Decoding a Decade:
A Temporal Analysis of India's Most Prominent News Topics

Introduction

Media serves as a powerful mirror of society, reflecting its aspirations, challenges, and priorities. News headlines, in particular, encapsulate the essence of public discourse, shaping and being shaped by the events they report. In India, a country characterized by its diversity and dynamic socio-political landscape, analyzing news content offers valuable insights into the evolving narrative of a nation navigating rapid change. From governance and public health to crime and cultural milestones, headlines capture a condensed yet impactful record of societal priorities. Leveraging Natural Language Processing (NLP), this study systematically analyzes a decade's worth of news headlines to uncover recurring themes and temporal trends that have defined India's discourse.

The motivation behind this research lies in the need to understand how media narratives influence public opinion and policymaking[1]. With millions of headlines spanning multiple years, manually identifying patterns and shifts in focus becomes impractical. NLP provides the tools to automate this analysis, enabling researchers to uncover latent themes, track their prominence over time, and interpret broader implications. In an era where media plays a critical role in shaping perceptions, this research contributes to a deeper understanding of societal issues that dominate public attention and the patterns that emerge in response to crises, elections, and transformative events[2].

Research Questions

This research addresses the following key question: *What dominant themes and trends emerge from Indian news headlines over the past decade, and how have these evolved temporally in response to societal, political, and economic changes?*

The objectives of this analysis are threefold:

1. To identify the dominant topics present in Indian news headlines over a decade.
2. To analyze temporal shifts in these topics, exploring how societal and media priorities have evolved over time.
3. To interpret the findings in the context of India's socio-political landscape, providing insights into the interplay between media narratives and societal changes.

This study is particularly relevant for policymakers, journalists, and scholars in media studies, public policy, and sociology. For policymakers, it highlights issues dominating public attention, enabling more responsive governance. For journalists, it underscores the responsibility of shaping societal priorities. For scholars, it demonstrates the utility of computational methods in analyzing unstructured data. By focusing on the past decade, this analysis bridges the gap between historical trends and contemporary challenges, shedding light on how media responds to pivotal events such as elections, pandemics, and economic reforms.

Structure of the Note

The structure of this research note is as follows: First, the dataset and preprocessing steps are described, detailing how the text was prepared for analysis. This is followed by the methodology, outlining the NLP techniques applied, including topic modeling using Latent Dirichlet Allocation (LDA). The results section presents key findings, offering insights into temporal trends and dominant themes across the dataset. Finally, the discussion addresses the strengths, limitations, and potential areas for future research.

Submitted by: Manjiri Girish Satam (235789)

Data Description & Preprocessing

This research uses a comprehensive dataset of news articles published by the *Times of India*[3], one of the largest selling English-language daily India[4]. *The India News Headlines Dataset*[5] is available on Kaggle and it captures a wide range of events and stories from the Indian subcontinent, spanning over two decades—from January 1, 2001, to June 30, 2023. With nearly 3.8 million articles, it provides a rich historical record of societal trends, local priorities, and major events.

The content focuses primarily on Indian news, including national, city-level, and entertainment coverage. The scale of the dataset is impressive, with an average of 470 articles published daily. This large volume ensures a diverse representation of events, ranging from headline-grabbing stories to everyday developments, offering valuable insights into how the narrative of Indian society has evolved over time.

Here's what the dataset contains:

Time range: Articles span from 2001 to mid-2023.

Number of rows: 3,876,557 entries in CSV format.

Columns:

1. publish_date: The date the article was published, formatted as yyyyMMdd.
2. headline_category: A category label for the headline, using simple dot-separated ASCII values.
3. headline_text: The actual text of the headline, written in English with only ASCII characters.

Due to computational constraints, the scope of this analysis is limited to the data from the last decade, specifically from January 1, 2014, to June 30, 2023. This subset still contains a significant portion of the dataset, with approximately 2 million articles. Focusing on this time period allows for efficient processing while ensuring that the analysis remains relevant to more recent societal and media trends.

This dataset serves as a powerful resource for studying societal changes. The Times of India's widespread reach and consistent reporting mean this archive captures not only major national events but also local stories, reflecting the nuances of everyday life in India.

Overall, the scale and detail of this dataset make it an exceptional tool for analyzing long-term trends and understanding the evolving priorities and issues within Indian society.

