News Coverage of Libraries in India: An NLP-based Analysis of News indexed in Google

Sourav Mazumder **Tapan Barui**

Abstract

The purpose of the study is to analyse the news coverage of libraries in India from April 2023 to March 2024. A total of 166 news titles and metadata were collected manually using a structured query using Google. Natural Language Processing (NLP) techniques including topic modelling, Named Entity Recognition (NER), and Sentiment Analysis were used to explore trends in news coverage, news topics, entities, and sentiments from the news titles. Major Findings of the study revealed that August 2023 is the most common month in terms of news coverage; the Times of India is the significant news source; "Academic Library" related topics dominated the news discourse; Organisations and location-based entities as the top entities found in the news titles, and neutral and positive sentiments expressed mostly in the news titles. The findings hold implications for library stakeholders and media professionals and provide valuable information regarding the portrayal of libraries in the news media landscape. This study uses a unique approach using mixed NLP techniques to get insights into library news topics, named entities, and sentiments and shows how libraries in India receive media attention.

Keywords: BERTopic modelling; Entity Extraction; Sentiment Polarity; Library News; NLP Techniques

1. Introduction

Libraries play an important role in society for creating and disseminating knowledge. They provide access to a wide range of resources such as books, digital books, and educational programs. In addition, libraries offer lifelong learning, literacy, research, preserve cultural heritage, and make communication between individuals and the global knowledge community (Krolak, 2006). Understanding libraries' significance can be gleaned from scientific outputs made by researchers working in the area (Castelli, 2006; Krolak, 2006; Lee, 2005). Scientific outputs showcase libraries' activities and innovation towards contribution of the institute. On the other hand, news media (print and online) is another source of information for library news, events, and initiatives for all, particularly the general public.

Corresponding Author
Sourav Mazumder
Email: smazumderlis91@gmail.com

In the era of Information Communication Technology (ICT), online news can be accessed through smartphones, tablets, and computers via the internet. Cui and Liu (2017) mentioned three types of online news media: legacy, explanatory, and citizen news. These media report news regarding fact-based aspects, in-depth analysis of news stories, blogging, social media, and individual websites. Information is available in various multimedia source types such as web, images, maps, videos, books, flights, and finance through Google Search engine. For more than two decades, Google has been the dominated force in the market. According to Statista data, monthly visits to google.com are 84.20 billion along with 91.74% of worldwide market share across all devices which reflects the popularity of Google. (Online Search Market Worldwide - Statistics & Facts, 2024) Additionally, India stands at top position in terms of share of search traffic with 92.9%. Transitioning to online news aggregation, Google News is one of the popular news platforms which curate news stories from various news sources such as The Times of India, India Today, The Hindu, etc. (Similarweb, 2024) As an example, major statistics for the month of March, 2024 reveals the total visits to Google News are 375.60 million by majority of visitors belongs to 35-44 age groups and received the most traffic to news.google.com from USA. These statistics illuminate the significance of Google News and how it has been utilised for staying up to date on the latest developments in the world, science, technology, business, sports and many more sectors.

Library news indexing with Google can reach a large group of readers. News related to library services, programs, and events can make more impact on the usage and availing library services. This study focused on the news coverage about libraries in India published in various news media during April 2023 to March 2024. Employing Natural Language Processing (NLP) techniques, this study analyses news trends, topics, named entities, and sentiments from news titles.

1.1 Research Objectives

This study states mainly four objectives, which are as follows:

- ❖ To analyse the trend of news coverage about libraries in India from April 2023 to March 2024
- ❖ To identify topics discovered through news titles
- ❖ To explore the entities that exist in the news titles
- ❖ To measure the sentiment expressed in the news titles

2. Related Work

This section reviews related studies and works that explores libraries on the news, as well as NLP techniques utilised for analysing news headlines and articles. Cho (2018) analysed Korean news coverage of libraries using semantic network analysis. The study identified the emphasis on public libraries over school and university; and lifelong learning. Publicity about businesses and events was more immersed than library policy issues. Gilbert and Kelley (2024) found that most news analysis studies utilised text articles and the

use of subscription-based databases. There was a significant rise in news websites and the use of news content for research. Sherman and Oakley (2024) conducted a systematic discourse analysis of media coverage regarding the community engagement of small and rural libraries. The study discovered the evolving role of libraries and depicted libraries as "safe spaces." Zhang and Wei (2021) examined news articles from university presses and news sites of libraries based on collaboration in scholarly communication using content analysis. The results showed news articles aimed to cover collaboration background, advantages, and operations. Regardless, most articles focused on university presses.

