Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

ShinyTopic: R Shiny Interactive Topic Modelling for Public Policy Narrative Case Study

Lukmanul Hakim*, Fadhilah Dhinur Aini

Fakultas Teknik dan Bisnis, Program Studi Informatika, Universitas Madani, Yogyakarta, Indonesia
Email: [1,*]lukmanulhakim@umad.ac.id, [2]fadhilah.da@umad.ac.id
Correspondencial email: lukmanulhakim@umad.ac.id

Abstract−The collection of unstructured text often yields a huge pile of raw data that cannot be easily measured with qualitative analysis. It is necessary to use a more varied form of assessment in order to allow researchers to explore large-scale collection of texts in an efficient manner. Using computer-aided to automate textual analysis offers them new opportunities rather than manual coding specifically for large amount of data, thus can saving resources. Topic model is one of the most common approach within the methodological domain of text mining and NLP. However, researcher in particular studies face technical barrier that limit systematic and replicable text analysis. ShinyTopic is web application built with R and Shiny packages that enable non-programmer who wish to perform topic models unifies with data import, pre-processing, topic analysis, metric assessment and annotation, visualization and result presentation. Here, interactive topic modelling contributes to research community for collaborative benefit by simplifying complex methodologies for different subject and increase reproducible research. The application employs LDA and CTM as topic model algorithms to conduct topic analysis. For demonstration purpose, 200 documents from news outlet were processes to illustrate how user can perform complete analytical workflows-begin with data import and ended to the result summary. Based on the interim result, we performed LDA topic modelling with 5 topics k extracted from the dataset. The highest probability topic terms among 5 topics related to the current Indonesia vice president ($\beta = 0.0730$), while UCI metric get the highest topic coherence score (0.7663). Although the average coherence score quite low, we still able to identify topic inference which are related to political figures, political party, people's representative, and current president and vice president. Whereas, CTM get lower result compare to LDA, yet it is interpretable overall. In spite of its advantages, there are some limitations of ShinyTopic in term of language structure preprocessing and stemming of Indonesian text. To overcome the issue, a custom function was built inside core program of ShinyTopic yet less accurate for Indonesian language structures which still need to improve.

Keywords: Unstructured text; Topic modelling; R Shiny; Text Mining; Web Application

1. INTRODUCTION

Popularity of text mining approaches across scientific and professional domains has played a significant role for several years. Many studies organize various use cases using some common methods like topic modelling, sentiment, and co-word analysis with different sources of dataset such as social media, news or documents. In terms of method category, some researchers prefer to classify it as NLP techniques [1], [2]. Basically, the idea behind this technique is turning unstructured data into structured data in the form of numbers so that mathematical and statistical algorithms can be applied [3]. This is due to the fact that utilizing unstructured textual data for analysis is important in many disciplines. Some studies rely on statistical data or time series in many cases for forecasting, decision support, monitoring, analysis tools and more. However, it is definitely requiring extensive programming knowledge and skills that many researchers lack – especially in the social sciences, humanities, and applied disciplines [4], [5].

As one of the popular methods, topic modeling is used to identify topics from a set of documents [6]; It is a generative probabilistic approach that groups words into topics and compresses a corpus of thousands of documents into a concise summary while capturing the most prevalent subjects in the corpus [7]. Topic modeling categorized as unsupervised learning algorithm where no prior annotation and labelling requires for document because the topics itself emerge from the analysis of the original texts [8]. We put a quote of topic modelling definition generally according to [9] based on their in-depth analysis of topic modeling evolution. In short, we can say that documents are a mixture of topics, whereas topics are a mixture of words [10].

"We define a topic model to be an unsupervised mathematical model that takes as input a set of documents $\mathcal{D}$, and returns a set of topics $\mathcal{T}$ that represent the content of $\mathcal{D}$ in an accurate and coherent manner"

The availability of modern data gathering tools, big data, and complex analysis giving new challenges for research. Moreover, the same data with different approaches affect output validity because varied methodological work-flows can reverse conclusions [11]. Bridging this gap is necessary because data is the research backbone. Visualizing textual data with interactive dashboard and open-source language like R, python, and Quarto potentially open up concurrent benefits in communication between those who analyze data and other disciplines. So, qualitative text can be transformed into quantitative without technical burden where no barrier of programming skill for end-user. According to [4], applying Shiny for text analysis suits for many purposes like epidemic, education, and policy therefore a collaborative benefit worth attention from the wider community. To support accessibility and reproducibility goals, they developed an interactive R Shiny application named as Text Analysis for All (TALL) for exploring, modelling, and visualizing textual data.

Professionals and academics have been contributing to the development of shiny web applications since it was conceptualized in July 2012. A study conducted by [11] revealed that this popular web application framework of R has been utilized in varied research areas. Meanwhile, the use of R shiny in public sectors has become the vast majority of interactive web application development in the domain of social science. It can be seen from the showcases built by

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

government agencies, statistical offices, and research institutes around the world [12]. Most of these exhibited applications using R and Python utilized by the government to answer questions their citizens care about.

In this study, we mainly focused on developing a Shiny web app for textual data analysis tools related to previous work [13]. With enormous textual data published by mainstream media, it can be harnessed to visualize public policy narrative or specific policies using topic models. What we want to achieve in this current study is to transform methods of topic modelling from prior analysis into an interactive web application through integrated steps of workflows. The application is designed to consist of different capabilities such as import data, preprocessing, term frequency, topic modelling, topic coherence, visualization, result export, and help. This enables researchers to interact with ShinyTopic and reproduce the same data with various workflows to generate research insight. Therefore, those with no programming backgrounds can use it for different data use cases based on their specialization.

Literature reviews conducted by searching research articles that designed similar analysis tools for public policy narrative. According to [14], policy narrative means describing and constructing a story with characters, a plot and a moral based on reality in particular policy issue. In Indonesian Journals, topic modelling of public policy assessment was conducted in several papers. A study by [15] investigates public discourse on Indonesia's Free Nutritious Food (MBG) from Youtube comments, whereas, another article [16] focuses on Indonesia food security news extraction using LDA and pyLDAvis for visualization. In addition, [17] also conducted a static public policy analysis from online news media within the period of 2014 to 2019 in the context of infrastructure development. Both LDA and pyLDAvis are used for topic modelling and visualization.

In relation with interactive app, [18] proposed LDAShiny to review scientific literature using LDA. They designed the application with R Shiny packages using LDA as the main method to extract topics from published literature related to *Oreochromis niloticus.* The latest literature shows us the application designed by [4], they built a complete topic modelling application with R Shiny using LDA as model. Moreover, [19] also developed qualitative data tools called RCualiText which combine AI coding and manual coding. To get the result, LLM was used to validate the semantic analysis. Lastly, we found another application called RollingLDA to track research topics with continuous growing data [2].

Reflected to our study, we introduce interactive ShinyTopic for topic modelling by extending public policy narrative in media. LDA includes as probabilistic model where CTM use to strengthen topic relationship [20][21]. The application has perceived benefits to wider community collaboratively, tool for reproducible research, and enables researchers to extend it for various data subject or case and non-programmer friendly.