Preprocessing Steps

To prepare the dataset for analysis, I focused on cleaning and structuring the text data to make it suitable for topic modeling using LDA (Latent Dirichlet Allocation). The preprocessing steps were essential for ensuring the text was in a uniform format and free from noise.

Steps Taken

1. Text Normalization: All text was converted to lowercase. This ensures consistency, as words like "India" and "india" are treated as the same during analysis.
2. Removing Punctuation and Special Characters: Punctuation and special characters don't add meaning to the text for topic modeling. So these elements to simplify the text further.
3. Stopword Removal: Common words like "the," "is," and "and" (known as stopwords) were removed since they do not contribute to the topics. NLTK's built-in stopword list was used for this step.
4. Tokenization: Using NLTK's word_tokenize() function, the headlines were broken into individual word level and consider that word as token. This step helps in analyzing the data at the word level, which is necessary for topic modeling.

5. Lemmatization: To reduce words to their base form, NLTK's WordNetLemmatizer() was used. For example, "running" becomes "run," and "better" becomes "good." This ensures that variations of a word are treated as the same, which is critical for identifying coherent topics.

Why NLTK?

NLTK (Natural Language Toolkit) was used for the preprocessing steps because it is lightweight and relatively fast for text preprocessing tasks, especially for a dataset of this size. Initially, an attempt was made to use SpaCy for these steps, but the processing time was significantly higher, making it impractical given the size of my dataset. NLTK's tools, like word_tokenize() and WordNetLemmatizer(), provided efficient and reliable results while being easy to implement.

Challenges with Data and Preprocessing

Working with news headlines presented several challenges. Headlines are often short and lack full context, making it difficult to extract meaningful topics. Additionally, the wide range of topics covered in news articles results in a diverse and complex vocabulary, which requires careful cleaning and grouping of words. Punctuation, frequently used in headlines for emphasis—such as quotes or hyphens—added another layer of complexity, as it could interfere with tokenization if not handled properly.

Despite these challenges, the preprocessing methods I employed worked effectively. By converting the text into a clean, tokenized, and lemmatized format, it was ensured that the input for LDA topic modeling was well-structured and ready for analysis. These steps were particularly successful in handling the unique nature of headlines, reducing noise while retaining the core elements of the text. The preprocessing also enhanced the model's ability to identify recurring patterns and uncover underlying themes across millions of entries. Although brief, headlines often carry substantial information, and the chosen methods preserved this richness while simplifying the data for analysis. By addressing both technical and contextual challenges, a strong foundation for identifying valuable topics and trends was laid, offering deeper insights into the themes that have shaped the Indian news narrative over the years.

Methodology & Model Selection

Topic modeling, as a method, is uniquely suited to address the research questions because it excels at uncovering hidden structures and themes in large, unstructured textual datasets. The core strength of topic modeling lies in its ability to identify latent patterns without requiring prior knowledge or labels[6], making it ideal for exploratory tasks like analyzing a decade's worth of news headlines. Unlike other methods such as sentiment analysis, which focuses on emotional tone, or clustering algorithms, which often impose rigid categorizations, topic modeling provides a nuanced understanding of the data by treating each document as a mixture of topics[4]. This approach reflects the complexity of real-world news, where individual headlines often touch upon multiple, overlapping themes. Furthermore, topic modeling enables temporal analysis by capturing how the prominence of these themes shifts over time, aligning perfectly with the objective of tracking societal, political, and economic changes. It also provides insights into the dynamics of public discourse, helping to reveal how major events, crises, or policy shifts influence the media narrative. By organizing and summarizing vast amounts of text data into interpretable themes, topic modeling not only answers the research questions but also provides a robust foundation for drawing connections between media content and societal priorities, making it an indispensable tool for this type of analysis.

For this analysis, Latent Dirichlet Allocation (LDA) was used to perform topic modeling on news headlines. LDA is a popular probabilistic method that identifies latent topics within a collection of text documents. This approach is particularly well-suited for uncovering themes in large textual datasets, as it assumes that each document (in this case, a news headline) is a mixture of topics and each topic is a distribution of words. By leveraging LDA, the aim is to to uncover the dominant topics within the dataset and analyze how

they evolve over time. To facilitate temporal analysis, 10 topics were generated for each year over the last decade (2014–2023), which allowed tracking shifts in societal priorities and concerns.