NLP techniques have been implemented in many cases such as text mining, sentiment analysis, named entity recognition, and part-of-speech (POS) tagging (Mazumder & Barui, 2021; Nadeau & Sekine, 2007; Ramesh et al., 2023). For example, NLP is used to identify fake news and classify real and fake news on social media (Ramesh et al., 2023). Varol et al. (2022) analysed over 36,000 articles from CNN and The Guardian with the help of clinical and biomedical NLP models to examine medical concepts, key entities, biases, and changes over time in news coverage. Park et al. (2023) investigated news coverage of the Russian-Ukraine war from multiple countries utilising LDA and semantic network analysis. Singh and Singh (2021) explored top news items and measured the similarity between English and Hindi news articles using by vector space model using NLP. The study revealed efficient identification and comparison of top news articles and patterns in news reporting. Chen et al. (2023) compared LDA, Top2Vec, and BERTopic topic models for news impact analysis on financial markets. The results showed that BERTopic was the most effective technique for that kind of analysis. Singh and Jain (2021) utilised transformer-based sentiment analysis to evaluate sentiments regarding news headlines. The bert-base-cased model was identified as the best model in terms of performance. Bade Shrestha and Bal (2020) presented the popularity of politicians through time series graphs of positive and negative sentiments from Nepali news text data using NER-based sentiment analysis.

Although these studies have been employed NLP to examine various aspects of news. There is a research gap in analysing news indexed to Google in the context of libraries. This study aims to fill this gap by analysing news articles' titles of libraries indexed in Google News during April 2023 to March 2024 by using NLP techniques.

3. Methodology

This study comprises of exploratory and descriptive research. As depicted in Figure-1, the methodology used in the study are discussed in the following subsections.

3.1 Data Collection and Dataset Preparation

English news related to libraries in India were collected manually through Google by selecting the news tab on 01/04/2024. Employing a structured Google search query "allintitle:(library OR libraries) India" and filtering parameters, such as time for span from April 01, 2023 to March 31, 2024, the retrieved search results

were refined to hide duplicates and sorted by date. Initially, the search results provided a total of 138 news articles. After screening for relevancy, 120 article headings were included and the other 18 titles were excluded. Moreover, cross-checking was conducted within the Google News App using the same search query and found 46 additional unique article headings. All collected metadata was recorded into Google Sheets which includes news title, source, date, and URL, etc.

In the phase of preparing and cleansing of dataset, it has been pre-processed for removing the stopwords, some non-stopwords (e.g., library, libraries, and India), punctuations, special characters, and URLs from titles using NLTK (Bird et al., 2009) based on stopwords list and Regular Expression (RegEx). This approach enriches the quality of the dataset and enhances the data analysis.

3.2 Data Analysis and Visualisation

Data Analysis of the study is based on three NLP techniques: Topic Modelling, NER, and Sentiment Analysis. However, basic characteristics of the news headlines, such as the trend of monthly article publication and sources of the articles were also presented. In the first phase, BERTopic modelling (v.0.16.1) (Grootendorst, 2022) was used to extract latent topics from the news titles for topic modelling. NER was performed using spaCy (v.3.7.4) (Honnibal et al., 2020) for identifying and categorising named entities within the news titles in the next phase. There are 18 entity types in the spaCy model, such as date, persons, organisations, nationalities, and locations. Sentiment analysis was conducted using NLTK's (v.3.8.1) SentimentIntensityAnalyzer with VADER lexicon for sentiment classification (Hutto & Gilbert, 2014) in the third phase. Furthermore, compound score (CS) was measured to determine positive ($CS > 0$), neutral ($CS \approx 0$) and negative ($CS < 0$) sentiments expressed in the news headlines. This helped to understand the emotion of the news headings. Other data analyses and visualisations depicted in the results were created using various python libraries including Pandas (v.2.0.3) (McKinney, 2010) and Matplotlib (v.3.7.1) (Hunter, 2007).

