Journal of Education for Pure Science
Vol.15 No. 3 (2025)
DOI: https://doi.org/10.32792/jeps.v15i3.686

Website: jceps.utq.edu.iq *Email: jceps@eps.utq.edu.iq*

Topic Modeling for Web Page using LDA Algorithm and Web Content Mining

Mohammed Ali Mohammed[*,1,] **, Raad Mahmood Mohammed**[2,] **and Hasan Aqeel Abbood**[3,]

[1,2,3] College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq

[*] Corresponding email: mohammed.ali@uoitc.edu.iq

Received 9 / 3 /2025, Accepted 5 / 5 /2025, Published 1 / 9 /2025

Abstract:

Information on Internet and specially on website environment is increasing rapidly day by day and become very huge, this information plays an important role for discovering various knowledge in the Web. So, this web should be choosing the strong and efficient topic title. This paper is aim to generate topic model title for this web using LDA algorithm and based on web content mining. The compared (according to Topic Similarity, Title Length, Topic Coherence, Topic Saturation) between the LDA and NMF algorithms shows that the LDA algorithm is the best algorithm for this approach with our dataset.

Keywords: Topic Model, Web Content Mining, NMF algorithm, LDA algorithm.

1-Introduction Section

Web mining refers to the utilization of data mining methods for extracting valuable insights from web-related data, encompassing web documents, interconnections between documents, and website usage patterns. Its primary objective is to uncover patterns within the vast expanse of the World Wide Web (WWW). Web mining serves as an extension of data mining, incorporating diverse technologies from research domains such as artificial intelligence (AI), statistics, informatics, knowledge discovery, and computational linguistics. The ultimate aim of web mining is to develop algorithms and techniques that enhance the efficiency and convenience of accessing and utilizing web data[1][2]. Web Mining techniques are categorized (as shown in Figure 1) [3] into three classes depend on which part to be mined, which are Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM)[4].

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web. Web usage mining allows the collection of Web access information for Web pages, this information is often gathered automatically into access logs via the Web server. Web content mining involves the extraction of valuable information from diverse sources throughout the World Wide Web. This process focuses on discovering and analyzing data contained within web pages and other web-based resources. Web structure mining, on the other hand, involves analyzing the relationships and connections between web pages. By employing graph theory, it examines the interlinking structure of web pages, either through direct links or through the information web pages share. While Web usage mining aims to extract meaningful patterns and insights from server logs. These logs capture and record user activity on websites, providing valuable information about user behavior and preferences[5][6].

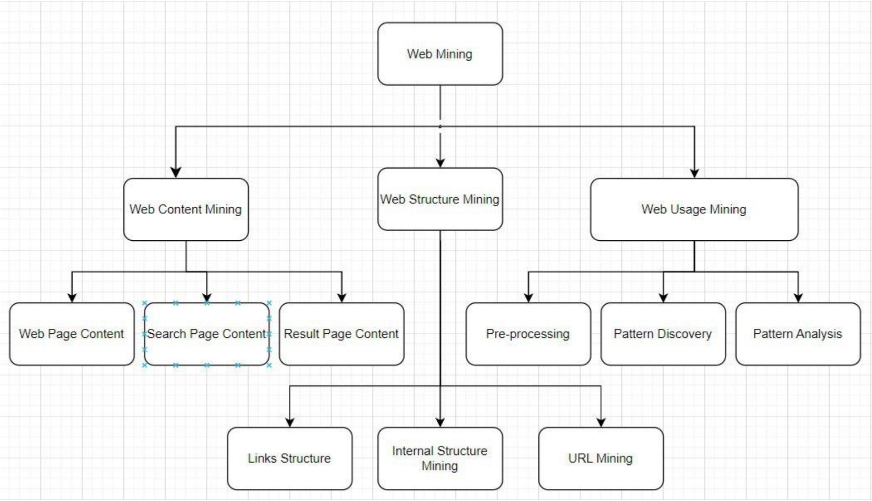

Figure 1: Web Mining Category[3]