ShinyTopic entirely build locally with R Studio and accessible via standard browser while Interactive GUI supported by R packages which available and can downloaded freely. Most important thing is, one of the packages called R Shiny has main capability to form a web layout [22]. Shiny make developers easy to architect web interface with minimum complication. Creating web prototype is now faster and easier without needed common web technologies like HTML, CSS, JavaScript and so on unless used for advance and far beyond shiny basic can provide.

In the architecture itself, R shiny made of single script called R.app which separating between UI and computational backend Server [23]. UI perform object control the layout and appearance while server function contains the instruction that need to build the app [4], [11], [22]. In order to host R shiny web application, we can publish it in www.shinyapps.io, a platform for deploying shiny app which need an infrastructure like physical or virtual machine. However, security concern should be considered firmly. For those who are novice to the shiny or web technology more likely make critical mistake if they not familiar with cybersecurity. So, relying this into expert in the field is more wise choice.

With the current architecture, not many studies mentioned about its limitation. As stated by [11], some issues noted in the previous article as follows. First of all, faster internet connection is needed when dealing with large dataset. More seriously, packages update in R can occur without warning, consequently may crash application. Another minor issue related to dashboard; it is not flexible when building complex application compare to other programming language like Java. This issue then addressed with creation of modularization. But, when it comes to a more complex application, it said that harder to add additional components in the future. In the other hand, a reactive programming paradigm may consider as non-trivial concept for some beginners. All of the stated technical issues could be fixed in the future release of RStudio or shiny core developer.

2. RESEARCH METHODOLOGY

The ShinyTopic comprises text mining techniques and analysis that enables users to obtain related topics from document collections without programming and coding needed. The users must prepare the data by themselves before begin text analysis because document collections is not part of the application. For demonstration purposes, 200 news text were used by aggregating it from various credible outlets such as Detik.com, Tempo.co, Kompas.com, and so on related to public policy of Indonesian Government. Prior data were collected through a web-scraping process implemented in Python using the "newspaper3k" library. Each URL was passed to the "Article" class to create an article object for subsequent extraction. To obtain approximately 200 relevant news articles, specific keywords were employed as search parameters in Google News. The keywords "Prabowo" and "Gibran" and "policy/kebijakan" were selected to identify and aggregate news articles related to the research topic and serve as the primary data sources [13]. The data collection process began

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

by creating an empty Python list to store successfully retrieved articles, with each article represented as a dictionary containing the extracted attributes. The URLs were then processed sequentially within a loop. For each URL, the corresponding article object downloaded the HTML content and parsed the webpage to extract relevant information, including the title, article text, publication date, and other metadata. The "newspaper3k" library automatically excludes non-content elements, such as advertisements, sidebars, footers, and navigation links, from the extracted article text. Nevertheless, the completeness of the extracted content was carefully considered because some news articles may be distributed across multiple webpages. After successful parsing, the title, publication date, URL, and article text were appended to the data collection list. During the experiment, however, "newspaper3k" was unable to consistently identify the publication date, resulting in missing values for this attribute in some records. Therefore, to prioritize data quality and ensure the accuracy and completeness of the dataset, the publication dates had to be manually coded. The collected list of dictionaries was subsequently converted into a structured data frame using the "pandas" library. The resulting data frame was inspected through its initial records to verify the consistency of the columns and rows with the defined attributes. Once the data structure and extracted fields were confirmed, the dataset was prepared for further processing and analysis using ShinyTopic.

To run ShinyTopic, user must prepare RStudio in their working environment and download app.R file from repository https://github.com/lukmanulhakim093/ShinyTopic.git or run the application directly from RStudio with the command `shiny::runGitHub('ShinyTopic', 'your_github_account_name')`. A publish version ready also available in www.shinyapps.io based on request. At initial step, we can choose the input data into "upload" tab name. A file must follow the correct formatting guidelines in *.csv extension. Users can choose to include column header when upload data so that it appears in the "preview data" output. The column name should at least contain one column and text using UTF-8 encoding format. There are also statistical summary of data underneath upload and preview data boxes that shows total documents/rows, number of columns, and average of character lengths for each row. The complete topic modelling processes display in Figure 1. Users are directed to upload data page when the application running for the first time.

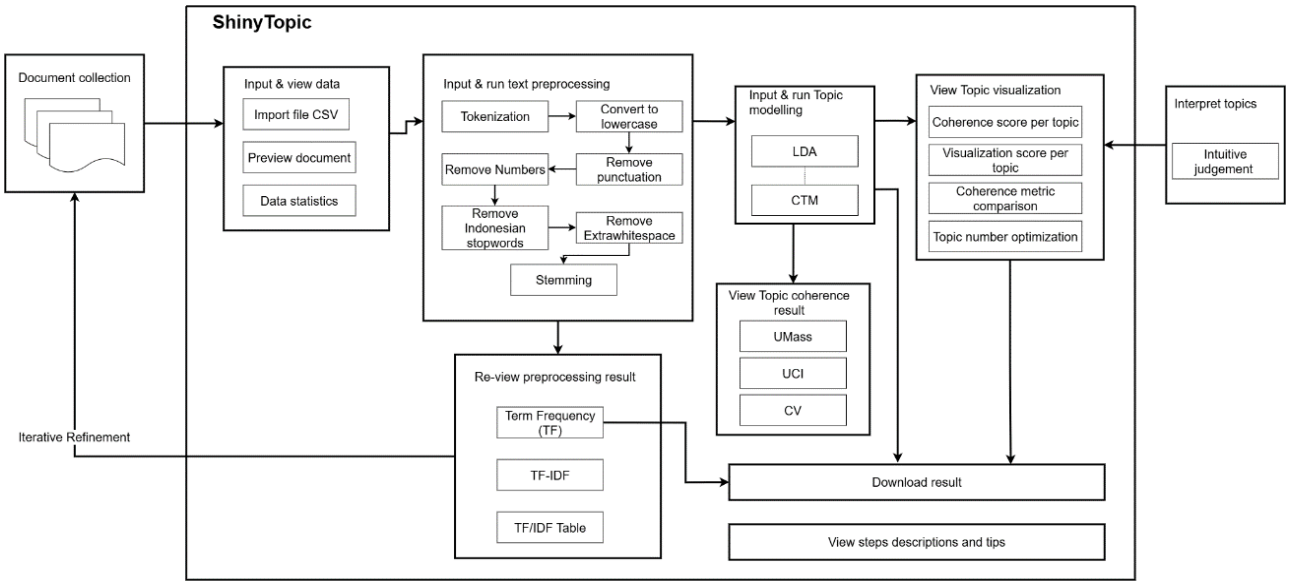

Figure 1. Research methodology and application scope

2.1 Text Preprocessing

Text pre-processing is part of text mining techniques. it's begun with tokenization, converting text into lower case, remove stop-words, special characters, and Indonesian stop-words which is specifically require for our case. We can also perform stemming (optional) which is part of text normalization by erasing prefix and suffix that mostly use in Indonesian textual data [3], [13] and more data cleaning steps available. However, there is no exact sequence of data cleaning steps so far; but it is depending on data format and language structure. Steps such as word contraction and remove accented characters considered important for English language [1] and suitable for Indonesian text. Whereas, a study by [24] put that abbreviation not relevant for their study, then should remove at this step. In contrast, [3] stated that scholars replace abbreviation into text format, so the terminology could be fully spelled. In addition, removing single character and decreasing occurrence frequency done by [25] to include only tokens corresponding to their study and manually refine the remaining tokens. Due to those reasons, we suggest the end-users to measure their data and fit their need when using ShinyTopic.