This study only focuses on news headlines of content indexed in one year rather than the full text content. Since, the news indexing is ongoing procedure, other researchers may find different results using the search method of this study. Additionally, very few search keywords were used to retrieve data. It was found that some articles were based on international context. These article titles were included for analysis. Apart from this, the study aimed to present unique outcomes about trends, topics, key entities, and sentiments expressed in the news titles.

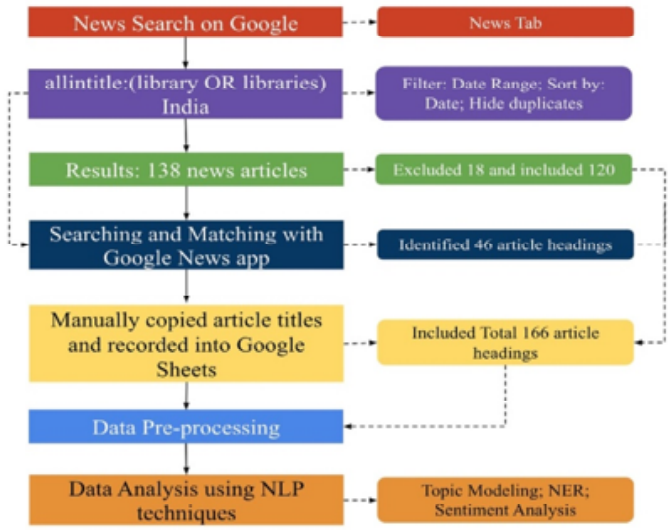

Figure 1: Schematic Representation of the Implemented Methods for Analysis

4. Results and Discussion

4.1 Trend of News Coverage on Libraries in India

The trend of news coverage on libraries in India over the months is illustrated in Figure-2(A). Generally, an upward trend was observed from April 2023 to August 2023. The peak was the month of August with 31 news coverage. The average coverage was almost 7 articles per week. A moderate coverage was found from September (n=13) to October (n=12). The lowest points were observed from November 2023 to January 2024, with a total of 23 articles. Subsequent months (i.e. February-March 2024) presented increasing with a total of 29 news articles. The trends might be influenced by potential factors such as media focus, events, and public perception. From a statistical perspective, the average news coverage was almost 14 articles per month, and the standard deviation was close to 8.

The analysis of the news source also provides valuable information about the importance of media focus. Figure-2(B) shows the top-10 media sources that focused coverage on libraries. The Times of India emerged as the most significant source of news (n=33) on libraries, featuring various types of news such as public libraries, opening libraries, and library issues. It was followed by India Education Diary and India Today, which covered 12 and 10 library news respectively. On the other hand Indian Express, The Hindu, The Telegraph, and Tribune India covered 7 articles each. Business Standard (n=6), Hindustan Times (n=5), and The New Indian Express (n=3) had fewer coverage during the period. It can be said that libraries are portrayed by top-tier news media in India.