Web content mining can be Known as an important procedure of extracting valuable and necessary information from web pages. It is linked to text mining due to the large part of the web content is text-based. Web content mining consider the semi-structured nature of the web. There are two types: the first directly extracts content of documents, the second enhance search result through other tools like search engines. Content mining is used to inspect data gathered by search engines and web spiders. The used technologies in Web Content Mining are Natural Language Processing (NLP) and Information Retrieval (IR) [7]. There are Two approaches that are used in web content mining are database and agent-based approach. The database assists in recapturing the semi-structured data from web documents [8][9]. The three types of agents are customized web agents, information filtering Categorizing agent and intelligent search agents [8]. Customized web agents seek to discover a web page on the user profile. Information filtering categorizing agents work on minimizing the time and effort of the user in finding the location of the relevant document through the Knowledge that is possesses. The filtering agent filters out irrelevant incoming documents and presents to the user only those documents which match the user's interest. Automatically, Intelligent search agents discover information according to a particular query utilizing user-profiles [8][9][10]. The techniques of web content mining showed in figure 2.

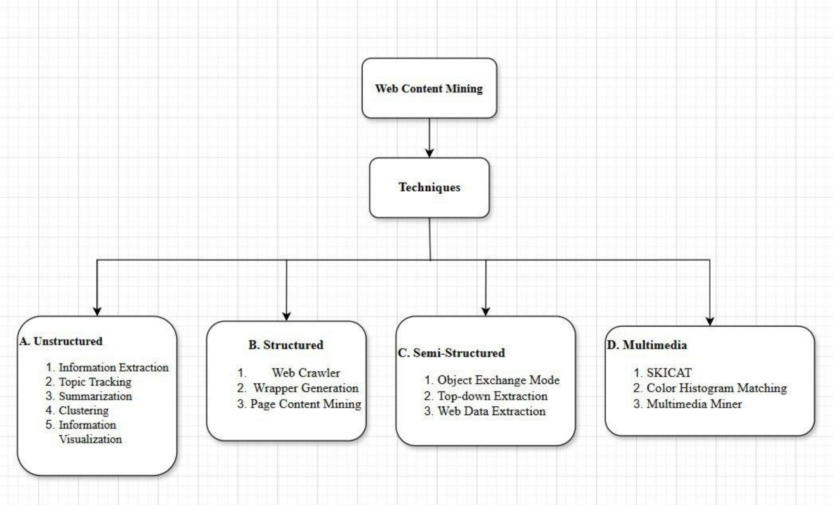

Figure 2: Web Content Mining techniques [10].

The purpose of Topic modeling is to find topics within a collection of digital textual documents. Several fields, many fields, like the Internet of Things, Block-chain, recommender systems, software engineering, digital humanities, and political science, topic modeling can be applied to efficiently discover topics in huge collections of data [11][12]. The benefits of applying that models web content mining audits many applications, like search engine optimization (SEO), web services, web filtering, and digital marketing [13]. Modeling topics in web content is considered as a critical and very important research issue in natural language processing. topic models benefits rely on the quality of out coming term patterns and subjects with high quality. Topic coherence is the primary metric to assess the quality of the currently available topic models.

The remaining sections of the paper are structured as follows: Section 2 presents a comprehensive literature review, examining existing research and studies related to web usage mining. Section 3 focuses on the view of our dataset used and its characteristics. Section 4 explains proposed methodology. Section 5 showed the result and section 6 view the discussion these result. Finally, Section 7 concludes the paper, summarizing the key findings, contributions, and potential future research directions.

2- Related Work Section

Bhat et al. (2019) [3] proposed two variants of Latent Dirichlet allocation (LDA) based on DNN, 2NN Deep LDA, and 3NN Deep LDA. The Bag of Words was input to both the plain LDA and DNN, to achieve better accuracy in the learning of topics. They derived that 2NN Deep LDA takes less time compared to LDA and 3NN Deep LDA.

