction as a Source of Labeling in Natural Language Processing. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 4312–4319. DOI: 10.1109/BIBM58861.2023.10385618.
- [29] Sweidan, A.H., El-Bendary, N., Elhariri, E., 2024. Autoregressive Feature Extraction with Topic Modeling for Aspect-based Sentiment Analysis of Arabic as a Low-resource Language. ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 23, No. 2, pp. 1–18. DOI: 10.1145/3638050.
- [30] Pallawala, D., Haddela, P.S., 2023. A Comparison of Topic Modeling Techniques for Sinhala. In Proceedings of the 2023 5th International Conference on Advancements in Computing (ICAC), IEEE, pp. 376–381. DOI: 10.1109/ICAC60630.2023.10417327.
- [31] Nanyonga, A., Wasswa, H., Wild, G., 2023. Topic Modeling Analysis of Aviation Accident Reports: A Comparative Study between LDA and NMF Models. In Proceedings of the 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), 2023. DOI: 10.1109/SMARTGENCON60755.2023.10442471.
- [32] Ghaly, M.Z., Laksito, A.D., 2023. Topic Modeling of Natural Disaster in Indonesia Using NMF. In Proceedings of the 2023 8th International Conference on Informatics and Computing (ICIC), IEEE. DOI: 10.1109/ICIC60109.2023.10382064.
- [33] Porter, M.F., 1980. An algorithm for suffix stripping. Program, Vol. 14, No. 3, pp. 130–137. DOI: 10.1108/EB046814.
- [34] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
- [35] Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature, Vol. 401, No. 6755, pp. 788–791. DOI: 10.1038/44565.
- [36] Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemometrics and Intelligent Laboratory Systems, Vol. 37, No. 1, pp. 23–35. DOI: 10.1016/S0169-7439(96)00044-5.
- [37] Gillis, N., 2020. Nonnegative Matrix Factorization. Philadelphia, PA: Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611976410.
- [38] Guillamet, D., Vitri, J., 2002. Non-negative Matrix Factorization for Face Recognition. In Proceedings of the Fifth Catalanian Conference on Artificial Intelligence, Castellon, Spain, pp. 336–344.
- [39] Kim, D., Sra, S., Dhillon, I.S., 2007. Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem. In Proceedings of the Sixth SIAM Conference on Data Mining, Minnesota, USA, pp. 343–354.
- [40] Hofmann, T., 2017. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 211–218.
- [41] Kim, H., Park, H., 2008. Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. SIAM Journal on Matrix Analysis and Applications, Vol. 30, No. 2, pp. 713–730. DOI: 10.1137/07069239X.
- [42] Belford, M., Mac Namee, B., Greene, D., 2018. Stability of topic modeling via matrix factorization. Expert Systems with Applications, Vol. 91, pp. 159–169. DOI: 10.1016/J.ESWA.2017.08.047.
- [43] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing Semantic Coherence in Topic Models. Association for Computational Linguistics, pp. 262–272. Accessed: Nov. 23, 2023. [Online]. Available: <https://aclanthology.org/D11-1024>.
- [44] OpenAI, 2024. ChatGPT. Version 3.5. Accessed: Jan. 10, 2024.