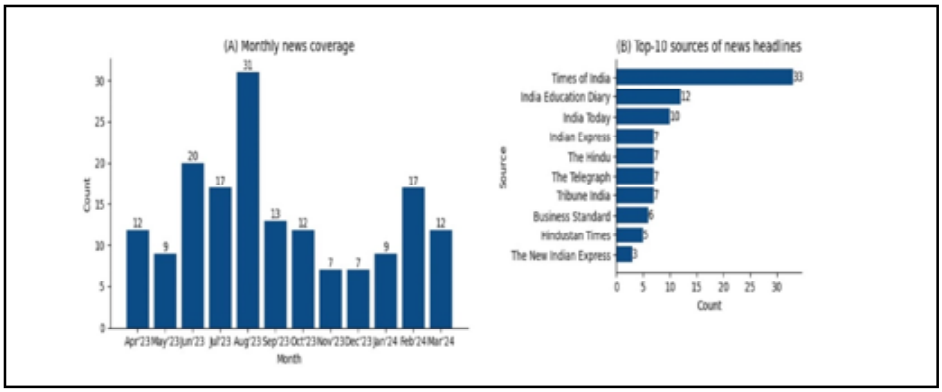

Figure 2(A): Monthly news coverage on libraries and (B) Top-10 sources of news

4.1 Topic Modelling

This section presents the results of topic modelling using BERTopic and shows topics derived from news titles as depicted in Figure-3. Five distinct topics were identified and characterised by ten representative keywords. The assignment of topic names is based on human judgement and there can be multiple topics represented by topics' keywords. The first topic labelled is 'Academic Library', with keywords such as 'university', 'public', 'director', 'college', 'ranchi', 'students', 'first', 'ever', 'collapses', and 'classical'. It gives an overview of activities and issues related to human resources, collection management, and administrative aspects in academic libraries. It encompasses 64 news articles. The second topic deals with 'Events and Public Library', representing keywords – 'inaugurates', 'hyderabad', 'delhi', 'list', 'biggest', 'memory', 'chennai', 'mansa', 'national', and 'public'. It can be ascertained that emphasis on geographical locations and the significance of public libraries has been depicted through the news media in 44 news articles. This particular result is also matched with the study of Cho (2018) who found the significance of public libraries. The third topic is 'Reading Habit' and library-related aspects. This was judged based on the keywords – 'book', 'beautiful', 'reading', 'world', 'public', 'biggest', 'digital', 'people', 'visit', and 'old'. Keywords of this topic indicate contexts of library use by citizens (children, adults, and old aged), reading books, awareness about libraries' value, and access to digital resources, such as e-books. A total of 33 news titles aligned to the topic 3. The fourth topic is characterised by 'Digital Library and Access' and rendered in 13 news by describing keywords- 'digital', 'awareness', 'indian', 'national', 'students', 'first', 'treasure', 'subjects', 'nlsiu', and 'underserved'. This suggests the growing interests on access to digital libraries and digitisation. In the age of ICT, digital libraries serve as a phenomenal tool for accessing digital resources. Lastly, topic five focuses on 'Renaming Libraries', encompassing keywords like 'museum', 'memorial', 'prime', 'renamed', 'ministers', 'society', 'name', 'dropped', 'government', and 'renames'. In some news (n=12), it was found that libraries' names were renamed. These topics provide information about existing contexts within the corpus of news titles. It also can stress the media coverage agendas.

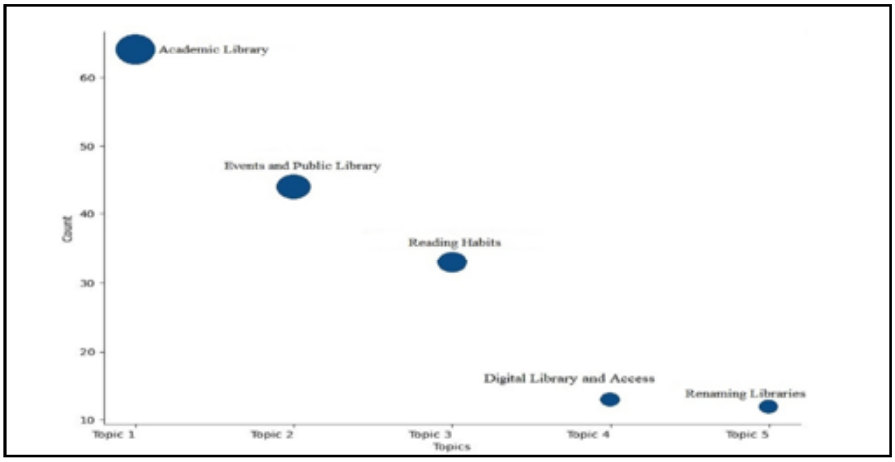

Figure 3: Five Topics assigned to Topic Numbers

4.2 NER analysis

The results of NER analysis for extracting and categorising named entities in the news titles are presented in this section. As mentioned in Section 3.2, NER helps identify key entities such as people, organisations, dates, and locations. Figure-4 illustrates examples of identified named entities in 5 news titles. The text highlighted with cyan colour indicates 'ORG' (organisations, companies, agencies, institutes, etc.), such as IIT Kharagpur, KK Handiqui Library, and Gauhati University. The light, yellow-coloured text specifies 'GPE' (i.e. Geo-Political Entities: countries, cities, states). The word 'Bengali' demonstrates the entity related to 'NORP' (i.e. Nationalities or Religious or Political Groups), highlighted with purple. The entity "PERSON" has been illuminated by a light purple colour. The light, grey-coloured text depicts 'CARDINAL' (i.e. numerals that do not fall under another type).

