Topic Modelling of Ukraine War-Related News Using Latent Dirichlet Allocation with Collapsed Gibbs Sampling

Nina Khairova[1,2], Yehor Holyk[3], Dmytro Sytnikov[3], Yurii Mishcheriakov[3] and Nadiia Shanidze[1]

[1] National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine
[2] Umeå University, UNIVERSITETSTORGET 4, Umeå, 901 87, Sweden
[3] Kharkiv National University of Radio Electronics, Nauky Ave. 14, Kharkiv, 61166, Ukraine

Abstract

The context of this research is the application of topic modeling to war-related news in the context of the Ukraine war. The objective of the research is to use Latent Dirichlet Allocation (LDA) with Collapsed Gibbs sampling to identify distinct content groups in war-related news. The method used in the research involves data scraping from a Ukrainian news website, data preprocessing, and applying the LDA with Collapsed Gibbs algorithm to infer the latent topics within the corpus. The results of the research include the identification of twelve distinct topics and the corresponding keywords that characterize each topic. The analysis of the results provides insights into the context of each topic, such as discussions on safety measures during wartime, consequences of military actions, and reports on military casualties. The research concludes that the application of LDA with Collapsed Gibbs is a valuable tool for identifying and understanding the context of war-related news. However, there may be discrepancies between the results of the model and human interpretation, which may be due to limitations in the results, model parameters, and the presence of noise data. Future research should focus on optimizing model parameters, filtering noise data, and improving the analysis of topic context to enhance the reliability and interpretability of the results.

Keywords

Topic modeling, Ukraine war, Latent Dirichlet Allocation

1. Introduction

With the progress of information technologies, concepts such as "information warfare", "information hygiene", and "hybrid warfare" have emerged in modern warfare. These terms have appeared not without reason, as wars and conflicts now take place not only on the battlefield but also in cyberspace and the information environment. Information warfare is used to manipulate public opinion, influence political processes, and destabilize countries or regions. This can include spreading disinformation, fake news, cyberattacks on critical infrastructure, and so on. Hybrid warfare combines military actions with unofficial, unconventional methods of warfare, such as subversive activities, psychological warfare, economic pressure, and more [1–2]. This can involve destabilizing countries through supporting internal conflicts, hybrid military operations, cyberattacks, and other methods of influence.

To counter information attacks, it is important to be able to classify types of information by their content. The use of machine learning algorithms has a wide range of applications, including natural language processing (NLP). They are widely used for tasks involving large volumes of data, which is advantageous when dealing with abundant information.

The research topic is the application of topic modeling to war-related news to identify distinct content groups. Topic modeling of news will allow for the separation of information by content

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

khairova.nina@gmail.com (N. Khairova); yehor.holyk@nure.ua (Y. Holyk); dmytro.sytnikov@nure.ua (D. Sytnikov), iurii.mishcheriakov@nure.ua (Y. Mishcheriakov); nashanidze@ukr.net (N. Shanidze)

0000-0002-9826-0286 (N. Khairova); 0009-0007-6325-1666 (Y. Holyk); 0000-0003-1240-7900 (D. Sytnikov); 0000-0002-5334-1808 (Y. Mishcheriakov), 0000-0002-9613-186X (N. Shanidze)

and observation of the context behind the keywords of these separate groups. In the future, the results could be valuable for comparing news from different news websites for the presence of disinformation or manipulation.

The objective of the research is to apply topic modeling to war-related news to identify distinct content groups. The research object is a news website. The subject of the research is Latent Dirichlet Allocation (LDA) with collapsed Gibbs sampling, belonging to the field of machine learning. LDA is a generative probabilistic model used for topic modeling in natural language processing (NLP). It is a technique used to discover hidden topics in a collection of documents by modeling how words are generated from these topics and how documents are generated from a mixture of topics.

2. Related works

2.1. Topic modelling usage in news

Topic modeling of news is a relevant topic for contemporary research. In the work [3], the possibility of applying the Latent Dirichlet Allocation (LDA) method to Indonesian news in the context of infrastructure development in the country is considered. The specificity of applying this method to news lies in finding the optimal number of topics among all news, as an excessive number of topics can lead to confusion and incomplete results. To find the optimal number of topics, it was proposed to determine the number of topics within a certain range and evaluate the coherence value for each number of topics. The highest value can be used for further analysis of topics. The conventional Latent Dirichlet Allocation (LDA) method was used in the work to conduct topic modeling and identify topics related to infrastructure development. In addition to the application of the method, the authors used data visualization models to display the obtained results.

In [4], two commonly used topic modeling methods are Latent Dirichlet Allocation (LDA) and BERTopic. They are employed to analyze the change in topics in Swedish newspaper articles about COVID-19. This allowed for obtaining more information about the main topics and topic changes in a large volume of data. The study processed 6515 articles, applying methods and tracking topic change statistics over approximately 1 year and 2 months from January 17, 2020, to March 13, 2021.

The article [5] describes the methodology used in a study to analyze the portrayal of urology in the media. The researchers collected data from news articles using a search term and extracted relevant information using Python's 'beautiful soup' library. They then preprocessed the data by segmenting the text and removing unnecessary words. The data was analyzed using Latent Dirichlet Allocation (LDA) topic modeling to identify key topics and associated words. The results showed that topics such as research and developments in new technologies, urinary conditions, health insurance coverage, and robotic surgery were frequently discussed in urology-related news.

2.2. Topic modelling usage in social media

The utilization of topic modeling in social media analysis has become increasingly prominent in contemporary research, offering valuable insights into the dynamic landscape of online discourse. [6] focuses on utilizing Twitter data to enhance disaster response and management efforts. The methods employed include natural language processing (NLP), particularly a supervised approach for classifying tweets into different categories to extract situational awareness (SA) information. However, it highlights the limitations of high-performing supervised models due to their reliance on domain knowledge and costly labeling tasks. To address these limitations, the research proposes a guided latent Dirichlet allocation (LDA) workflow to identify temporal latent topics from tweets during the 2020 Hurricane Laura disaster event. By integrating prior knowledge, coherence modeling, LDA topic visualization, and validation from

official reports, the guided approach reveals that tweets during Hurricane Laura contain multiple latent topics. This finding suggests that existing supervised models may not fully exploit tweet information, as they assign each tweet a single label. In contrast, the proposed model not only identifies emerging topics during different disaster events but also provides multilabel references to enhance classification accuracy. Additionally, the results can aid in quickly extracting SA information for responders, stakeholders, and the general public to facilitate timely response strategies and resource allocation during hurricane events.