Yang et al. (2020) [14] introduced a topic model called the Named Entity Topic Model (NETM) to identify factors driving the growth of web content popularity. While NETM outperforms the LDA model in terms of accuracy, it does not consider the HTML structure of webpage content.

In this paper, Youngseok et al. (2021) [16], suggest a web document sorting method by using topic modeling to collect effective information and classification. The suggested processes are applied to the document sorting technique to prevent duplicated crawling when crawling at high speed. By the suggested document ranking technique, it is possible remove repeated documents, and assure that the crawler service is working.

Hamza H.M et al. (2023) [17] conducted a study to introduce a novel topic model designed to learn coherent topics from web content data. They proposed the HTML Topic Model (HTM), which incorporates HTML tags to better understand the structure of web pages. The study involved two sets of experiments to highlight the limitations of existing topic models and evaluate the topic coherence of HTM in comparison to the widely used LDA model and its variants, such as the Correlated Topic Model.

Simra Shahid et al. (2023) [18] provide HyHTM- a Hyperbolic geometry relay on Hierarchical Topic Models that deals with this restriction by combine hierarchical information from hyperbolic geometry to clearly model hierarchies in topic models. the results with four baselines reveal that HyHTM can better care to parent-child relationships among topics.

3- Dataset Used Section

The Author’s own Dataset: the new dataset will be created for proposed system which should contain set of web page content with titles. We use all URLs of the “geeksforgeeks.org” website to create this dataset and saved as .csv file. Table 1 content sample of these dataset. the new dataset with the following properties:

- Id: unique number
- Title: Title of web page
- Content: Content of web page
- URL: URL of the web page.

Table 1: Sample of our dataset

Id	Title	Content	URL
1	Data Structures Tutorial	Page Content as Text	https://www.geeksforgeeks.org/data-structures/
2	Algorithms Tutorial	Page Content as Text	https://www.geeksforgeeks.org/fundamentals-of-algorithms/
3	System Design Tutorial	Page Content as Text	https://www.geeksforgeeks.org/system-design-tutorial/
4	Complete Machine Learning & Data Science Program	Page Content as Text	https://www.geeksforgeeks.org/courses/data-science-live?itm_source=geeksforgeeks&itm_medium=gfg_submenu&itm_campaign=DS_Submenu

4- Methodology Section

New design and implementation of proposed topic model with web content mining system which is used to create topic name form a content of web page. Our system is characterizing with many features such as a new dataset will be create containing a web content mining data type.
Our system consists of the following steps:

1. Create data set as mention in section 3 and saved in Dataset.csv file.

2. preprocessing step: In general, most documents, texts, and articles have several unnecessary (redundant) words like stop-words, slang, misspellings, etc. However, some techniques for text cleaning and preprocessing text datasets are explained in the following (see Fig. 3).

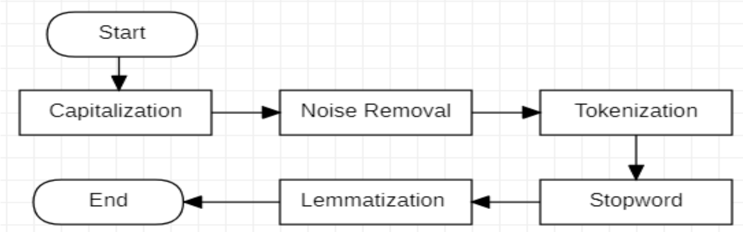

Fig. 3: The utilized techniques of the preprocessing module

a) Capitalization handling: converting all words to upper or lower case letters [19].
b) Noise removal: redundant characters will be remove. [20].
c) Tokenization: text will be breakdown into individual tokens [21], [19].
d) Stop-word removal: removing stop words (such as “is”, “the”, “that”, etc.) addresses from document [22].
e) Lemmatization: The suffixes words that transform into their root or base forms are known as lemmas [23].