2.2 Term Frequency – Inverse Document Frequency (TF-IDF)

ShinyTopic includes term frequency and inverse document frequency as sequence of topic modelling process. At this step, users can inspect and choose between term frequency or TF-IDF. Based on the result, they can do iterative refinement

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

by removing infrequent words or observe word's weigh. The following article [3] mentioned that scholars used to spot word frequencies and discriminative words to reduce document-term matrix (DTM) while [24] implemented this on his research for information retrieval based on the importance of a keyword. Term Frequencies (TF) measure how frequently a term t occurs in document d. Where $f_{t,d}$ denotes the number of occurrences of term t in document d. A higher TF value indicates that the term occurs more frequently in the document and may therefore be more representative of its content. It is defined as formula (1). Inverse Document Frequency (IDF) measures the degree of specificity of a term across the entire document collection. It is calculated as formula (2) where *N* represents the total number of documents in the corpus, while df_t denotes the number of documents containing term *t*. A term that appears in many documents obtains a relatively low IDF value, whereas a term occurring in only a small number of documents receives a higher IDF value. To calculate the TF-IDF weight, it is obtained by multiplying the TF and IDF values as defined in formula (3).

$$TF(t,d) = f_{t,d} \tag{1}$$

$$\mathrm{IDF(t)} = \log\left(\frac{\mathrm{N}}{\mathrm{df_t}}\right) \tag{2}$$

$$TF - \mathrm{IDF(t,d)} = \mathrm{f_{t,d}} \times \log\left(\frac{\mathrm{N}}{\mathrm{df_t}}\right) \tag{3}$$

2.3 Topic Modelling

In the context of ShinyTopic, 2 topic models are available such as Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM). LDA is one of the best topic modelling techniques used for text summarization and information retrieval. In the other hand, there is limitation on LDA model that put topic's with no correlation each other where CTM address this issue [21]. A study by [9] mentioned that, CTM replaced the Dirichlet distribution with a logistic normal distribution and add a covariance matrix between topics to model correlation. Based on the prior statement, ShinyTopic give a choice to users whether they start with LDA as starting point and perform CTM for comparison. Moreover, ShinyTopic adopt an iterative approach where users can specify the number of iteration and topics based on their judgement and tune appropriate number of topics as per topic coherence score in the next step.

2.4 Topic Coherence

As mentioned also by [9] in their research article, topic coherence is important to measure high quality topics where topic model itself not sufficient to guarantee topic noisy and intelligible. For that reason, topic coherence also included in ShinyTopic to provide scores of optimal topic number. Pointwise Mutual Information (PMI) is one of the common coherence metrics that attempt to measure the closeness of words in each topic based on their relative cofrequencies with each other. ShinyTopic includes variants of PMI such as UCI and Cv. UCI metric based on PMI used to measure the strength association between word pairs based on their co-occurrence within a sliding window of length l [26]. Where Cv is a combination of cosine similarity and normalized PM under a Boolean sliding window to capture proximity between words as well as the mutual information and vector similarity [9]. Another metric called UMass used to evaluate the quality of a topic based on the co-occurrence of top-ranked words within the same document means that higher degree of corresponds to greater topic coherence [26]. Sara et al [27] provide a tutorial in their paper how to select and label the number of topics. Determining the number of topics is challenging, therefore still require human coding.

2.5 Topic Visualization

Within visualization menu, topic models visualized into graphical and numerical result. Users can observe wordcloud per topic, top ten word per topic, document distribution per topic, topic comparison between two topics, and heatmap of topic-word. User must interpret and label each topic based on their intuitive and judgement because there are no pre-defined categories in unsupervised learning method like LDA or CTM [13].

2.6 Result and Help

Under result and help menu, user then able to conclude the result of text analysis for every topic throughout sequence steps in ShinyTopic. It shows the summary of topic assignment per documents and which topics are dominant in every documents. The downloadable button available to export Document Topics, Topic terms and coherence score in CSV format. User guideline available to explain the goal of each phase and some tips in order to gain better result, optimal topic coherence, scoring threshold, and additional library for more accurate stem.

3. RESULT AND DISCUSSION

For For demonstration purpose, we used 200 document collections from prior study to simulate a proper handling of topic modelling illustration. In this section, we upload CSV file into upload tab to get data statistic dan data preview as show as Figure 2. The data consists of id and text columns where the "id" used to specify news articles and "text" is textual content to be analyze. From the data preview box, we can set how many data records to be display in the preview table and searching each or consecutive word in every news articles. Thus, we can eliminate the non-significant words from preliminary stage. We can also inspect length of characters in each row.