Another example of using LDA in social networks is described in the [7]. This study focuses on analyzing Instagram data related to the Healthy Living Community Movement (GERMAS) in Indonesia. The country is currently facing a double burden of disease, with a shift in disease patterns due to changes in people's lifestyles. The researchers used Data Mining techniques, specifically Latent Dirichlet Allocation (LDA), to model topics from the data captions on Instagram. They collected 80,745 data captions with the "#germas" keyword and performed preprocessing and feature extraction before applying LDA. The evaluation of the number of topics was done using topic coherence, and the results showed that eight topic segments were most appropriate. The content analysis revealed that the most dominant topic related to GERMAS was a healthy lifestyle diet. This study highlights the importance of Instagram data in providing new media information for the community and the health department, and it can help promote a healthy lifestyle among the population.

3. Methods

To address the task of topic modeling news related to the theme of war, the following algorithm of actions is proposed, depicted in the form of a flowchart in the Figure 1:

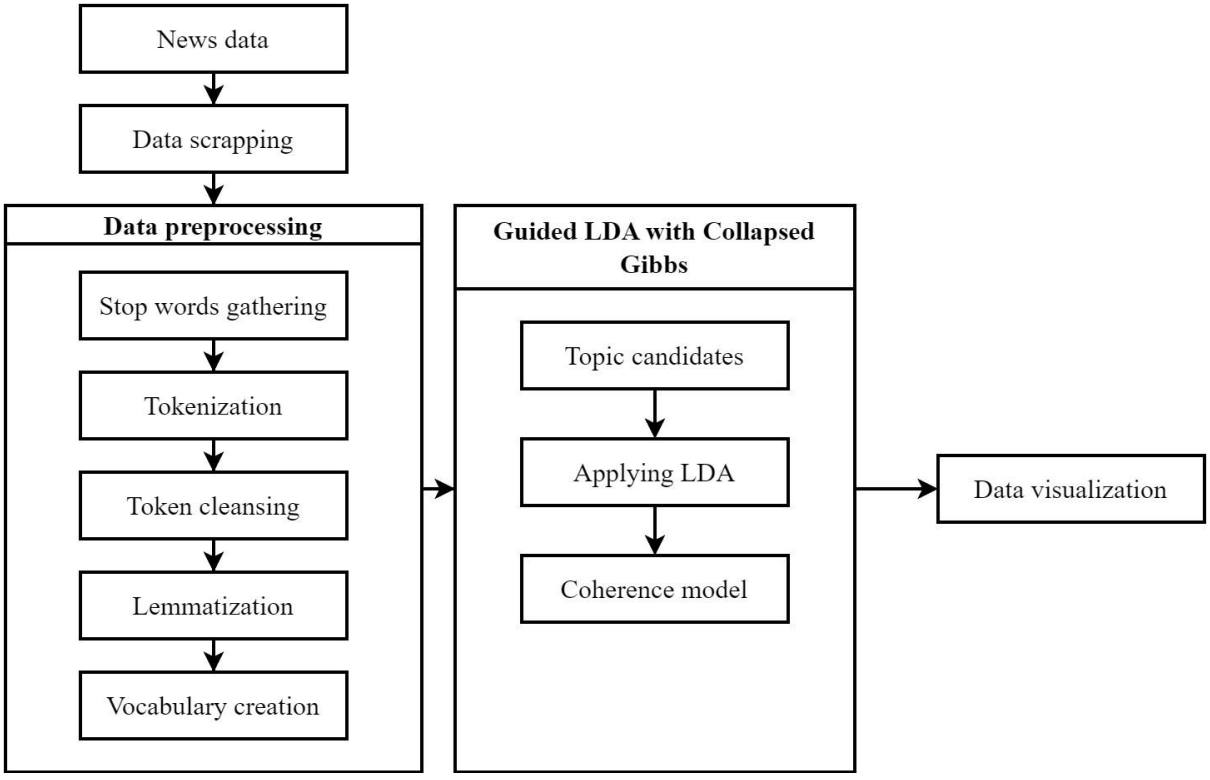

Figure 1: Overview of the proposed workflow of Guided LDA with Collapsed Gibbs model.

3.1. Data scrapping

The process of data collection and extraction is unique and requires an individualized approach for each data source. The easiest way to obtain data from news sources is by extracting

data from news websites, as they are stored in a structured format. The complexity lies in the need to familiarize oneself with the structure of the HTML code of the news website, and locate the tags and classes corresponding to the news content. Additionally, one must also consider the structure of requests to the news website, especially if the news is stored across multiple pages, requiring separate requests for each page. The LxmlSoup library in the Python programming language is a convenient tool for extracting data from HTML documents. Its versatility and robust functionality make it an indispensable asset for web scraping and data extraction tasks. With LxmlSoup, it can be easy to navigate through HTML documents, swiftly locating specific elements based on class names, identifiers, or other attributes.

For the study, the Ukrainian news website 'Ukrainska Pravda' [8] was selected. This news website covers events related to various domains, including war, military actions, and shelling, as well as political, social, and economic developments within Ukraine and internationally.

This website stores news on pages by specific date, namely day, month, and year. The structure of queries to such a news website is very simple, and by using the Python programming language, it will be possible to extract the necessary data.

3.2. Data preprocessing

To analyze the news, it is necessary to first perform a preprocessing step. This step is designed according to the specificities of applying the Latent Dirichlet Allocation method with Collapsed Gibbs.

Text preprocessing is a crucial step in natural language processing (NLP) that involves transforming raw text data into a format that can be easily understood and analyzed by machine learning algorithms. It involves several stages of data cleaning and transformation to prepare the text for feature extraction and analysis.

When using Latent Dirichlet Allocation, it is important to customize preprocessing steps to match the needs of the algorithm. One example of this is preprocessing text to maintain the structure of phrases or multi-word combinations, as this can provide useful information on topic consistency and thematic connections.

The main goals of text preprocessing are to simplify the text, reduce noise and variability, and extract meaningful features. This is important because machine learning algorithms typically require numerical inputs and unstructured text data needs to be converted into a structured format.