Figure 4: Examples of Entities Identified in News Headings

This study found 13 entities out of 18 entities (Figure-5) in 153 news titles. A total of 13 news titles were labelled as 'N/N' (No Entities), as they do not contain any entities. According to Figure-5, the most frequent

entity type is 'ORG (Organisation)' (n=125). This implies that 'ORG', such as libraries, colleges, and universities are predominantly mentioned across news titles. The second most frequent entity type is 'GPE' (n=105), which indicates the strong presence of location-based entities in the news titles. One potential factor could be that this study's search string involved the geographic name 'India'. As a result, the occurrence of 'GPE' is high since most of the titles contain the term 'India'. An in-depth analysis revealed 17 news titles having locations without the term 'India'. In addition, 'PERSON' (e.g., people, including fictional) (n=34) and 'CARDINAL' are also identified as frequent entities. This suggests a focus on individuals and numerical data within the text (see Figure-4). Entities like 'NORP' and 'FAC' (i.e. facility, e.g., building, airport) were also present in the news titles, with 10 each. Furthermore, there are some other types of entities, such as 'DATE' (i.e. absolute or relative dates or periods; n=9), 'ORDINAL' (e.g., first, second, etc.; n=5), 'MONEY' (monetary values, including unit; n=4), 'TIME' (i.e. times smaller than a day; n=3), and 'PERCENT' (n=1). These results reveal the presence of entities in the corpus of the news titles and highlight the prominence of organisations, geographical locations, numerical data, persons, nationalities, monetary, etc.

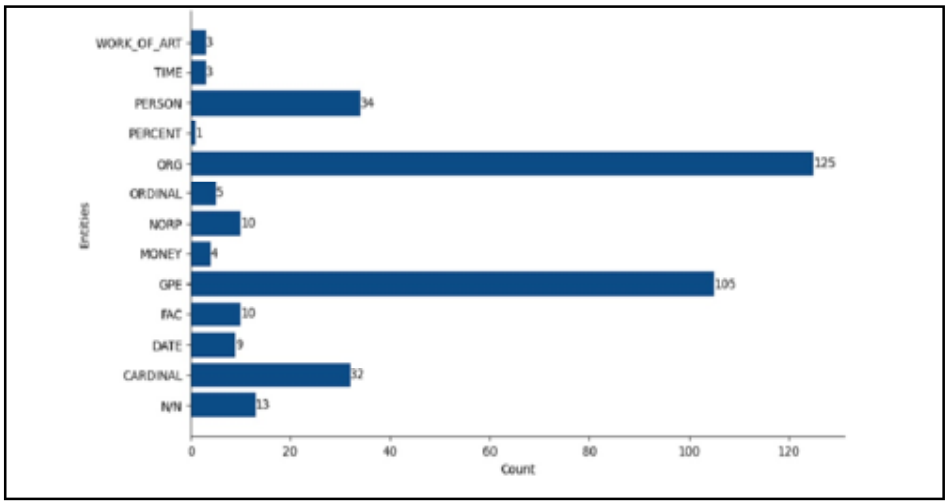

Figure 5: Counts of Entities Identified in News Titles

4.3 Sentiment Analysis

This section aims to draw a sentiment analysis of news titles with overall sentiment polarity (positive, neutral and negative). Table-1 provides a few titles along with their corresponding sentiment scores. It offers a clear understanding of the emotion of a text. Based on the compound score, the levels of positive, neutral, and negative can be aggregated for an informative outcome.

Table 1: Examples of Text along with Sentiment Score

Title	Scores				
	Sentiment	Compound	Positive	Negative	Neutral
Visits to public libraries increase	Positive	0.3182	0.535	0	0.465
Dispute rocks Murty Classical Library of India; editorial board dismissed without explanation	Negative	-0.4019	0	0.278	0.722
Library to accept old book donation: The library at Swapnabhor, the senior citizens' club, is building on its stock with donated books	Positive	0.3818	0.191	0	0.809
Connecting People With Books: The Rise of Free Libraries In India	Positive	0.5106	0.452	0	0.548
Talk at the Library: C. G. Jung and India: A Conversation with Cultural Historian Sulagna Sengupta	Neutral	0	0	0	1
Bangalore man arrested for stealing rare, century-old book from Nilgiri library	Negative	-0.7783	0	0.493	0.507
Why libraries are liberating spaces	Neutral	0	0	0	1

Further, the analysis reveals three types of sentiments expressed within corpus (Figure-6). The majority of the sentiments of the news titles are neutral (n=84, 50.60%). Positive sentiments were expressed in 61 (36.75%) titles. For instance, news on library visits, new establishments, access to physical and digital libraries, individuals' success stories, and community engagement have been covered for a balanced aspect of sentiments. A significant portion of the news' emotional tone is either neutral or pointing a positive stance on libraries. On the other hand, a total of negative sentiments were discovered in 21 (12.65%) news titles. Negative sentiments can help in identifying the reasons and areas of concern. For example, issues and challenges related to libraries are important to know about for sophisticated library services. Some news articles do not reflect any library-related issues (e.g., India Today, 2023) which indicate mixed sentiments expressed through the news titles.