The number of topics (k) that the model should generate from the corpus is a crucial LDA parameter. The text in the corpus will be divided into k themes when the LDA model is run using k as input. Because several different themes may be combined into one, setting k too small will produce results that are overly generic, missing out on important nuances. On the other hand, if k is set too large, there will be too many topics that are unclear on their own. The LDA model was constructed multiple times with k=5, 10, 15, 20 in order to determine the ideal number of topics k. The quality of the topics that were produced was then compared. Many works that employed LDA as a theme adopted this strategy[7]. Additionally, dividing the dataset into yearly subsets and generating 10 topics for each year allowed to conduct a temporal analysis, helping to identify how societal concerns and trends have shifted over time.

LDA was chosen for its ability to handle large-scale, unstructured text data effectively and its interpretability. Unlike other topic modeling techniques such as Non-Negative Matrix Factorization (NMF), LDA provides a probabilistic framework that assigns probabilities to words belonging to specific topics, making the results more intuitive and easier to analyze. Moreover, LDA does not require pre-defined labels or training data, making it ideal for exploratory tasks like this one, where the goal is to discover latent themes.

In terms of vectorization, I opted for the Bag of Words (BoW) approach instead of TF-IDF (Term Frequency-Inverse Document Frequency). While TF-IDF is effective for giving less weight to common words, it can dilute the importance of frequent terms that might be crucial for headlines. Headlines are typically short and rely on key terms to convey meaning, so BoW is better suited for retaining these terms without penalizing their frequency. To refine the vocabulary further, I used a maximum document frequency threshold of 0.9 (to exclude overly common words) and a minimum document frequency of 5 (to remove infrequent, less meaningful terms).

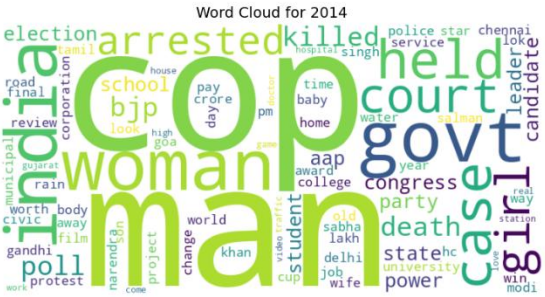
Results and Analysis

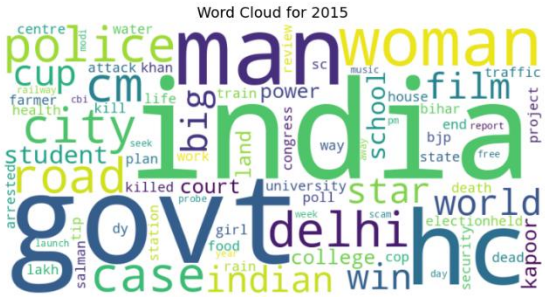
India's political vibrancy is a consistent thread throughout the decade. Major elections, both national and regional, dominate specific years, such as 2014, 2019, and 2023. Keywords like *poll*, *assembly*, *election*, *BJP*, and *Congress* capture the democratic fervor, while references to leaders like *Narendra Modi* highlight the rise of centralized, personality-driven politics. The COVID-19 pandemic (2020–2022) reshaped the news landscape, bringing public health and social welfare to the forefront. The challenges of healthcare infrastructure were laid bare, and the pandemic's impact extended to *education*, with *exam* and *student* becoming prominent keywords. Throughout the decade, infrastructure projects have been a recurring theme. Terms like *road, project, metro*, and *train* frequently appear in headlines, reflecting a governance priority across administrations. In post-pandemic years, infrastructure projects have taken center stage as tools for economic recovery, signaling a push for urbanization and employment generation. The focus on cities like *Mumbai* and *Delhi* emphasizes the centrality of urban development in national priorities. The decade also witnessed significant social and environmental movements. The farmer protests of 2019–2021 stand out, with terms like *protest* and *farmer* highlighting rural discontent with agricultural reforms.