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

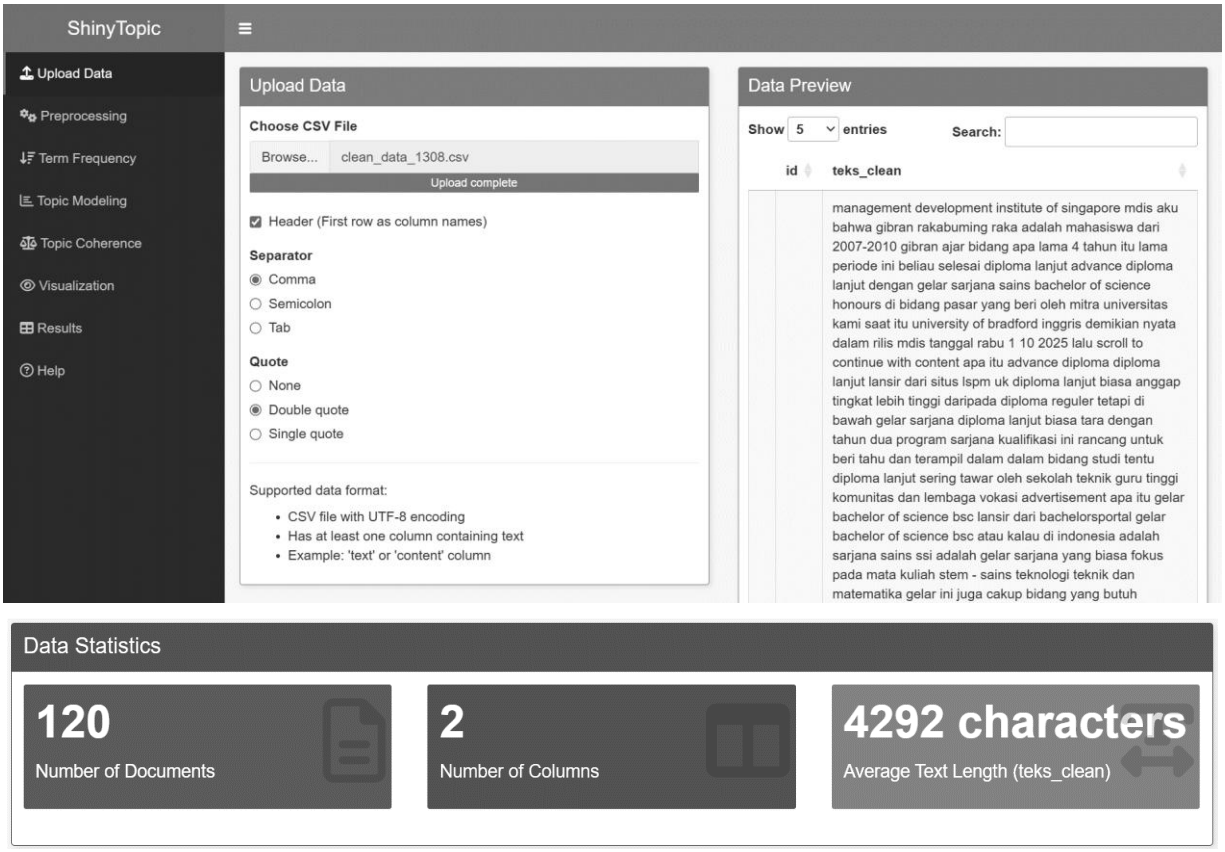

Figure 2. Data upload and summary

Next section, preprocessing stage done by choosing selective options to clean data such as lowercase conversion, punctuation removal, removing numbers, Indonesian stopwords, decreasing extra spaces, and more steps available for data cleaning. The sequence step depicted in Figure 3.

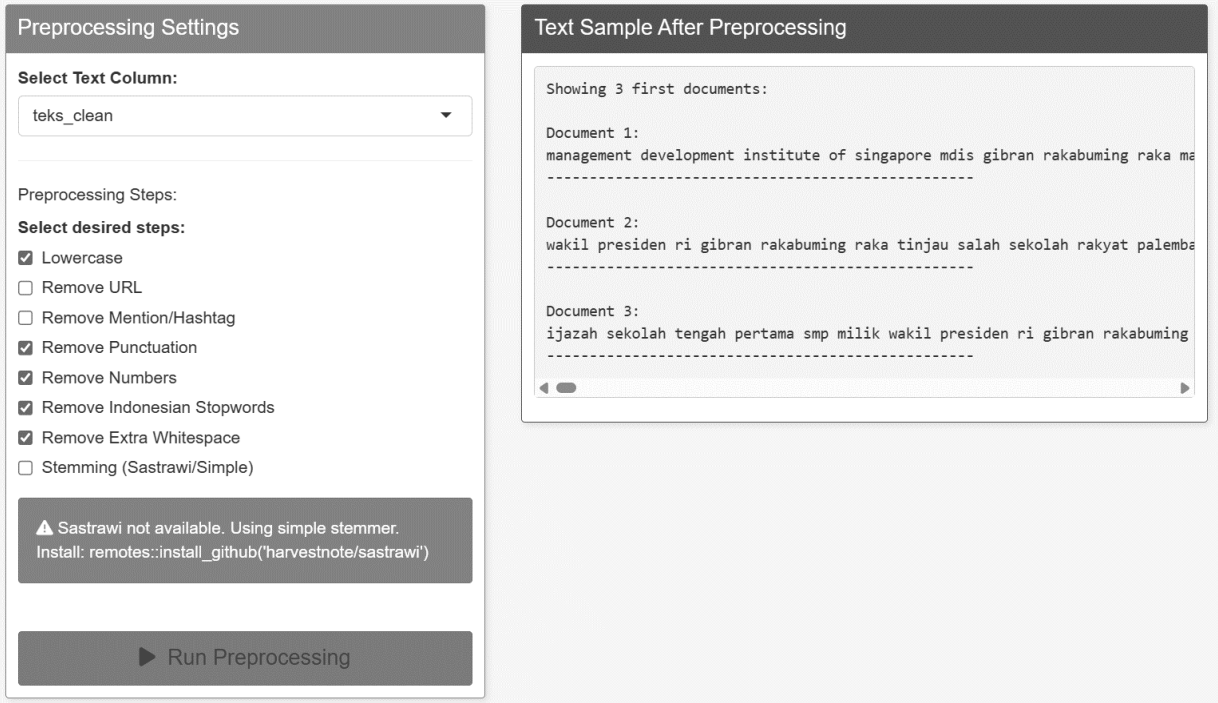

Figure 3. Preprocessing steps

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

We can validate each document of data testing by comparing the number of documents, total column and characters average for each document before and after preprocessing. Going further, next things to do is analyzing Term Frequency and Inverse Document Frequency by measuring total weight of TF-IDF. Term frequencies obtain from total appearance of terms T in documents D and total document D containing terms T. TF-IDF weight scores calculated based on formulas as attached in subchapter 2.2. If terms exist in every document, meaning that terms are not discriminating or valueless as shown in the Figure 4. Scholars tend to remove infrequent or weight words [3].

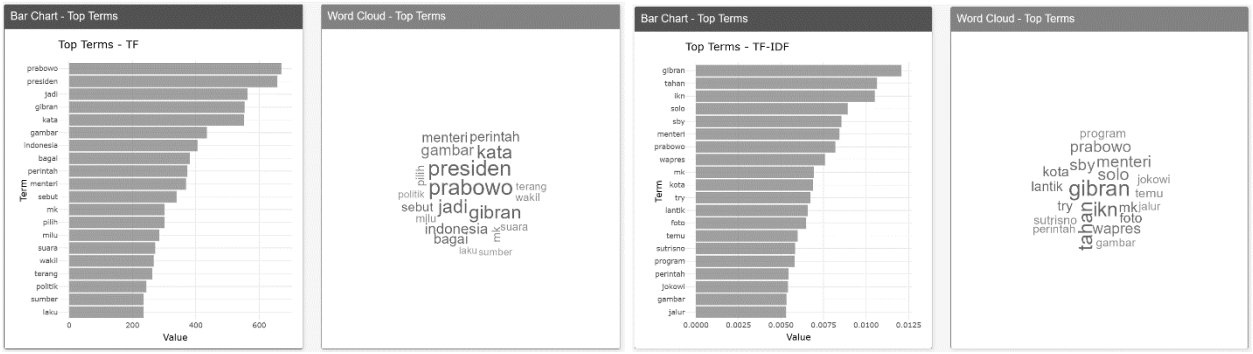

Figure 4. Term Frequency and Inverse Document Frequency bar chart dan world cloud

Moving forward, we can now apply text clustering algorithm by selecting between LDA (Latent Dirichlet Allocation) and CTM (Correlated Topic Model). Users must determine the number of topics based on their own judgement or tune to fit topic modelling needs as shown in figure 5. To get topic summary, built-in Gibbs's sampling is embedded for LDA. The number of iterations is customizable and the default random seeds value set to 1234. For burn-in parameter, we configured it not exceed 200 and thin parameter set to 100. It is suggested to increase iteration value for more precise output but it will take longer and more loads.