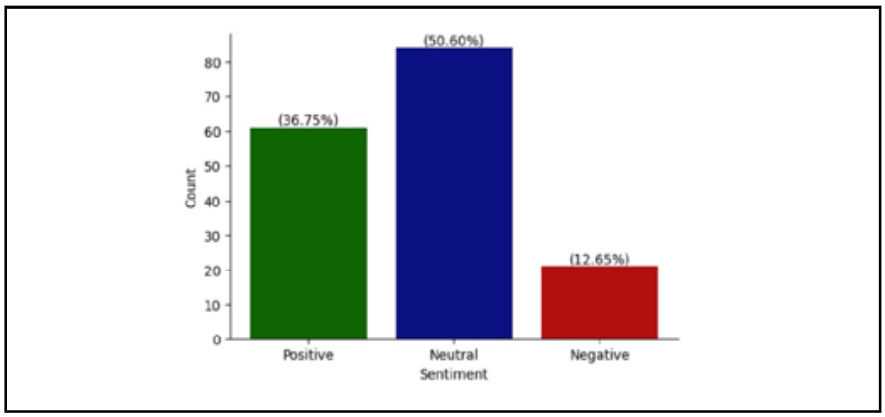

Figure 6: Distribution of sentiments expressed in the news titles

5. Conclusion

This study utilised a hybrid approach using NLP to analyse news articles' titles that covered various perspectives of libraries in India. The findings of the trends analysis revealed that August 2023 received the highest media attention and the average frequency of news coverage was 14 per month. The Times of India was the top source for library-related news. Using BERTopic, news topics were discovered, and the most prominent topic is 'Academic Library', closely followed by 'Events and Public Library', and 'Reading Habit' which can help in identifying areas of media focus. A total of 13 named entities were recognised out of which topmost were 'ORG' and 'GPE' which highlight the geographical and institutional dimensions of the media discourse. The dominant sentiment across all news titles was neutral but there was a significant presence of positive sentiment and a lesser portion of negative sentiments.

The results hold several implications for library stakeholders as well as media professionals. First, library stakeholders like library professionals, researchers, and policymakers can comprehend the findings and concentrate on community engagement. Second, the media professionals can show the interests and issues regarding libraries. Third, from the methodological perspective, researchers can conduct a more in-depth analysis of various contexts of libraries with more years. This can project the evolution of media coverage on libraries and how the patterns changed over time.

There are some limitations in this study. Only one year was taken as a time frame using a couple of search keywords, which might restrict the comprehensiveness of the search results related to library news in India. Search results might be changed based on ranking over time. Some entities were not recognised correctly. For the quality of analysis, manual screening was undertaken for each title to keep consistency. A more compact and fine-tuned NER approach could tackle this situation and reduce the bias over entity extraction. Nevertheless, the results provide significant outcomes regarding media coverage of libraries in India; and can guide future research prospects for advancing the field of study using more titles, including full-text articles.

References

1. Bade Shrestha, B., & Bal, B. K. (2020). Named-Entity Based Sentiment Analysis of Nepali News Media Texts. In E. YANG, E. XUN, B. ZHANG, & G. RAO (Eds.), Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications (pp. 114–120). Association for Computational Linguistics. https://aclanthology.org/2020.nlptea-1.16

2. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. https://github.com/nltk/nltk/blob/develop/CITATION.cff

3. Castelli, D. (2006). Digital libraries of the future – and the role of libraries. Library Hi Tech, 24(4), 496–503. https://doi.org/10.1108/07378830610715365

4. Chen, W., Rabhi, F., Liao, W., & Al-Qudah, I. (2023). Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. Electronics, 12(12), Article 12. https://doi.org/10.3390/electronics12122605