The preprocessing steps typically involved in text preprocessing include:

1. Stop word gathering. Stop words are commonly used words in a language that do not carry much meaning and are often removed during text preprocessing in natural language processing (NLP). These words, such as "the," "is," "and," etc., occur frequently in text but do not provide much information for analysis. By removing stop words, we can reduce the size of the dictionary of unique words, which can improve the efficiency and performance of natural language processing (NLP) algorithms. Gathering a list of stop words helps in identifying and removing these common words from the text data before further analysis and feature extraction
2. Tokenization. This involves breaking the text into individual words or tokens. It is the first step in feature extraction and involves splitting the text at the word level
3. Token cleansing. Stop words are common words that do not carry much meaning. These words are often removed from the text as they can add noise and do not contribute much to the overall meaning of the text
4. Lemmatization. This step involves reducing words to their base or root form. This helps in reducing the dimensionality of the data and ensures that variations of the same word are treated as the same entity. This can be done through stemming or lemmatization techniques
5. Vocabulary creation. A master dictionary is a collection of all the unique words in the corpus. It helps in creating a standardized representation of the text data and serves as a reference for feature extraction

Once the text has been preprocessed, various techniques can be used to represent the text as numerical features. One popular technique is the bag-of-words (BOW) representation, where each word in the text is treated as a separate feature [9]. The values of these features can be modified based on techniques such as term count, term frequency (TF), and term frequency-inverse document frequency (IDF) [9]. These techniques assign different weights to words based on their frequency and importance in the corpus.

Overall, text preprocessing is a critical step in NLP as it helps in transforming unstructured text data into a structured format that can be easily analyzed using machine learning algorithms. It involves several stages of data cleaning and transformation to extract meaningful features and simplify the text.

3.3. Guided LDA with Collapsed Gibbs

Selecting the optimal number of topics for a given corpus is a critical step in topic modeling. The coherence model, particularly when integrated with guided Latent Dirichlet Allocation (LDA) using Collapsed Gibbs sampling, serves as a robust method for this purpose.

Topic modeling aims to uncover the latent thematic structure within a corpus of text. Determining the appropriate number of topics is pivotal for obtaining meaningful insights. The coherence model, coupled with guided LDA employing Collapsed Gibbs sampling, facilitates this task effectively. This process includes:

1. Topic Candidates. This step begins by generating a range of potential topic numbers, often referred to as topic candidates. These candidates represent the spectrum of possible thematic structures that the corpus might encapsulate. The range is typically predefined based on domain expertise or through iterative exploration
2. Applying LDA. This involves utilizing guided LDA with Collapsed Gibbs sampling to model each candidate's number of topics on the corpus. The process entails iteratively inferring the topic-word distributions and document-topic assignments, thereby uncovering the underlying thematic structure of the text
3. Coherence Model. After applying LDA to each candidate topic number, the coherence of the resulting topics using a coherence model is calculated. The coherence model assesses the semantic coherence and interpretability of the topics by measuring the relatedness of the top words within each topic
4. Optimal Number of Topics Selection. This step involves selecting the optimal number of topics based on coherence scores. Higher coherence scores indicate more coherent and interpretable topics. Thus, the number of topics corresponding to the peak coherence score is considered optimal for representing the thematic structure of the corpus
5. Refinement and Validation. This involves refining the selected number of topics if necessary, considering contextual relevance and domain-specific requirements. It is essential to validate the chosen number of topics through qualitative analysis and expert judgment to ensure alignment with the corpus's underlying themes

3.3.1. Latent Dirichlet Allocation with Collapsed Gibbs

Latent Dirichlet Allocation (LDA) is a generative probabilistic model commonly used for topic modeling, which aims to discover the latent topics present in a collection of documents and the distribution of words within each topic. Collapsed Gibbs sampling is a method used to estimate the posterior distribution of latent variables in a Bayesian model, such as LDA. The procedure for LDA with Collapsed Gibbs unfolds as follows:

1. Initialization. The process commences with initializing model parameters, encompassing topic distributions for each document, word distributions for each topic, and hyperparameters such as the number of topics and Dirichlet priors for the distributions [10]
2. Gibbs Sampling. At the core of the LDA algorithm lies a Gibbs sampling iteration that continually updates topic assignments for every word in the corpus. During each iteration, the algorithm samples a new topic assignment for a randomly selected word, influenced by

existing topic assignments of all other words in the document and the prevailing topic distributions [10]

3. Convergence. The Gibbs sampling process persists until reaching convergence, typically identified through various convergence criteria like a predetermined number of iterations, minimal alterations in the model's log-likelihood, or slight adjustments in topic assignments [10]
4. Estimation of Parameters. Upon achieving convergence, the model's parameters can be estimated, encompassing topic distributions for each document, word distributions for each topic, and hyperparameters [10]
5. Inference. Utilizing the estimated parameters facilitates inference tasks, such as determining the most probable topics for a new document, estimating a document's likelihood given the model, or identifying the most likely words associated with a given topic [10]

4. Experiment

The experiment on topic modeling of war-related news was conducted within the Jupyter Notebook environment with the capability of utilizing GPU. The implementation of the program code was done using the Python programming language and the following packages: LxmlSoup, requests, datetime, json, numpy, spacy, nltk, and random.