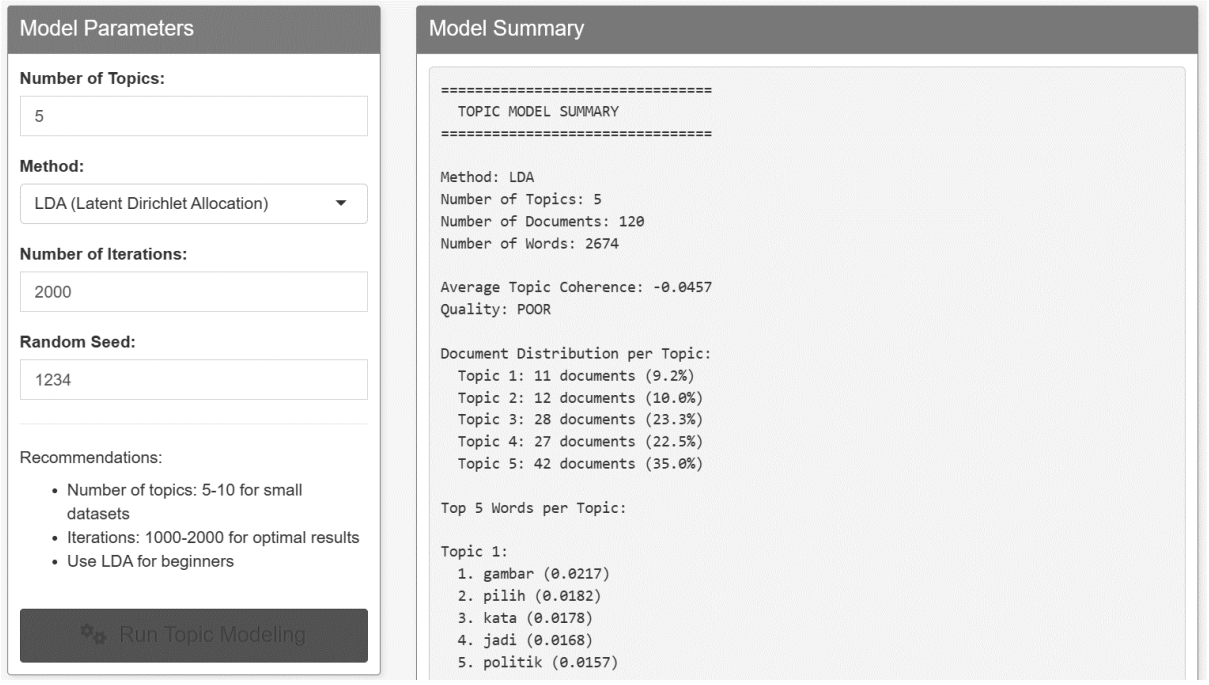

Figure 5. Topic Modelling analysis

Beside of human annotation, an automated topic coherence metrics was employed in ShinyTopic to capture word co-occurrence. Users should not rely on coverage and coherence evaluation because it is not sufficient to judge a topic model [9]. Consequently, qualitative evaluation should be part of topic modelling analysis during experiments. Initially, LDA have been evaluated using likelihood and perplexity since it shows a negative correlation compare to human judgement. For that reason, a study then proposed topic coherence metric such as UCI, UMass, NPMI, and Cv which more align with human interpretation [26]. The Figure 6, topic coherence and metric comparison visualize in line and radar plot. According to our experiment, UCI showing us that the highest score achieves by "topic 2" and the lowest score pointed to "topic 3". Meanwhile "topic 4" and "topic 5" ranked in the middle with slightly higher scores compare to "topic 3". The other metrics show similar pattern although the score is lower than UCI.

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

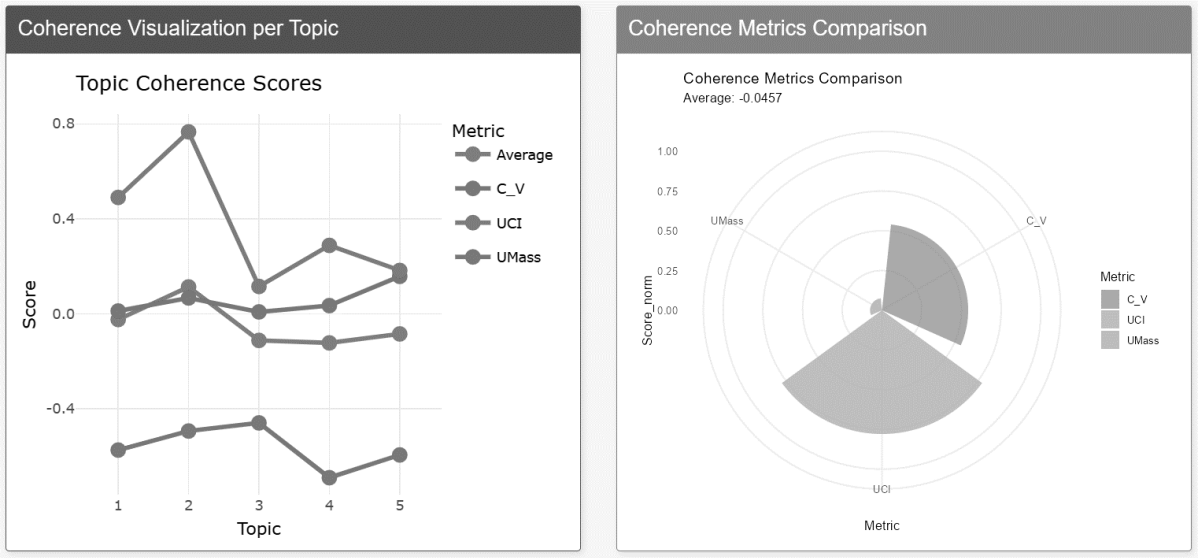

Figure 6. Topic Coherence

Last but not least, users able to retrieve topic word-cloud, top ten terms per topic, document distribution per topic, topic comparison and heatmap between topics and terms as represented in Figure 7.