5. Cho, J. (2018). The Trends of Media Coverage about Libraries in Korea: Using Semantic Network Analysis of Portal News. Libri, 68(4), 291–300. https://doi.org/10.1515/libri-2018-0068

6. Cui, X., & Liu, Y. (2017). How does online news curate linked sources? A content analysis of three online news media. Journalism, 18(7), 852–870. https://doi.org/10.1177/1464884916663621

7. Gilbert, S., & Kelley, R. (2024). A Content Analysis of News Analyses: Examining Trends in News Content and Resources. Journal of New Librarianship, 9(1), Article 1. https://doi.org/10.33011/newlibs/15/1

8. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv. https://maartengr.github.io/BERTopic/index.html#visualizations

9. Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. 10.5281/zenodo.1212303

10. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

11. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14550

12. India Today. (2023). Student dies after roof of library collapses in Jharkhand's Ranchi. https://www.indiatoday.in/india/story/student-dies-after-central-librarys-roof-collapses-in-jharkhands-ranchi-2392940-2023-06-14

13. Krolak, L. (2006). The role of libraries in the creation of literate environments. International Journal of Adult and Lifelong Education, 4(1–4), 5–28. https://unesdoc.unesco.org/ark:/48223/pf0000210034.locale=en

14. Lee, H.-W. (2005). Knowledge Management and the Role of Libraries. The 3rd China-US Library Conference. https://www.white-clouds.com/iclc/cliej/cl19lee.htm

15. Mazumder, S., & Barui, T. (2021). Discovering Topics from the Titles of the Indian LIS Theses. Library Philosophy and Practice (e-Journal). https://digitalcommons.unl.edu/libphilprac/5924

16. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

17. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3–26. https://doi.org/10.1075/li.30.1.03nad

18. Online search market worldwide - Statistics & Facts. (2024, March 8). https://www.statista.com/topics/1710/search-engine-usage/#topicOverview

19. Park, H., Pak, J., & Kim, Y. (2023). Analysis of News Article Various Countries on a Specific Event Using Semantic Network Analysis. In S. Latifi (Ed.), ITNG 2023 20th International Conference on Information Technology-New Generations (pp. 229–235). Springer International Publishing. https://doi.org/10.1007/978-3-031-28332-1_26

20. Ramesh, A., Thube, G., & Jadhav, S. (2023). Realtime News Analysis using Natural Language Processing. 2023 4th International Conference for Emerging Technology (INCET), 1–6. https://doi.org/10.1109/INCET57972.2023.10170350

21. Sherman, M., & Oakley, S. (2024). Small & Rural Libraries Transforming Communities: A Discourse Analysis of Media Coverage. The Library Quarterly. https://doi.org/10.1086/730467

22. Similarweb. (2024). News.google.com Traffic Analytics, Ranking & Audience [July 2024]. Similarweb. https://www.similarweb.com/website/news.google.com/

23. Singh, A., & Jain, G. (2021). Sentiment Analysis of News Headlines Using Simple Transformers. 2021 Asian Conference on Innovation in Technology (ASIANCON), 1–6. https://doi.org/10.1109/ASIANCON51346.2021.9544806

24. Singh, R., & Singh, S. (2021). Text Similarity Measures in News Articles by Vector Space Model Using NLP. Journal of The Institution of Engineers (India): Series B, 102(2), 329–338. https://doi.org/10.1007/s40031-020-00501-5

25. Varol, A. E., Kocaman, V., Haq, H. U., & Talby, D. (2022). Understanding COVID-19 News Coverage using Medical NLP (arXiv:2203.10338). arXiv. https://doi.org/10.48550/arXiv.2203.10338

26. Zhang, M., & Wei, X. (2021). What Can "Marriage Announcements" Tell Us? A Content Analysis of News Articles on Library-Press Collaboration | Zhang | College & Research Libraries. https://doi.org/10.5860/crl.82.7.959

About Authors

Sourav Mazumder, Research Scholar, Dept of Lib. & Info. Sci., University of North Bengal, India
Email: smazumderlis91@gmail.com
ORCID: https://orcid.org/0000-0003-0956-661X

Tapan Barui, Assistant Professor, Dept of Lib. & Info. Sci., University of North Bengal, India
Email: tapanbarui@nbu.ac.in
ORCID: https://orcid.org/0000-0002-8023-4987