3. Bag of words: which keeps a count of the total occurrences of most frequently used words [19].

4. LDA algorithm

LDA is a probabilistic generative model used to identify multiple topics. The key idea is that documents can be viewed as random mixtures of underlying topics, with each topic being characterized by a distribution over words. Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan in 2003, is one of the most popular methods in topic modeling. LDA defines topics based on word probabilities, and the most probable words within each topic typically reveal a clear understanding of the topic, as determined by the word distributions produced by LDA [24]. The proposed methodology used the LDA algorithm in processing step to get the weighted words for the document, LDA is a best choose for topic label compared with NMF according to results.

5. Topic Label

The set of words that generated from topic creating step for each document should be labeling as a main topic field. The Gemini AI will be used. Gemini, a family of highly capable multimodal models developed at Google. the trained Gemini models jointly across image, audio, video, and text data for the purpose of building a model with both strong generalist capabilities across modalities alongside cutting-edge understanding and reasoning performance in each respective domain [25].

The proposed system calls Gemini API with top 5 words for a content of web page as a parameter. The question is: "please, i need the main topic name only in computer science for these words = (5 words)".

The final flowchart of the proposed system shown in figure 4 and the algorithm shown in algorithm 1.

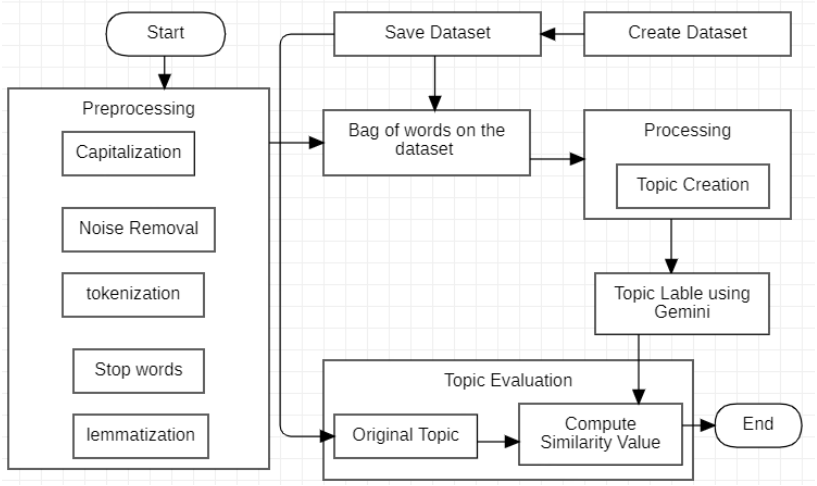

Figure 4: proposed system

```
Algorithm 1: Proposed Methodology
Input:    Document
Output:   Topic_Name
Steps:
1.        Preprocessing Step (Tokenization, Stop words, Lemmatization, Data Cleaning, etc.).
2.        Compute Bag of words.
3.        Processing Topic Creation using LDA algorithm.
4.        Topic Labeling using Gemini with top 5 words for a document
5.        Compute Similarity value between (generate topic, original topic).
```

5- Results Section

As mentioned earlier, the intended system consists of with the following steps. The preprocessing includes cleaning and removing unwanted words. In addition, a lemmatization step is then applied, where the suffix is removed, and many other phases. Then the Bag of words will be computed for dataset. Apply the LDA algorithm (and NMF to compare) to create the topic words. Finally, the Gemini AI will be used to create Topic label according to words. The evaluation step compute similarity value between the original title topic (which saved in database) and generated topic with NMF and LDA. Table 2 shows the result of topics name using NMF algorithm (sample: 5-pages).