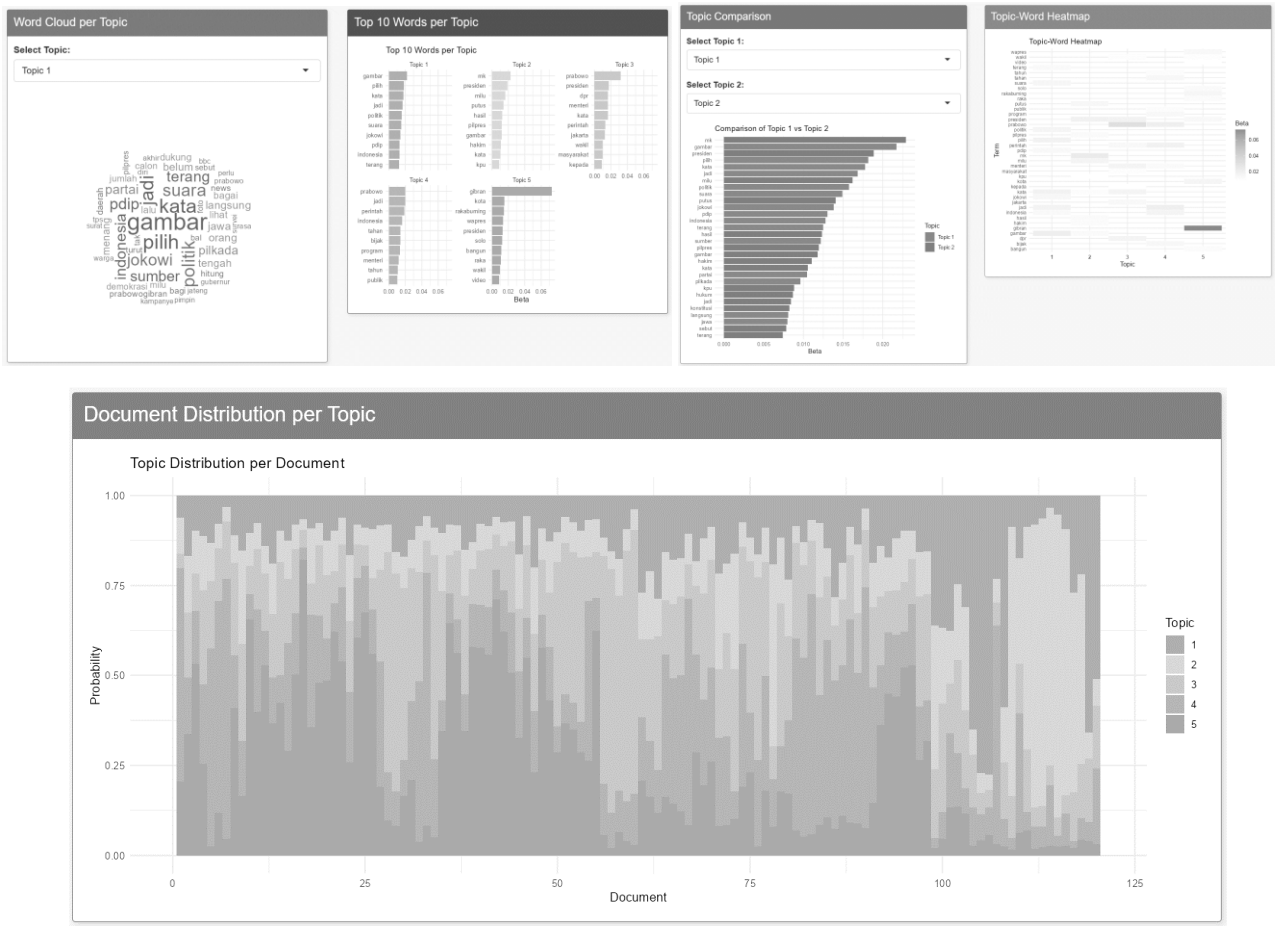

Figure 7 Topic Visualization

Finally, there is a list of which topic assign to document ID and what topic is contributed dominantly into the document as shown in figure 8. Underneath to a list, we can download each result independently in CSV format attached in figure 9.

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

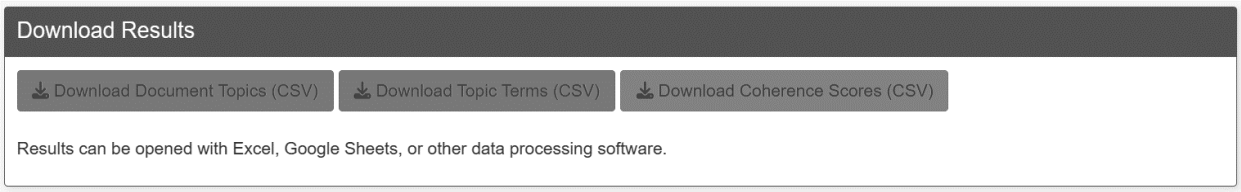

Figure 8 Download result

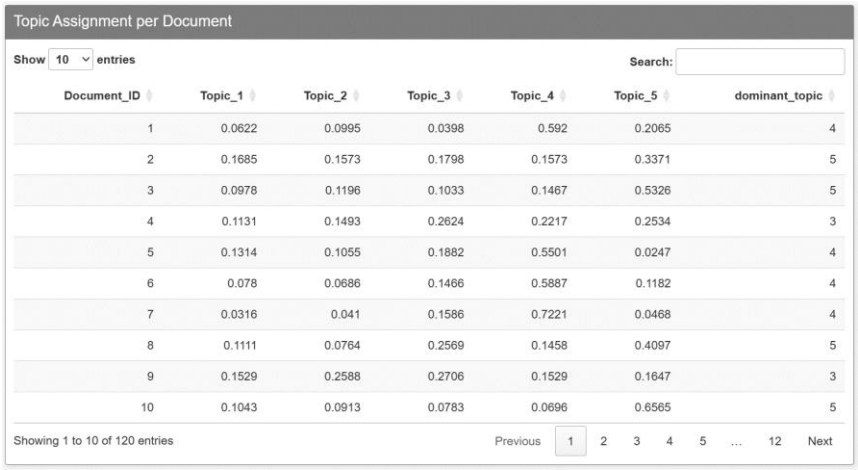

Figure 9. Summary of topic assignment

To demonstrate the implementation of interactive topic modeling, we present a public policy narrative case study that simplifies methodological complexity into an integrated application. The methodologies and analytical procedures adopted from previous studies were transformed into an integrated workflow within ShinyTopic. Using the same dataset, the proposed approach produces higher probability (β) scores for several topic terms, although preprocessing steps such as n-gram construction and stemming using Sastrawi were not incorporated into the current methodology. Table 1 presents five topics related to public policy narratives identified in the previous study, while Table 2 presents the results obtained from the current study. Both sets of topics remain interpretable; however, the topics presented in Table 2 were not selected for further analysis because only five topics were extracted from the corpus.