Initially, an algorithm was developed to generate a set of dates within a specified range, and then, using HTTP requests, the HTML code of the news page from the “Ukrainska Pravda” [8] website was obtained. From each news page, its title and article text were extracted. To obtain news related to the theme of war, each article was checked for the presence of the following phrases, words, or abbreviations: “attack on”, "explosions rock","attack in", "war", "agression", "millitary", "support", "ukraine", "europ"," "united", "states". These words were manually selected by analyzing which phrases or combinations of words were most frequently encountered in articles related to the theme of war. The retrieved data was saved in a file named “news.json”. This dataset will undergo the preprocessing process. Thus, within the specified date range between 02/17/2023 and 02/17/2024, a total of 2364 news articles were obtained, containing a total of 423,251 words, and 13,414 unique words. Fragment of the extracted data is shown in Figure 2.


```
[
  {
    "aritcle_header": "Ukraine's Air Force downs 9 out of 10 Shahed UAVs launched by Russians",
    "article_text": "Ukrainian air defence forces shot down 9 out of 10 Russian drones during the afternoon attack on Monday, 1
    January.\nSource: Air Force of the Armed Forces of Ukraine\nQuote: \"At about 14:00 on 1 January 2024, Russian invaders launched
    10 assault UAVs of the Shahed-136/131 type from the north.\nAir defence destroyed nine enemy Shahed drones.\nOne Kh-59 guided
    aircraft missile was also destroyed in the eastern direction.\"\nBackground:\nSupport UP or become our patron!\nAll content
    posted on this website with reference to the \"Interfax-Ukraine\" news agency are not subject to further reproduction and/or
    distribution in any form, except with the written permission of the \"Interfax-Ukraine\" news\u00a0agency.\nFounder: Georgiy
    Gongadze\nEditor-in-chief: Sev\u011fil Musaieva\nFounding Editor: Olena Prytula\nContact us: upeng@pravda.ua"
  },
  {
    "aritcle_header": "Explosions rock Odesa and Dnipro",
    "article_text": "Explosions were heard in the cities of Odesa and Dnipro on the night of 31 December-1 January.\nSource:
    Dumskaia, an Odesa-based local news outlet; Suspilne on Telegram\nDetails: Local journalists stated that air defence systems were
    responding to the attack in Odesa.\nBackground: On the evening of 31 December, several dozen Russian attack drones were flying in
    Ukrainian airspace.\nSupport UP or become our patron!\nAll content posted on this website with reference to the
    \"Interfax-Ukraine\" news agency are not subject to further reproduction and/or distribution in any form, except with the written
    permission of the \"Interfax-Ukraine\" news\u00a0agency.\nFounder: Georgiy Gongadze\nEditor-in-chief: Sev\u011fil
    Musaieva\nFounding Editor: Olena Prytula\nContact us: upeng@pravda.ua"
  },
```


Figure 2: Sample of News Articles from “news.json” file

Next, during the data preprocessing stage, the "news.json" file obtained in the previous step was read. For each article, the title and text of the article were combined and placed into a document container. In order to improve text analysis, a carefully selected list of stop words was created. This list was compiled using an additional file called "stopwords-en.json" and utilizing the features of the English language text processing tool "en_core_web_sm" from the spacy package. By utilizing these resources, a thorough stop word list was developed to eliminate redundant and unhelpful terms that could mask important patterns in the text data. After filling the document container, we carefully cleaned the data in each document to improve its quality

and relevance. This involved a series of important operations designed to enhance the text corpus, such as:

- Tokenization. This step involves breaking the text into smaller units, like words or phrases, to make it easier to study and work with. It is an essential first step in preparing the text for more in-depth analysis and handling. This initial phase lays the groundwork for other tasks involved in processing the text
- Removal of tokens belonging to the compiled list of stop words. Stop words were removed from the text to reduce noise and focus on important content for analysis
- Removal of tokens belonging to punctuation marks. Extraneous punctuation marks were stripped from the text to ensure consistency and readability, mitigating any potential interference with subsequent processing steps
- Removal of tokens consisting of numbers. Tokens that only contained numbers were removed from the text to keep the attention on language-based content, avoiding the inclusion of numeric information that could potentially impact the analysis
- Removal of tokens representing email addresses. Tokens representing email addresses were filtered out to maintain data privacy and integrity, ensuring that personal information did not influence the analysis
- Lemmatization of the text to bring words with the same root to a unified form. The process of lemmatization was applied to the text data, standardizing words to their base or root form. This step aimed to reduce lexical variation and enhance the coherence of the text corpus, thereby facilitating more accurate analysis and interpretation

Through careful execution of these preprocessing steps, the text underwent a thorough refinement process to ensure it was suitable for analysis and modeling. This laid the foundation for extracting valuable insights and patterns from the text, ultimately promoting a deeper understanding and exploration of the data.

Next, a structure was created with a key-value pair where the key is the word and the value is the total count of the word appearing in the documents. For each token in each document in the container, the token was added to the structure containing the word count, and the value was incremented by one. This way, a dictionary of all possible words encountered in the documents was obtained. This dictionary was then filtered so that only those words whose count exceeds 5 remained in the dictionary. From this set, a vocabulary was formed where each word contains its identifier, and a mirror vocabulary was created where each identifier contains the word. The mirror vocabulary will be necessary at the end of the experiment for the reverse conversion of identifiers to words, as part of the decoding process. Once the dictionaries were established, the corpus was preprocessed to convert each word token into its corresponding identifier using the dictionary of word-to-identifier mappings. This transformation facilitated the subsequent modeling steps by representing the corpus in a numerical format suitable for analysis. Additionally, the mirror dictionary ensured that the original words could be reconstructed from their identifiers when interpreting the model's output or evaluating its performance. This bidirectional mapping between words and identifiers formed a crucial component of the data preparation phase, enabling seamless integration of the corpus into the modeling pipeline.

Next, a corpus was created, which is a collection of documents to be analyzed. For each token in each document in the container, word-to-identifier mapping was applied, as Latent Dirichlet Allocation with Collapsed Gibbs sampling typically requires numerical representations of the words instead of their textual forms for efficient processing. This mapping converts each word token into a unique identifier, allowing the algorithm to operate on numerical data. With this preprocessing step completed, the LDA with Collapsed Gibbs sampling algorithm can proceed to infer the latent topics within the corpus and the associated word distributions, facilitating a deeper understanding of the underlying thematic structure. In Figure 3, an example of a document from the corpus before mapping is shown, and in Figure 4, an example of a document from the corpus after mapping is shown.

```
['ukraine', 'air', 'force', 'down', 'shahe', 'uav', 'launch',
'russians', 'ukrainian', 'air', 'defence', 'force', 'shoot',
'russian', 'drone', 'afternoon', 'attack', 'monday', 'january',
'source', 'air', 'force', 'armed', 'forces', 'ukraine', 'quote',
'january', 'russian', 'invader', 'launch', 'assault', 'uav',
'shahed-136/131', 'type', 'north', 'air', 'defence', 'destroy',
'enemy', 'shahed', 'drone', 'kh-59', 'guide', 'aircraft',
'missile', 'destroy', 'eastern', 'direction', 'background',
'support', 'patron', 'content', 'post', 'website', 'reference',
'interfax', 'ukraine', 'news', 'agency', 'subject',
'reproduction', 'and/or', 'distribution', 'form', 'write',
'permission', 'interfax', 'ukraine', 'news', 'agency',
'founder', 'georgiy', 'gongadze', 'editor', 'chief', 'sevğil',
'musaieva', 'founding', 'editor', 'olena', 'prytula', 'contact']
```