Table 2: Topic Name using NMF algorithm

No	Original Topic Page	NMF Words	Gemini Topic Label
1	System Design Tutorial	design, interview, tutorial, data, question	Data Interview Design
2	Machine Learning Tutorial	machine, learning, tutorial, data, python	Machine Learning
3	Web Design Tutorial | Learn Web Design from Basics	design, web, tutorial, graphic, python	Web Design
4	Software Design Patterns Tutorial	design, pattern, link, tutorial, method	Design Patterns
5	What is SEO Optimization or Search Engine Optimization?	search, engine, google, optimization, tutorial	Search Engine Optimization (SEO)

The proposed system will be compare with the NMF algorithm which is the one of the algorithm used in Topic model. Table 3 shows the result of topic name using LDA algorithm (sample: 5-pages).

Table 3: Topic Name using LDA algorithm

No	Original Topic Page	LDA Words	Gemini Topic Label
1	System Design Tutorial	design, system, interview, tutorial, data	System Design
2	Machine Learning Tutorial	machine, learning, tutorial, data, python	Machine Learning
3	Web Design Tutorial | Learn Web Design from Basics	design, web, tutorial, graphic, data	Web Design
4	Software Design Patterns Tutorial	design, pattern, link, tutorial, method	Design Patterns
5	What is SEO Optimization or Search Engine Optimization?	search, engine, google, optimization, tutorial	Search Engine Optimization (SEO)

6- Evaluations and Discussion Section

This section evaluation and discussion the proposed methodology, among different evaluation metrics, the similarity method is the mostly closest metric, and it is obtained by comparing word vectors (so-called “word embeddings”). These vectors are multiple dimensions of meaning representations of a word [26]. Cosine Similarity, This metric measures the cosine of the angle between two vectors (the original title and the generated title). A cosine similarity score of 1 indicates perfect similarity, while a score of 0 means no similarity. Table 4 and Figure 5 shows that the LDA algorithm is a best from NMF algorithm according to the cosine similarity value, the similarity will be compute between the original topic and generated topic.

Table 4: Similarity values between topics

No	Original Title	NMF Title	Similarity Value	LDA Title	Similarity Value
1	System Design Tutorial	Data Interview Design	0.3333	System Design	0.8164
2	Machine Learning Tutorial	Machine Learning	0.8164	Machine Learning	0.8164
3	Web Design Tutorial | Learn Web Design from Basics	Web Design	0.8302	Web Design	0.8302
4	Software Design Patterns Tutorial	Design Patterns	0.7071	Design Patterns	0.7071
5	What is SEO Optimization or Search Engine Optimization?	Search Engine Optimization (SEO)	0.6666	Search Engine Optimization (SEO)	0.6666

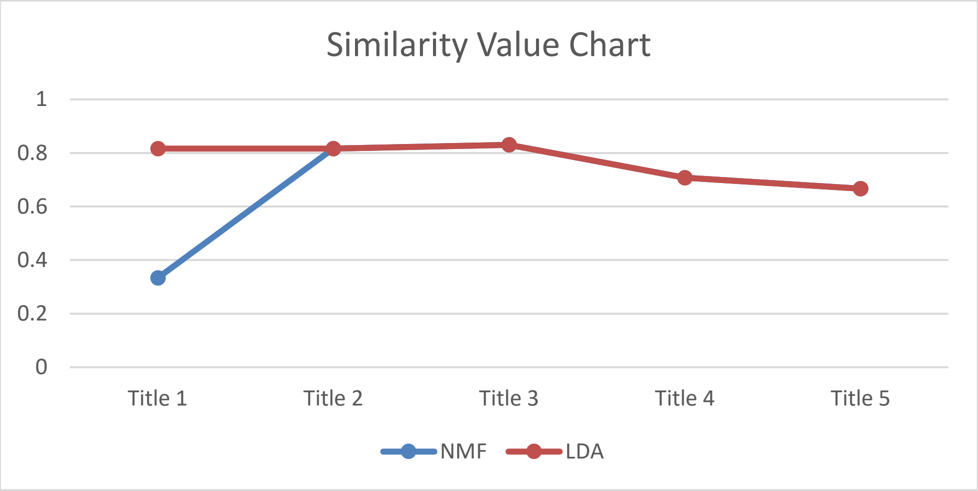

Figure 5: Similarity values between topics

Discussion and Analysis for Title:

1. **Cosine Similarty:**

- **Machine Learning Tutorial (0.8164 for both NMF and LDA)**: Both models show a high cosine similarity for this pair. This makes sense since both titles are focused on "Machine Learning," and the small variations (like "Tutorial" vs. "Machine Learning") are still closely related.
- **System Design Tutorial (0.3333 with NMF, 0.8164 with LDA)**: The NMF model's low cosine similarity to "Data Interview Design" suggests that the topic may not be fully aligned with the original context (system design tutorial). The LDA model does much better because it retains the key subject ("System Design").
- **Web Design Tutorial (0.8302 for both NMF and LDA)**: Both models return a high cosine similarity, which indicates that the terms "Web Design" are dominant in both the original title and the simplified model output.
- **Software Design Patterns Tutorial (0.7071 for both NMF and LDA)**: The cosine similarity of 0.7071 indicates that both models recognize the term "Design Patterns," although they discard the "Software" part. The score is moderate but still reflects the connection between design patterns in both titles.
- **SEO Optimization (0.6666 for both NMF and LDA)**: While the term "SEO" is central to both titles, the phrase "What is SEO Optimization or Search Engine Optimization?" is more verbose. Both NMF and LDA produce a shorter title, which results in a moderate similarity score.

2. **Topic Coherence:** Topic coherence is a metric used to measure how meaningful the generated topics are. It checks whether words within the topics co-occur frequently in documents and can give insight into how well the model understands the topic. Analysis for Title: System Design Tutorial: The discrepancy in similarity scores (NMF: 0.3333, LDA: 0.8164) might indicate that LDA is more effective at identifying coherent topics related to "System Design" due to its probabilistic approach. The NMF model might not fully capture the broader context of the tutorial format, hence a lower coherence. Both NMF and LDA are likely to show high topic coherence for others topic.

3. **Title Length:** A longer original title could lead to lower similarity scores; as more complex phrasing may reduce the focus on the central topic. For instance, "What is SEO Optimization or Search Engine Optimization?" is a more complex title compared to "Machine Learning.".

4. **Topic Saturation:** Titles like "Machine Learning Tutorial" and "Web Design Tutorial" are more saturated with common phrases related to their topics, which makes them easier for both NMF and LDA to match.

7- Conclusion Section

NMF is generally more sensitive to specific word choices and might show lower similarity when titles are phrased differently (e.g., "System Design Tutorial" to "Data Interview Design"). LDA, being more flexible in its probabilistic approach, often performs better when titles vary in structure but still share a common underlying topic. In terms of cosine similarity, both models perform well for clear topics like "Machine Learning" or "Web Design," but the topic coherence and preprocessing steps show how each model responds to complexity and word variation.

Similarity Score: Generally, the models show reasonable agreement between the original titles and the generated ones, with LDA slightly outperforming NMF in terms of matching the original intent and content. Relevance to Core Topic: Both models tend to reduce the title to its core concept, for instance, "Machine Learning" and "Web Design." However, the differences in similarity scores (e.g., 0.3333 for "System Design Tutorial" with NMF) reflect how much the models retain or lose context during topic modeling

References

[1] P. Sukumar, L. Robert, and S. Yuvaraj, "Review on modern Data Preprocessing techniques in Web usage mining (WUM)," in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2016, pp. 64-69.

[2] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey," *J. Web Semant.*, vol. 36, pp. 1–22, 2016.

[3] K. Sharma, G. Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," in *2011 3rd International Conference on Electronics Computer Technology*, 2011, vol. 1, pp. 399–403.

[4] P. S. Sharma, D. Yadav, and R. N. Thakur, "Web page ranking using web mining techniques: a comprehensive survey," *Mob. Inf. Syst.*, vol. 2022, pp. 1–19, 2022.

[5] S. Yadao and A. Vinaya Babu, "Usage of Web Mining for Sales and Corporate Marketing," in *Communication Software and Networks: Proceedings of INDIA 2019*, 2021, pp. 55–60.