Table 1. LDA topic cluster from previous study [13]

Keywords based on topics	Topic labelling
Topic 3: 0.057*"gibran" + 0.031*"bangun" + 0.023*"kota_solo" + 0.019*"solo" + 0.016*"program" + 0.011*"persen" + 0.011*"wali_kota_solo" + 0.010*"masyarakat" + 0.010*"proyek" + 0.009*"kerja"	Gibran program, project for Solo city
Topic 9: 0.020*"gibran" + 0.018*"negara" + 0.015*"milik" + 0.010*"kait" + 0.009*"tni" + 0.009*"lihat" + 0.009*"pres" + 0.009*"189rabowo" + 0.008*"wapres" + 0.008*"usul"	Gibran and military
Topic 11: 0.031*"bijak" + 0.014*"masyarakat" + 0.013*"negara" + 0.012*"program" + 0.012*"189rabow" + 0.010*"kerja" + 0.008*"jalan" + 0.008*"putus" + 0.007*"hadap" + 0.007*"pulau"	Topic related to Island and national program
Topic 12: 0.105*"ikn" + 0.044*"bangun" + 0.040*"wapres" + 0.037*"rencana" + 0.031*"selesai" + 0.029*"wapres_gibran" + 0.027*"gibran" + 0.025*"kantor" + 0.025*"kota_nusantara_ikn" + 0.022*"usul"	Gibran new office at Indonesia new capital city
Topic 14: 0.051*"tahan" + 0.035*"189rabowo_tahan" + 0.020*"negara" + 0.018*"189rabowo189_tahan" + 0.016*"menhan" + 0.015*"alutsista" + 0.013*"kuat_tahan" + 0.012*"angkat" + 0.012*"tni" + 0.011*"189rabowo_tahan_prabowo_subianto"	Prabowo, military equipment, and defense industry

Table 2 Five topics *k* in the current analysis

Keywords based on topics	Topic labelling
Topic 1: gambar (0.0201), jadi (0.0197), kata(0.0173),politik(0.0165),pilih(0.0158)	Public election and political party

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

Keywords based on topics	Topic labelling
Topic 2: mk(0.0210),milu(0.0175),presiden(0.0136),gambar(0.0121),putus(0.0116),suara(0.0116)	Constitutional court of the Republic of Indonesia, vice president candidacy, and court decision
Topic 3: perintah(0.0197),masyarakat(0.0176),indonesia(0.0167),bijak(0.0143),dpr(0.0125)	Related to people representative, policy, public and government
Topic 4: gibran(0.0752),kota(0.0170),rakabuming(0.0139),wapres(0.0139),solo(0.0135)	Gibran and solo city and his candidacy as vice president of Indonesia
Topic 5: prabowo(0.0156),presiden(0.0393),menteri(0.0337),tahan(0.0160),subianto(0.0146)	Related to Prabowo's position as defense ministry and defense

3.1 Discussion

An in-depth comparison of the empirical outputs from the previous study and the current study reveals major shifts in quantitative performance, semantic granularity, and lexical distribution, despite utilizing the same underlying national news corpus. From a quantitative standpoint, the two studies show contrasting behavior across different topic modeling algorithms and evaluation metrics. In the previous study, the baseline Latent Dirichlet Allocation (LDA) model yielded a relatively low coherence score of 0.3709, which prompted the researchers to implement Non-negative Matrix Factorization (NMF) to achieve a more interpretable and higher coherence score of 0.6826. In contrast, the current study, utilizing the LDA algorithm under a fixed $k = 5$ topic constraint, achieved a significantly high local coherence score of 0.7663 using the UCI metric, although the overall average coherence remained relatively low. Furthermore, while the Correlated Topic Model (CTM) was introduced in the current study to capture topic relationships, it yielded lower coherence scores than the current LDA model. These findings demonstrate that although NMF previously offered superior performance over standard LDA, iterative parameter tuning and alternative evaluation metrics, such as UCI, can enable LDA to capture highly coherent word associations.

Semantically, the choice of topic granularity k substantially altered the scope of the public policy narratives captured. By scanning a wider range of $k = 5$ to 20 and qualitatively selecting five specific clusters, the previous study mapped highly specialized, micro-level issues, such as local infrastructure projects in Solo "topic 3", Gibran's interactions with the military "topic 9", and the logistical relocation of the Vice President's office to the New Capital City (IKN) "topic 12". Conversely, the direct extraction of $k = 5$ topics in the current study captured broader, macro-level structural narratives of Indonesian politics. These include national election dynamics "topic 1", Constitutional Court (MK) disputes concerning vice-presidential candidacy "topic 2", and institutional legislative–executive checks and balances between the DPR and the government "topic 3". This shift suggests that a smaller, fixed k-value can be effective for distilling high-level constitutional and electoral policy narratives, whereas a larger k-value may be necessary to isolate more localized and issue-specific development policies.

Finally, differences in text-preprocessing strategies directly influenced the word-probability distributions β. The previous study employed intensive n-gram construction, which preserved the contextual meaning of compound terms such as kota_nusantara_ikn ($\beta = 0.025$) and menteri_tahan_prabowo_subianto ($\beta = 0.011$). In the current study, the omission of n-grams resulted in a basic bag-of-words representation. Although this approach reduced the semantic specificity provided by multi-word expressions, it increased the probability of individual core tokens. This pattern is particularly evident in the representation of key political actors: the term "gibran" in "topic 4" of the current study reached a peak probability of 0.075210, compared with 0.057 in "topic 3" of the previous study. Consequently, while the bag-of-words representation in the current study provides less semantic specificity than the n-gram approach, it generates a more focused and readable thematic structure, with stronger emphasis on prominent political actors.

4. CONCLUSION

In conclusion, this study developed ShinyTopic as an interactive topic modeling application that integrates the major stages of the topic modeling process within an R Shiny environment, with the aim of simplifying methodological complexity and making topic modeling more accessible to users with different levels of programming and text-mining expertise. The application also demonstrates that R and its extensive package ecosystem can provide a flexible computational environment for integrating statistical, text-mining, and visualization procedures into a single application. The public policy narrative case study further demonstrates the applicability of ShinyTopic for extracting interpretable topics from a textual corpus and examining topic-term probabilities. The results show that the application can produce meaningful topic representations even when several advanced linguistic preprocessing procedures are not yet incorporated. However, the study also identifies several linguistic and computational aspects that require further

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

development. In particular, the current bag-of-words representation may treat compound expressions such as "open innovation" and "innovation culture" as separate tokens, potentially reducing their contextual meaning. The integration of n-gram techniques is therefore important for preserving meaningful multi-word expressions during preprocessing. In addition, Indonesian text can generate high-dimensional document-term matrices because of morphological variations, making stemming and lemmatization important for reducing vocabulary dimensionality. Although ShinyTopic currently provides a basic stemming procedure, integrating the Sastrawi stemming algorithm would improve Indonesian-language preprocessing; because an official Sastrawi implementation is not currently available as an R package, interoperability with Python may be considered in future development. Further improvements should also include abbreviation normalization, Named Entity Recognition (NER), and Part-of-Speech (POS) tagging to provide a more linguistically informed representation of the corpus. Overall, the study demonstrates that ShinyTopic provides a practical foundation for interactive topic modeling while establishing a clear direction for improving linguistic preprocessing, analytical accuracy, and the general applicability of the application.