Figure 3: Corpus before mapping

```
array([   0,   18,   19,   89,   11, 1082,   26,  529,   78,    0,   18,
         19,  128,  824,    7,  101,  807,   92,  568,  113,   89,   21,
         26,  567,  564, 1082,  915,   17,  569,  570,   18,   19,    0,
        571,   75,  171,  258,  520,   11,  622,  355,   81,  572,  186,
        573,  574,  575,  297,  520,  576,  124,  143,   36,   37,   38,
         39,   40,   41,   42,    0,   43,   44,   45,   46,   47,   48,
         49,   50,   51,   42,    0,   43,   44,   52,   53,   54,   55,
         56,   57,   58,   59,   55,   60,   61,   62])
```

Figure 4: Corpus after mapping

According to [11], the implementation of the software code executing the Latent Dirichlet Allocation with Collapsed Gibbs algorithm was performed, which takes the following parameters:

- CORPUS – the corpus being analyzed
- NUM_ITER – the number of iterations
- ALPHA – Dirichlet prior parameter for the distribution of topics in documents
- BETA – Dirichlet prior parameter for the distribution of words in topics
- NUM_TOPICS – the number of topics

Implemented function initializes topic assignments for each word in each document randomly. Then, it estimates document-topic counts (NDK), topic-word counts (NKW), and topic counts (NK) based on the initial assignments. Next, it iterates through the specified number of iterations, updating topic assignments for each word in each document using the collapsed Gibbs sampling algorithm. Finally the function returns the final topic assignments (Z), document-topic counts (NDK), topic-word counts (NKW), and topic counts (NK).

In addition to the LDA with Collapsed Gibbs algorithm, using [12], an auxiliary function has been developed to determine the coherence score for the model depending on the number of topics. This function will be used to analyze the coherence score of models with different numbers of topics to select their optimal quantity. The coherence score calculation function takes the following parameters:

- NKW – topic-word count matrix obtained from LDA Gibbs sampling
- TEXTS_FOR_LDA – textual representation of documents
- CORPUS_FOR_COHERENCE – corpus in the required format for coherence calculation
- DCT – dictionary mapping words to their integer indices
- NUM_OF_TOPICS – number of topics inferred from the corpus

This function calculates coherence scores for the inferred topics, computes the topic coherence by considering the top 20 words per topic, and constructs a coherence model using the specified coherence measure. In this case, 'c_v' coherence was chosen because this coherence measure is a widely used metric for evaluating topic coherence in Latent Dirichlet Allocation (LDA) models [12]. By selecting 'c_v' coherence, the function aims to provide a coherence score that reflects the interpretability and semantic coherence of the topics, making it easier for researchers or practitioners to assess the quality of the topics generated by the LDA model and compare different models or parameter settings effectively [13].

With the ALPHA = 0.1, BETA = 0.1, NUM_ITER = 200 and the NUM_TOPICS ranging between 7 and 12, an LDA with Collapsed Gibbs algorithm was executed to evaluate the coherence score of the processed models. Table 1 depicts the coherence scores of the models depending on the number of topics. Figure 5 illustrates the coherence score graph of the processed models.

Table 1
Coherence score of the processed models

Number of topics	Coherence score
7	0.4486
8	0.3656
9	0.3965
10	0.317
11	0.3893
12	0.477

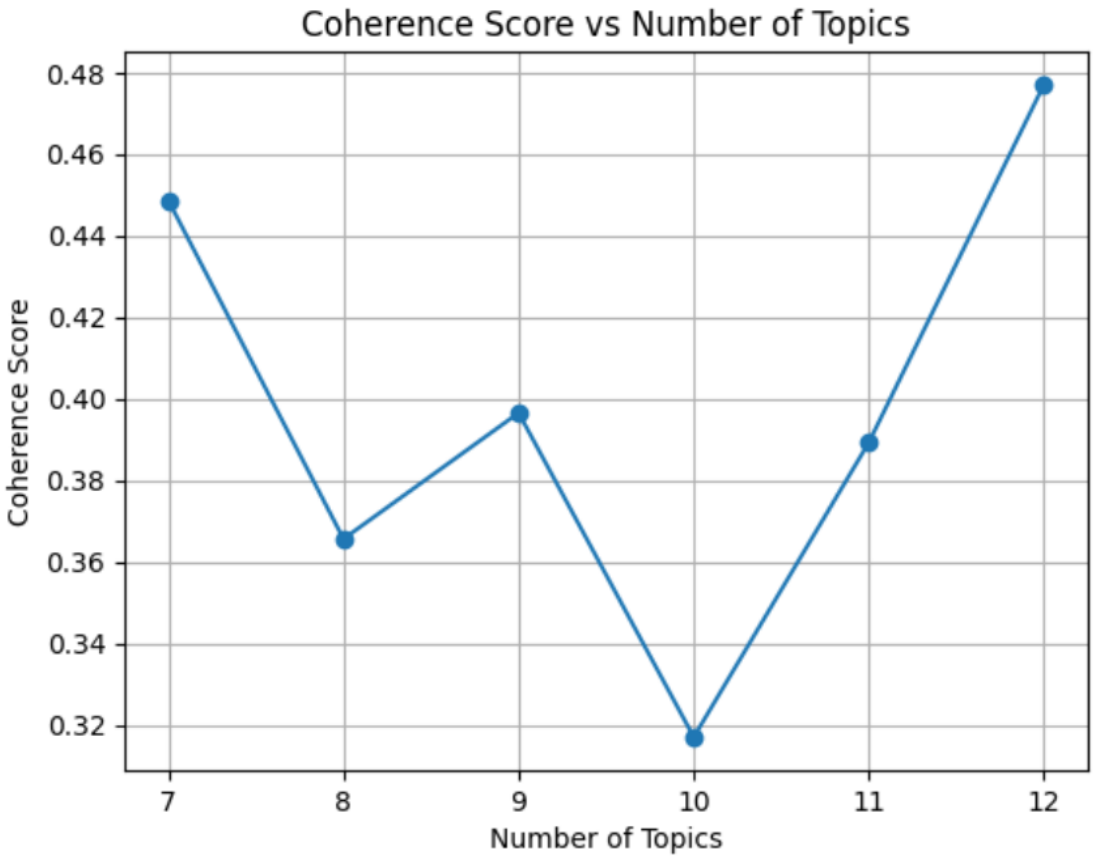

Figure 5: Coherence score graph of the processed models

Based on Table 1, it can be seen that the most appropriate number of topic segments is twelve topic segments. Therefore, to analyze the results of the Latent Dirichlet Allocation with Collapsed Gibbs, a model with 7 topics will be used.