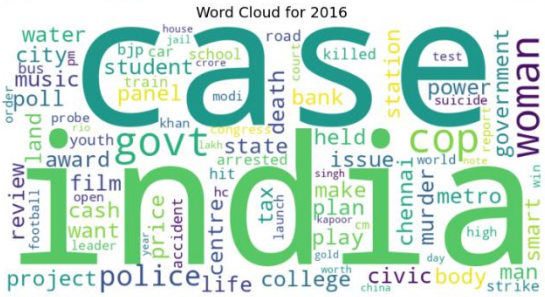
Dominant Themes:

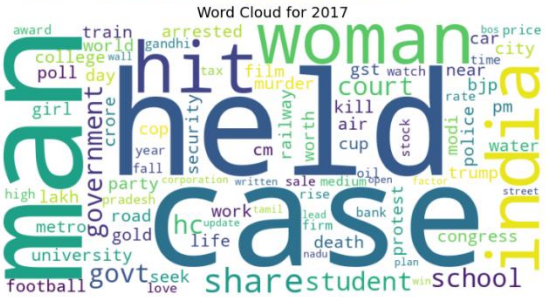
Each year's top topic highlights what dominated public discourse, offering valuable insights into the issues that shaped the nation. This temporal analysis traces the trajectory of India's priorities, unpacking recurring themes, transformative events, and subtle shifts in focus.

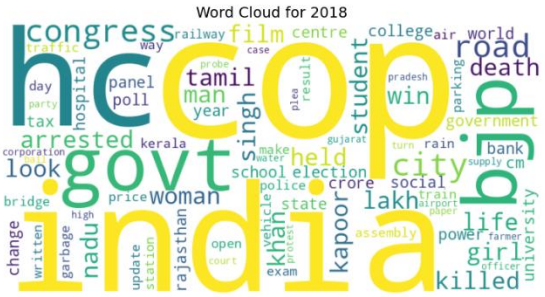
Submitted by: Manjiri Girish Satam (235789)

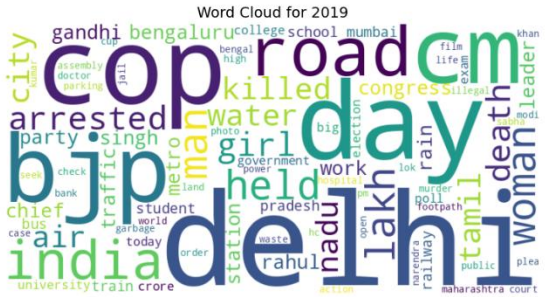


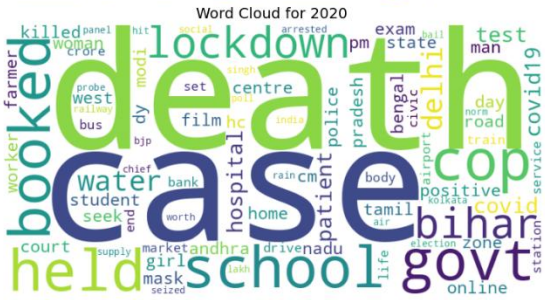


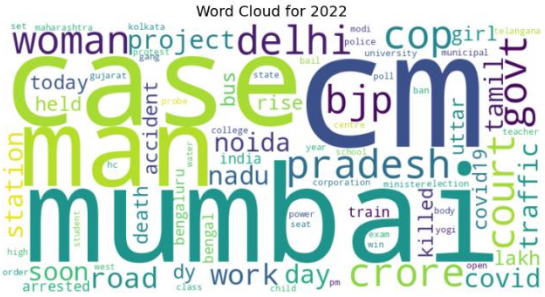


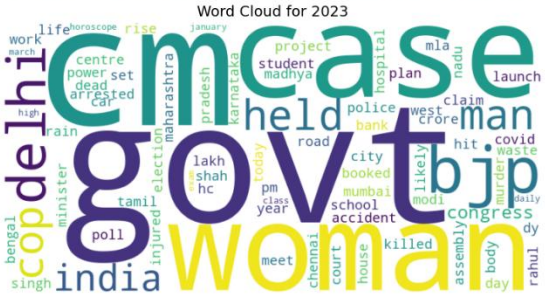












Crime and Justice: A Persistent Concern

In 2014, headlines were dominated by crime and justice, with terms like case, court, cop, girl, woman, arrested, and death featuring prominently. This focus reflects heightened public awareness and outrage over gender-based violence and high-profile criminal cases. The justice system's role in addressing societal concerns remained under scrutiny, setting a tone for years to come.