[6] M. A. Mohammed, R. A. Hamid, and R. R. AbdulHussein, "Data Collection and Preprocessing in Web Usage Mining: Implementation and Analysis," *Iraqi Journal for Computers and Informatics*, vol. 50, no. 2, pp. 54–74, 2024.

[7] D. Navadiya and R. Patel, "Web Content Mining Techniques – A Comprehensive Survey," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 10, pp. 1–6, 2012.

[8] S. A. Inamdar and G. N. Shinde, "An agent based intelligent search engine system for web mining," in *Research, Reflections and Innovations in Integrating ICT in education*, 2000.

[9] K. R. Srinath, "An Overview of Web Content Mining Techniques," 2017.

[10] R. H. Salman, M. Zaki, and N. A. Shiltag, "A Study of Web Content Mining Tools," *Al-Qadisiyah Journal of Pure Science*, vol. 25, no. 2, pp. 1-10, 2020.

[11] K. Chung, H. Yoo, D. Choe, and H. Jung, "Blockchain network based topic mining process for cognitive manufacturing," *Wireless Personal Communications*, vol. 105, no. 2, pp. 583–597, 2019, doi: 10.1007/s11277-018-5979-8.

[12] F. Guo, A. Metallinou, C. Khatri, A. Raju, A. Venkatesh, and A. Ram, "Topic-based evaluation for conversational bots," *ArXiv preprint*, arXiv:1801.03622, 2018.

[13] J. L. Boyd-Graber, Y. Hu, and D. Mimno, *Applications of Topic Models*, vol. 11, Norwell, MA: Now Publishers Incorporated, 2017.

[14] Y. Yang, Y. Liu, X. Lu, J. Xu, and F. Wang, "A named entity topic model for news popularity prediction," *Knowledge-Based Systems*, vol. 208, art. no. 106430, 2020.

[15] M. R. Bhat, M. A. Kundroo, T. A. Tarray, and B. Agarwal, "Deep LDA: A new way to topic model," *Journal of Information and Optimization Sciences*, vol. 41, no. 4, pp. 1–12, 2019.

[16] Y. Lee and J. Cho, "Web document classification using topic modeling based document ranking," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 2386–2392, 2021.

[17] H. H. Altarturi, M. Saadoon, and N. B. Anuar, "Web content topic modeling using LDA and HTML tags," *PeerJ Computer Science*, vol. 9, art. no. e1459, 2023.

[18] S. Shahid, T. Anand, N. Srikanth, S. Bhatia, B. Krishnamurthy, and N. Puri, "HyHTM: Hyperbolic Geometry based Hierarchical Topic Models," *ArXiv preprint*, arXiv:2305.09258, 2023.

[19] R. R. Tated and M. M. Ghonge, "A Survey on Text Mining- techniques and application," *vol. 3, no. 1*, pp. 380–385, 2015.

[20] B. Pahwa, S. Taruna, and N. Kasliwal, "Sentiment Analysis- Strategy for Text Pre-Processing," *Int. J. Comput. Appl.*, vol. 180, no. 34, pp. 15–18, 2018, doi: 10.5120/ijca2018916865.

[21] H. Jain, A. Jain, C. Maji, A. Tiwari, and N. Jain, "TiZen: Neural Title Generation for Scientific Papers," 2024.

[22] H. Saif, M. Fernandez, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," in *Proc. 9th Int. Conf. Lang. Resour. Eval. Lr.*, 2014, pp. 810–817.

[23] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[24] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, pp. 15169–15211, 2019.

[25] G. Team, R. Anil, S. Borgeaud, J. B. Alayrac, J. Yu, R. Soricut, et al., "Gemini: a family of highly capable multimodal models," *ArXiv preprint*, arXiv:2312.11805, 2023.

[26] [pypi], "https://pypi.org/project/keytotext/," Accessed on Jan. 17, 2025.