5. Results

Based on the results of Latent Dirichlet Allocation with Collapsed Gibbs on war-related Ukrainian news, a comprehensive analysis was conducted to generate Table 2, which illustrates the topic number alongside 20 keywords that distinctly characterize each of the identified topics.

After obtaining the results of the LDA with Collapsed Gibbs method, namely twelve identified topics and twenty keywords for each topic, a thorough analysis of the keywords was conducted to understand the context of news related to the topic of war. As a result of the analysis, Table 3 was formed, which illustrates the approximate context for each topic.

Table 2
Topic Modeling Results of War-related Ukrainian News Using LDA with Collapsed Gibbs

Topic number	Top 20 words
Topic 1	[agency, spirne, protective, safe, education, threat, metro, simultaneously, senate, house, representatives, direct, shutdown, negative, impact, engage, internal, regulatory, underground, exhaust]
Topic 2	[forces, armed, destroy, kill, loss, agency, russian, february, form, late, tank, contact, russia, total, represent, invader, russians, troop, continue, artillery]
Topic 3	[armed, forces, kill, russian, destroy, agency, loss, december, parenthesis, represent, armoured, tank, reproduction, contact, confirm, distribution, continue, invader, figure, combat]
Topic 4	[chicherina, senate, protective, safe, education, threat, metro, simultaneously, house, underground, representatives, direct, shutdown, negative, impact, engage, regulatory, exhaust, chamber, pivnichne]
Topic 5	[armed, forces, loss, russian, destroy, kill, agency, support, represent, late, parenthesis, liberation, carrier, combat, armoured, artillery, russians, form, total, continue]
Topic 6	[forces, armed, agency, loss, kill, destroy, russian, liberation, carrier, russians, tank, support, war, total, confirm, figure, combat, contact, source, form]
Topic 7	[forces, armed, loss, destroy, russian, agency, kill, carrier, russia, liberation, figure, february, invader, total, reproduction, support, continue, parenthesis, confirm, armoured]
Topic 8	[troop, reproduction, armed, senate, safe, education, threat, metro, simultaneously, house, regulatory, representatives, direct, shutdown, negative, impact, engage, internal, protective, exhaust]
Topic 9	[armed, forces, russian, agency, loss, destroy, kill, continue, source, figure, liberation, distribution, february, combat, war, confirm, represent, form, artillery, russia]
Topic 10	[forces, armed, kill, destroy, loss, agency, russian, late, contact, combat, source, invader, armoured, reproduction, war, total, support, distribution, figure, artillery]
Topic 11	[forces, armed, russian, kill, agency, destroy, loss, december, troop, source, distribution, february, russia, parenthesis, artillery, war, tank, figure, confirm, armoured]
Topic 12	[forces, armed, kill, agency, destroy, loss, russian, war, russians, february, carrier, troop, invader, armoured, artillery, continue, reproduction, russia, combat, confirm]

Table 3
Topic context analysis

Topic number	Topic context analysis
Topic 1	Discussion of safety measures during wartime. Utilization of shelters, underground infrastructure as shelters.
Topic 2	Consequences of military actions. Report on enemy military operations in February.
Topic 3	Consequences of military actions. Report on enemy military operations in December.
Topic 4	Discussion of the consequences of shelling on the civilian population
Topic 5	Report on military casualties due to defensive or offensive actions.
Topic 6	Report on military casualties due to defensive or offensive actions.
Topic 7	Report on military losses resulting from defensive or offensive actions. Discussion on support for Ukraine in conducting military operations.
Topic 8	Discussion of threats resulting from military actions.
Topic 9	Consequences of military actions. Report on enemy military operations in February.
Topic 10	Report on military casualties due to defensive or offensive actions.
Topic 11	Report on military casualties due to defensive or offensive actions.
Topic 12	Consequences of military actions. Report on enemy military operations in February.

Table 3 shows the results of the analysis of the approximate context of the topics based on keywords.

For topic 1, the most characteristic keywords are "safe", "education", "threat", "metro", "shutdown", "negative", "impact". Based on these keywords, the context for topic 1 is likely to be Discussion of safety measures during wartime and the usage of shelters or underground infrastructure as shelters.

Topic 4 has very similar keywords to topic 1, although some of them, such as "threat", "direct", "shutdown", "negative", "impact", most likely indicate consequences of shelling on the civilian population.

Topic 8, based on the keywords, is similar to topics 1 and 4, but also includes the keywords "troop" and "protective", which most likely indicate reports in the news about future threats to the civilian population due to military actions.

Topics 2, 3, 5, 6, 7, 9, 10, 11, 12 were the most difficult to analyze in terms of context. In these topics, about half of the keywords overlap, such as "forces", "armed", "destroy", "kill", "loss", "russian", "invader", "artillery". All these words are directly related to the war, as they reflect events taking place on the territory of Ukraine. However, it is possible to distinguish separate groups of topics, such as:

- Topics 5, 6, 10, 11, which reflect reports on military losses due to offensive or defensive actions. Characteristic keywords: "liberation", "combat", "invade"
- Topics 2, 3, 9, 12, which reflect the consequences of military actions. Characteristic keywords: "total", "loss", "represent", "destroy"
- Topic 7, which likely highlights discussion on support for Ukraine in conducting military operations. Characteristic keywords: "support", "continue", "armoured"

Thus, based on the conducted analysis of the context, it can be said that among the 12 topics identified by the Latent Dirichlet Allocation with Collapsed Gibbs method, there are topics that are semantically similar to each other based on keywords, and in reality, only 6 distinct topic groups are likely to be distinguished.

For visual interpretation of the results, a distance map of the obtained outcomes was constructed. Pairwise distances between topics were calculated based on their word distributions. Common distance metrics were computed using cosine similarity. Subsequently, Multidimensional Scaling (MDS) was applied to project the topics into a two-dimensional space while maintaining the pairwise distances between them. MDS transforms the high-dimensional topic space into a lower-dimensional space, facilitating visualization. After the initial iteration of data visualization, it was decided to display topics 1, 3, 5, 6, 7, 8, 9, 10, and 12 in a separate plot. The results of data visualization in the form of a Distance Map are provided on Figure 6 and Figure 7, respectively. To emphasize the lexical patterns within war-related news discourse, a word cloud was constructed from the set of terms extracted from the news data. As shown in Figure 8, this graphical tool visually highlights the most common terms to offer a snapshot of the thematic focus of the field.