REFERENCES

[1] P. Ghasiya and K. Okamura, "Investigating COVID-19 News across Four Nations: A Topic Modeling and Sentiment Analysis Approach," *IEEE Access*, vol. 9, pp. 36645–36656, 2021, doi: 10.1109/ACCESS.2021.3062875.

[2] A. Bittermann and J. Rieger, "Finding Scientific Topics in Continuously Growing Text Corpora," in *Proceedings ofthe Third Workshop on Scholarly Document Processing*, Oct. 2022, pp. 7–18.

[3] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "The application of text mining methods in innovation research: current state, evolution patterns, and development priorities," in *R and D Management*, Blackwell Publishing Ltd, Jun. 2020, pp. 329–351. doi: 10.1111/radm.12408.

[4] M. Aria, M. Spano, L. D'Aniello, C. Cuccurullo, and M. Misuraca, "TALL: Text analysis for all–an interactive R-shiny application for exploring, modeling, and visualizing textual data," *SoftwareX*, vol. 34, Jun. 2026, doi: 10.1016/j.softx.2026.102590.

[5] A. Bleichrodt, A. Phan, R. Luo, A. Kirpich, and G. Chowell, "StatModPredict: A user-friendly R-Shiny interface for fitting and forecasting with statistical models," *PLoS One*, vol. 20, no. 8 August, Aug. 2025, doi: 10.1371/journal.pone.0329791.

[6] R. Debnath and R. Bardhan, "India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling," *PLoS One*, vol. 15, no. 9 September, Sep. 2020, doi: 10.1371/journal.pone.0238972.

[7] E. Mayor and A. Miani, "A topic models analysis of the news coverage of the Omicron variant in the United Kingdom press," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12889-023-16444-7.

[8] J. De la Hoz-M, M. J. Fernández-Gómez, and S. Mendes, "Ldashiny: An r package for exploratory review of scientific literature based on a bayesian probabilistic model and machine learning tools," *Mathematics*, vol. 9, no. 14, Jul. 2021, doi: 10.3390/math9141671.

[9] R. Churchill and L. Singh, "The Evolution of Topic Modeling," *ACM Comput. Surv.*, vol. 54, no. 10, Jan. 2022, doi: 10.1145/3507900.

[10] U. Chauhan and A. Shah, "Topic Modeling Using Latent Dirichlet allocation: A Survey," *ACM Comput. Surv.*, vol. 54, no. 7, Sep. 2022, doi: 10.1145/3462478.

[11] P. Kasprzak, L. Mitchell, O. Kravchuk, and A. Timmins, "Six Years of Shiny in Research: Collaborative Development of Web Tools in R," 2020. [Online]. Available: https://digitalcommons.unl.edu/r-journal/144

[12] Jeremy Allen, "Public Sector Shiny Showcase: Government Agencies Worldwide Building Data Science with R, Python, and Shiny," https://posit.co/blog/public-sector-shiny-showcase-government-agencies-worldwide-building-data-science-r-python.

[13] L. Hakim, A. Y. Aditya, and M. Lubis, "Topic modelling analysis of public policy narratives on prabowo-gibran in national news," *Journal of Soft Computing Exploration*, vol. 6, no. 4, pp. 266–273, Dec. 2025, doi: 10.52465/joscex.v6i4.632.

[14] R. Mu, Y. Li, and T. Cui, "Policy narrative, policy understanding and policy support intention: a survey experiment on energy conservation," *Policy Studies*, vol. 43, no. 6, pp. 1361–1381, 2022, doi: 10.1080/01442872.2021.1954609.

[15] C. Suhaeni, L. Nissa, A. Mualifah, and H. Wijayanto, "LDA Topic Modeling Analysis of Public Discourse on Indonesia's Free Nutritious Meals Program (MBG)," *International Journal on Informatics for Development*, vol. 14, no. 1, pp. 587–600, 2025, doi: 10.14421/ijid.2025.5211.

[16] A. Afiyati, I. Rochmad, S. Budiyanto, B. Jokonowo, H. Santoso, and K. Budiana, "Topic Modeling Analysis of Indonesia Food-Security News: Methods,Interpretations, and Trend Insights," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 25, no. 2, pp. 335–344, Mar. 2026, doi: 10.30812/matrik.v25i2.5784.

[17] A. F. Hidayatullah, M. R. Ma'arif, M. Habibie, and S. Khomsah, "Indonesia Infrastructure Development Topic Discovery on Online News with Latent Dirichlet Allocation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1077, no. 1, p. 012012, Feb. 2021, doi: 10.1088/1757-899x/1077/1/012012.

[18] J. De la Hoz-M, M. J. Fernández-Gómez, and S. Mendes, "Ldashiny: An r package for exploratory review of scientific literature based on a bayesian probabilistic model and machine learning tools," *Mathematics*, vol. 9, no. 14, Jul. 2021, doi: 10.3390/math9141671.

[19] J. Ventura-León and M. Barboza-Palomino, "RCualiText: An Open-Source R/Shiny Web Application for Qualitative Analysis," Feb. 2026. doi: https://doi.org/10.31234/osf.io/46zgx_v1.

[20] S. Syahrial and R. P. F. Afidh, "Fine-Tuning Topic Modelling: A Coherence-Focused Analysis of Correlated Topic Models," *Infolitika Journal of Data Science*, vol. 2, no. 2, pp. 82–87, Nov. 2024, doi: 10.60084/ijds.v2i2.236.

[21] D. M. Blei and J. D. Lafferty, "Correlated Topic Models," in *NIPS'05: Proceedings of the 19th International Conference on Neural Information Processing Systems*, MIT Press, Dec. 2005, pp. 147–154. [Online]. Available: www.jstor.org

[22] N. Satyahadewi and H. Perdana, "Web Application Development for Inferential Statistics using R Shiny," in *1st International Conference on Mathematics and Mathematics Education (ICMMEd 2020)*, Atlantis Press SARL, May 2021. doi: 10.2991/assehr.k.210508.099.

Bulletin of Information System Research (BIOS)
Volume 4, No 3, Agustus 2026, Page: 182−192
ISSN 2963-2455 (media online)
DOI: doi.org/10.62866/bios.v4i3.398,
https://journal.grahamitra.id/index.php/bios

[23] C. Sievert, "Interactive Web-Based Data Visualization with R, plotly, and shiny," 1st Edition., New York, 2020, ch. 34, p. 470. doi: 10.4324/9780429447273.
[24] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.
[25] C. Meaney *et al.*, "Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada," *J. Biomed. Inform.*, vol. 128, Apr. 2022, doi: 10.1016/j.jbi.2022.104034.
[26] C. Yang and Y. S. Kim, "A Context-Driven Topic Coherence Evaluation Approach Using LLM-Based Embeddings," *IEEE Access*, vol. 13, pp. 182257–182275, 2025, doi: 10.1109/ACCESS.2025.3623881.
[27] S. J. Weston, I. Shryock, R. Light, and P. A. Fisher, "Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial," *Adv. Methods Pract. Psychol. Sci.*, vol. 6, no. 2, Apr. 2023, doi: 10.1177/25152459231160105.