By 2018, this theme resurfaced with a renewed intensity, as terms like man, woman, cop, arrested, and killed again took center stage. Unlike 2014, the emphasis on public safety had widened, likely driven by publicized systemic failures in ensuring security for vulnerable groups. Similarly, in 2019, crime remained

a key concern, with the addition of Delhi suggesting region-specific events that captured public attention, such as protests or law enforcement controversies.

These recurring topics highlight a deep-rooted issue: despite societal progress in some areas, crime, particularly gender-based violence, and public safety remain unresolved challenges. This persistence may signal either systemic inadequacies in law enforcement and governance or an increasingly vocal public demanding accountability.

Cultural and Entertainment Milestones

In stark contrast to the heavy themes of crime and justice, 2015's dominant topic revolved around India's achievements on the global stage. Terms like India, world, film, win, and cup reflect a celebratory mood, with the Cricket World Cup and cultural accomplishments shaping national discourse. This brief respite from more somber topics suggests that moments of collective pride in sports and entertainment resonate deeply with the public. However, these lighter narratives may also indicate a tendency to momentarily shift focus from deeper societal challenges.

Governance and Education: Shifting Priorities

The mid-decade years of 2016 and 2017 marked a transition toward governance and education as key themes. In 2016, words like state, student, govt, poll, and issue suggest public discourse focused on policies, state elections, and civic issues. This likely reflected an increasing interest in government accountability and public engagement in shaping state-level politics.

By 2017, education became more central, with terms like school, student, university, and case emphasizing judicial and governance challenges within the education sector. This period reflects a growing recognition of education as a cornerstone for social mobility and economic growth. The overlap with judicial terms like court and HC suggests that significant policy or governance failures led to legal interventions, further amplifying the public discourse on education.

This evolution indicates a growing awareness of systemic issues, such as unequal access to education and inefficiencies in governance. Public expectations of reforms in these areas grew, with education emerging as a recurring focus for development.

The COVID-19 Disruption

The years 2020 and 2021 stand out for their unprecedented focus on public health, driven by the COVID-19 pandemic. In 2020, the headlines were dominated by words like COVID19, test, hospital, death, and positive, underscoring the crisis's devastating impact. The focus was on managing the immediate healthcare challenges, exposing gaps in India's medical infrastructure.

Infrastructure and Post-Pandemic Recovery

In the years following the pandemic, the narrative shifted toward rebuilding and development. By 2022, terms like work, road, project, Mumbai, traffic, and station reflect an emphasis on infrastructure projects. This shift highlights the government's focus on economic recovery through job creation and urban development.

The trend continued into 2023, where terms like govt, road, project, power, and school suggest a broader governance focus, balancing infrastructure with public welfare initiatives. The inclusion of school points to a renewed emphasis on education, possibly reflecting efforts to address the disruptions caused by the pandemic.

This period illustrates how infrastructure projects often serve as visible markers of governance success, signaling both progress and political priorities. The focus on urban centers like Mumbai and Delhi also underscores regional inequalities, as development often concentrates in already-privileged areas.

This decade of headlines tells a story of a nation in transition—shaped by its democratic processes, challenged by crises, and driven by aspirations for development. The analysis of these topics not only highlights what captured public attention but also offers insights into the priorities and challenges that define contemporary India. By exploring these trends, we better understand the evolving narrative of a nation poised between tradition and modernity, grappling with its past while striving for a brighter future.

Discussion & Future Work

The dataset used in this analysis is undoubtedly one of its key strengths. With nearly 2 million news headlines spanning over a decade, it offers robust coverage of events, capturing both major national trends and local developments. This vast scale ensures a rich historical archive of societal priorities, political shifts, and cultural dynamics. By narrowing the focus to the last decade, the analysis maintains its relevance to current societal trends while leveraging a significant portion of the data to provide a comprehensive view of public discourse in recent years.

The decision to adopt a temporal, year-by-year approach further strengthens the study by allowing for a granular understanding of how public discourse has evolved over time. This method illuminates shifts in societal priorities, highlighting the ebb and flow of topics like governance, healthcare, and crime. Such a temporal focus provides deeper insights into the dynamics of Indian society, showing how national events, policy changes, and global phenomena have influenced media narratives.

However, the inherent nature of headlines poses certain challenges. Headlines are brief and often lack detailed context, which can limit the richness of topics extracted. Their ambiguity or vagueness, compounded by the absence of accompanying article text, means that certain nuances of the issues discussed may remain obscured. This brevity restricts the depth of analysis, as the true extent of a topic's complexity may not always be captured in a single sentence.