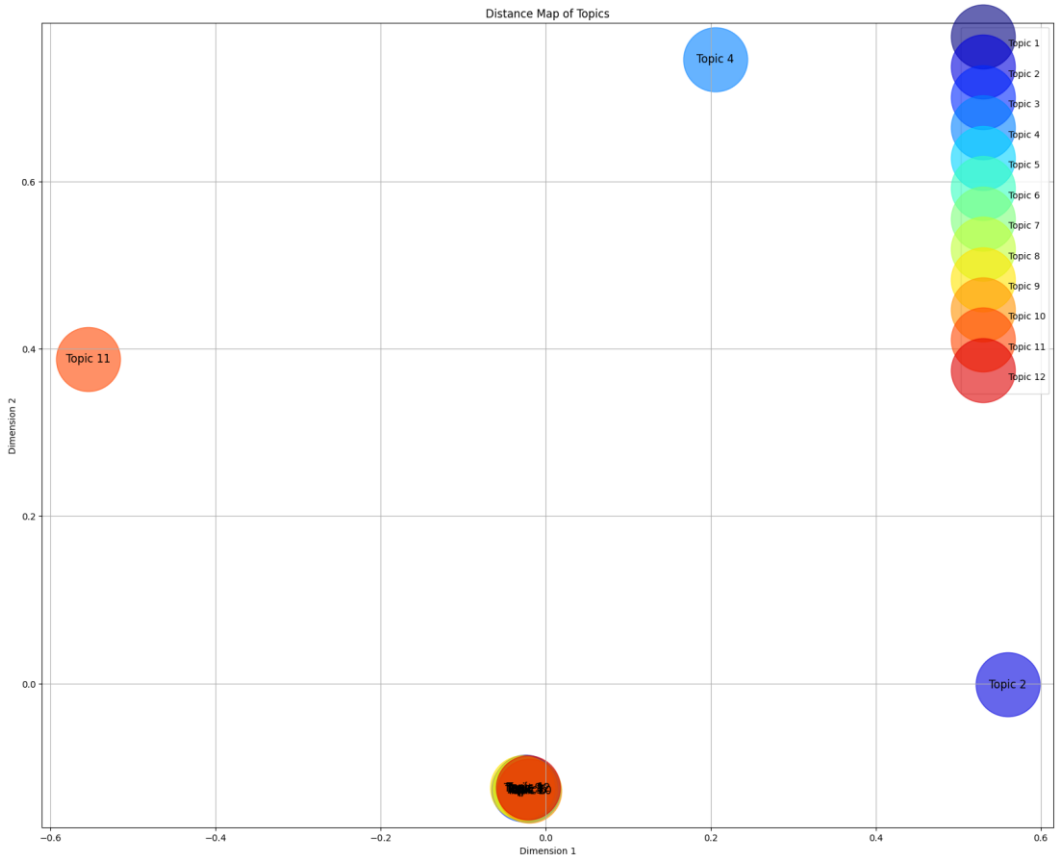

Figure 6: Distance Map of All Topic Segments

Figure 6 and Figure 7 depict the Distance Map of the created topic segments. Based on Figure 6, it can be concluded that the most distant topics in terms of semantic content are topics 2, 4, and 11. The rest of the topic segments are closest in semantic content and form a cluster on the graph. Therefore, topic segments numbered 1, 3, 5, 6, 7, 8, 9, 10, 12 were separated and displayed on a separate Distance Map. Based on Figure 7, it can be concluded that although the topic segments are closest in semantic content, they do not intersect on the graph and have their own significance among other topics.

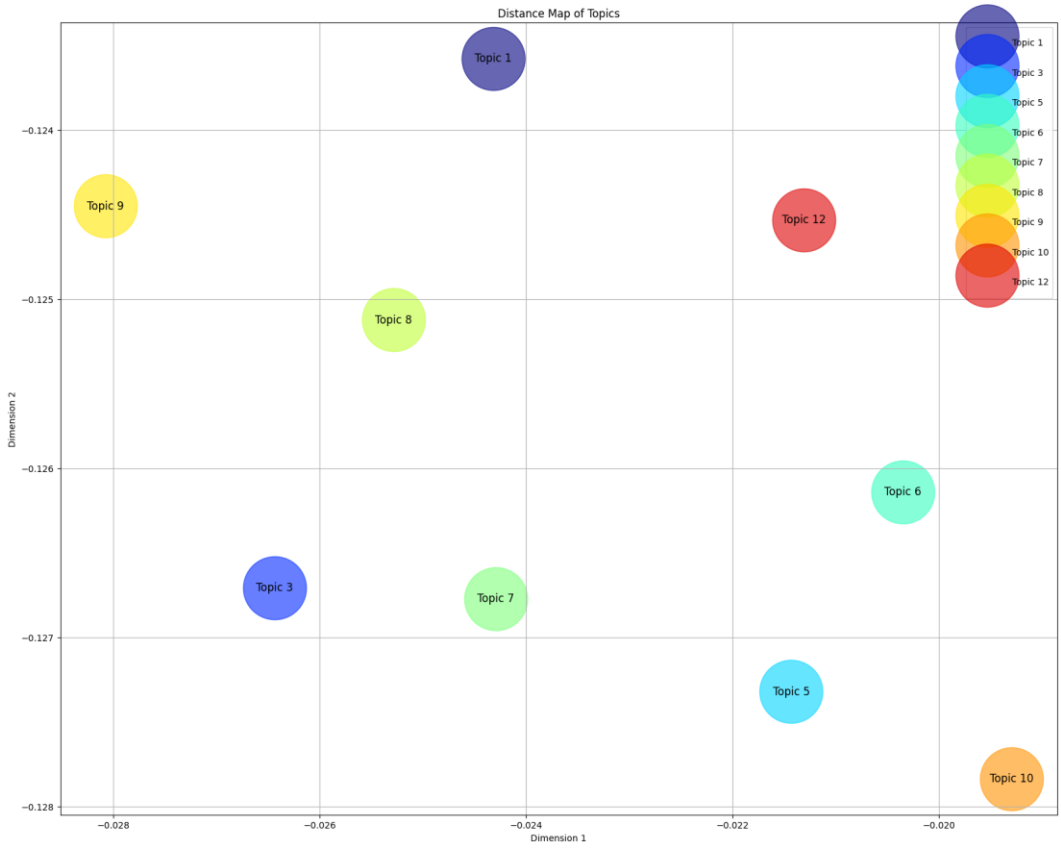

Figure 7: Distance Map of Topic Segments 1, 3, 5, 6, 7, 8, 9, 10, 12

Figure 8: Word cloud

6. Discussions

Comparing the results of the context analysis of topics based on keywords and the results of displaying topic segments on the Distance Map graph, it can be said that there are discrepancies between the conclusions made by humans and those made by the LDA with Collapsed Gibbs method. This could be due to several factors:

1. Limitation of results. Since only 20 keywords were obtained for each topic, this may lead to loss of context and insufficient completeness in representing each topic segment. Some keywords may be common across multiple topics, complicating their interpretation. Additionally, excluding parts of the vocabulary that are not key to any topic may lead to loss of contextual information important for a complete understanding of the topic. Thus, analyzing the results of LDA with Collapsed Gibbs requires careful consideration and interpretation taking into account these limitations
2. Model parameters. Discrepancies in results may stem from the model being either overfit or underfit, as the number of iterations of the LDA with Collapsed Gibbs method and the number of topics were incorrectly set. Insufficient iterations may result in instability in the results of topic modeling, while too many iterations may lead to overfitting the model and formation of overly complex or insufficiently generalized topics. Furthermore, improper selection of the ALPHA and BETA parameters of the LDA with Collapsed Gibbs method may also distort the results. ALPHA controls the distribution of topics in documents, while BETA influences the distribution of words in topics. Improper tuning of these parameters may lead to underestimation or overestimation of the importance of topics or words in the model, affecting its accuracy and interpretability [14]. Thus, it is important to consider these parameters when analyzing the results of topic modeling
3. Noise data. Discrepancies in results may also arise from the processed data containing words that do not carry meaningful content but were not included in the list of stop words. This is because depending on the context, some words may be important for understanding the text, while others may not convey significant information. Despite the model being carefully checked multiple times for the presence of noise words, the results were difficult to interpret as the obtained keywords for topics were often semantically distant from each other

The question for future research is how these factors affect the reliability and interpretability of the results of topic modeling using the LDA with Collapsed Gibbs method. Limitation of results, model parameters, and the presence of noise data may affect the accuracy and completeness of topic and keyword detection, as well as their interpretability. For future research, it will be important to explore optimal strategies for model parameter selection, develop more effective methods for filtering noise data, and devise new approaches to analyzing the context of topic models to improve the quality and reliability of the results.

7. Conclusions

We can conclude that the application of topic modeling to war-related news using Latent Dirichlet Allocation (LDA) with Collapsed Gibbs sampling is a valuable tool for identifying distinct content groups and understanding the context behind key words in these groups. The research conducted in this study demonstrated the effectiveness of this approach in analyzing war-related news from a Ukrainian news website.

The analysis of the results showed that the LDA with Collapsed Gibbs method was able to identify distinct topics related to safety measures during wartime, consequences of military actions, discussions on threats resulting from military actions, and reports on military casualties. However, there were discrepancies between the results of the model and the human interpretation of the topics, which may be attributed to limitations in the results, model parameters, and the presence of noise data.

To improve the reliability and interpretability of the results, future research should focus on optimizing model parameters, developing more effective methods for filtering noise data, and

exploring new approaches to analyzing the context of topic models. Additionally, further investigation is needed to understand how these factors affect the accuracy and completeness of topic and keyword detection in order to enhance the quality of topic modeling in the field of war-related news analysis.

References

[1] O. Fridman, 'Information War' as the Russian Conceptualisation of Strategic Communications, RUSI J. 165.1 (2020) 44–53. doi:10.1080/03071847.2020.1740494.

[2] N. Tytova, K. Mereniuk, Digital literacy of future teachers in the realities of large-scale military aggression (Ukrainian experience), Futur. Educ. (2022) 43–54. doi:10.57125/fed/2022.10.11.33.

[3] A. F. Hidayatullah, M. R. Ma'arif, M. Habibie, S. Khomsah, Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation, IOP Conf. Ser. 1077.1 (2021) 012012. doi:10.1088/1757-899x/1077/1/012012.

[4] B. Griciūtė, L. Han, G. Nenadic, Topic Modelling of Swedish Newspaper Articles about Coronavirus: a Case Study using Latent Dirichlet Allocation Method, in: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), IEEE, 2023. doi:10.1109/ichi57859.2023.00110.

[5] Y. W. Oh, J. Kim, Insights Into Korean Public Perspectives on Urology: Online News Data Analytics Through Latent Dirichlet Allocation Topic Modeling, Int. Neurourol. J. 27.Suppl 2 (2023) S91–98. doi:10.5213/inj.2346288.144.

[6] S. Zhou, P. Kan, Q. Huang, J. Silbernagel, A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura, J. Inf. Sci. (2021) 016555152110077. doi:10.1177/01655515211007724.

[7] M. Habibi, A. Priadana, A. B. Saputra, P. W. Cahyo, Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA), in: International Conference on Health and Medical Sciences (AHMS 2020), Atlantis Press, Paris, France, 2021. doi:10.2991/ahsr.k.210127.060.

[8] G. Gongadze, Ukrainska pravda. URL: https://www.pravda.com.ua/eng/.

[9] K. Juluru, H.-H. Shih, K. N. Keshava Murthy, P. Elnajjar, Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists, RadioGraphics 41.5 (2021) 1420–1426. doi:10.1148/rg.2021210025.

[10] Y. E. Ogunwale, M. O. Ajinaja, Application Research on Semantic Analysis Using Latent Dirichlet Allocation and Collapsed Gibbs Sampling for Topic Discovery, Asian J. Res. Comput. Sci. 16.4 (2023) 445–452. doi:10.9734/ajrcos/2023/v16i4404.

[11] M. O. Ajinaja, A. O. Adetunmbi, C. C. Ugwu, O. S. Popoola, Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling, Iran J. Comput. Sci. (2022). doi:10.1007/s42044-022-00124-7.

[12] B. Srinivasa-Desikan, Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras, Packt Publishing, 2018.

[13] S. Mifrah, Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus, Int. J. Adv. Trends Comput. Sci. Eng. 9.4 (2020) 5756–5761. doi:10.30534/ijatcse/2020/231942020.

[14] A. Panichella, A Systematic Comparison of search-Based approaches for LDA hyperparameter tuning, Inf. Softw. Technol. 130 (2021) 106411. doi:10.1016/j.infsof.2020.106411.