Another critical limitation arises from the dataset's exclusive reliance on the Times of India. While this publication is a leading news outlet, its editorial priorities and biases reflect the perspectives of urban, English-speaking audiences. The exclusion of regional and vernacular media reduces representativeness, potentially overlooking issues and concerns prevalent in rural areas or among non-English-speaking communities. This bias underscores the need for broader data sources to provide a more holistic view of public discourse in India.

The interpretation of the results also introduces a layer of ambiguity. While the top words for each topic provide valuable insights, their meanings are often context-dependent. Words like work, project, and development can signify vastly different themes based on the year or event they are associated with. This subjectivity in interpretation may lead to varying conclusions, highlighting the need for more precise contextual analysis.

Future research could address these limitations by incorporating the full text of articles rather than relying solely on headlines. This would provide a richer context for analysis, enabling the identification of more nuanced themes and offering a deeper understanding of the issues covered. Additionally, expanding data sources to include multiple publications—especially regional and vernacular outlets—would help address the current bias, ensuring a more balanced representation of India’s diverse public discourse.

Another promising avenue for future work is comparative analysis. Combining media coverage with public opinion data, such as social media posts or surveys, could validate the findings and offer insights into how the media shapes and reflects societal priorities. Finally, network analysis of topics could deepen our understanding of how themes intersect and evolve over time. For instance, exploring how governance and healthcare are connected or how education and infrastructure overlap could provide a richer, multidimensional perspective on societal trends. These enhancements would not only strengthen the analytical framework but also broaden the scope and applicability of the insights generated.

Submitted by: Manjiri Girish Satam (235789)

Conclusion

This research illustrates how Indian news headlines, over the past decade, have mirrored the country's socio-political and economic evolution. Through topic modeling, the study reveals recurring themes such as governance, crime, public health, and infrastructure development. These themes highlight the persistent challenges and shifting priorities that define public discourse. The analysis demonstrates how events like the COVID-19 pandemic caused significant disruptions in societal focus, temporarily sidelining other concerns like crime and governance. It also shows how certain issues, such as gender-based violence and judicial matters, remain enduring concerns, reflecting systemic challenges that continue to demand attention.

By tracking the evolution of these topics year by year, this research uncovers not just the key narratives but also the underlying forces shaping them—be it elections, policy reforms, or external crises. This temporal mapping offers a nuanced understanding of how public discourse adapts to changing circumstances while retaining core issues of national importance.

The findings from this study have broader implications for how media, society, and policy intersect. They underscore the central role of media in shaping public attention and influencing perceptions, revealing both its strengths in amplifying critical issues and its limitations in providing balanced coverage across regions and topics. Moreover, the analysis highlights the potential for computational tools like NLP to provide a structured and scalable approach to understanding societal priorities. By extending this work to include regional, vernacular, or alternative media sources, future studies could address the biases inherent in relying on a single source and uncover a more inclusive picture of India's public discourse.

---

[1] Scope of Journalism in India
LNCT. (n.d.). Scope of Journalism in India: The Role of Media.
Retrieved from https://lnct.ac.in/scope-of-journalism-in-india-the-role-of-media/
[2] Press Council of India Report
Press Council of India. (n.d.). The Role of Media in Present Day Context.
Retrieved from https://www.presscouncil.nic.in/Pdf/MEFI_46th.pdf
[3]Times of India. https://timesofindia.indiatimes.com/
[4] Audit Bureau of Circulation Report
Audit Bureau of Circulation. (2022). Highest Circulated Newspapers (Across Languages). Retrieved from https://www.auditbureau.org/files/JD%202022%20Highest%20Circulated%20(across%20languages).pdf
[5] Source Dataset
Kaggle. (n.d.). India Headlines News Dataset.
Retrieved from https://www.kaggle.com/datasets/therohk/india-headlines-news-dataset/data
[6] Topic Modeling: A Comprehensive Review
Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. EAI Endorsed Transactions on Scalable Information Systems, 6(21), e2. https://doi.org/10.4108/eai.13-7-2018.159623
[7] Determining the Number of Topics in Topic Modeling
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics, 16(Supplement 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

GitHub Repository: https://github.com/Manjiri-Satam/NLP_ResearchNote

Submitted by: Manjiri Girish Satam (235789)