

Text and Context in Understanding and Interpreting a State's Preferences through Text-mining of the UN General Assembly Speeches

Fasih Ahmed, Sabahat Khan and Sohail Ahmad
COMSATS University, Islamabad

This research paper explores India's speeches in United Nations General Debate (UNGD) sessions to investigate the nation's priority settings and challenges faced during the last five decades. The study adopted Latent Dirichlet Allocation (LDA), a computational linguistics technique to uncover hidden topics in a selected corpus. The data were accessed from the United Nations General Debate Corpus. The research underpins Computational Grounded Theory, which relies on the computational exploration of the text. The results reveal that world peace, nation-building, poverty eradication, economic development, women empowerment, dialogue to resolve issues, and terrorism as the global threat are the main issues and challenges faced by India. Through topic modeling and word clustering, it becomes clear that India wants to eradicate poverty and terrorism through the joint effort of the whole world .

Keywords: topic modeling, natural language processing, computational grounded theory, UNGA Speeches, Poverty Eradication

The UN General Debate (UNGD) session is a reliable forum where all the countries express issues and concerns. It is a forum to resolve global matters through negotiations (Baturu et al., 2017). All the member countries bring their commitments and concerns to this platform to inform the world and to find possible solutions through common consensus (Peterson, 2006). The study has been delimited to the UN general assembly speeches of India that have been investigated to discover the hidden topics to understand the nation's priority setting. The statements made during the General Debate (GD) sessions are an invaluable and untapped source of information on the government's policy preferences across various issues over time (Baturu et al., 2017).

The rationale for selecting the time frame from 1970 to 2018 is that multiple new developments took place during this time, especially in the context of South Asia. The conflict between Afghanistan and Russia rose, especially after 1970 (Arnold, 1985; Sajad, 2022). The existence of a new neighboring country in the form of Bangladesh resulted in strenuous relations with the other neighboring country, Pakistan (Bose, 2005). The Cold War lasted for multiple decades and affected international relations (Gordon, 1995). In 1990, the fall of the USSR and the government of the Taliban also brought new challenges to many South Asian countries (Lamb et al., 2014). Lastly, the waves of terrorism and the event of 9/11 brought multiple new narratives (Ashraf, 2007). The common challenges faced by South Asian countries may be summed up as poverty, unemployment, malnutrition, unplanned urbanization, and illiteracy (Malhotra, 2011). India is the biggest and economically better country compared to the other countries in South Asia, and it has a vital role in the world (Wagner, 2016). From this point of view, it is indispensable to understand the preferences and issues conveyed to the World by Indian leaders through the platform of UN General Assembly sessions. The strategic socio-economic and demographic position of India makes it the centre of attention for other countries in the world (Crane, 1966; Kapur, 2010).

On the other hand, topic modeling is a novel approach to exploring hidden topics from a corpus. It relates to specific words occurring in any document (Stoica et al., 2020). In social sciences, topic modeling helps find the hidden topics from a selected corpus spanned over a specific time (Ahmed, et al., 2022; Ahmed, et al., 2022). Likewise, topic modeling is a machine learning technique that scans a set of documents by detecting the word and phrase pattern. It automatically forms a word cluster defining a document set (Reisenbichler & Reutterer, 2019). Therefore, this approach combines similar ideas and presents them in one topic from a large corpus (Jacobs & Tschötschel, 2019). This study uses topic modeling to investigate the hidden topics in speeches delivered from 1970 to 2018.

The study is significant because it provides a country's priority settings, including its issues and concerns raised at the world's highest forum. Various researchers have investigated the UN General Debate sessions using machine learning techniques (Baturu et al., 2017; Gurciullo & Mikhaylov, 2017) to investigate the priority settings and policy

preferences of all the UN member countries. Compared to these studies, the present study focuses on the priority settings regarding one country (i.e., India). It is helpful because it provides the priority settings and preferences of the Indian government.

Literature Review

The analysis of speeches usually involves discourse analysis. In simple words, discourse analysis is the analysis of language use that focuses on co-textual relations in the language (Brown et al., 1983; Widdowson, 2007). Discourse analysis encompasses the language used and the proper context in which it reveals meaning (Boréus & Bergström, 2017). It is a widely used research method to analyze political discourse, whether spoken or written (Bonyadi, 2019). Different studies have been conducted to analyze political discourses (Chen, 2018; Latupeirissa et al., 2019; Mohammadi & Javadi, 2017).

Various researchers have conducted the discourse analysis of the UN speeches, which is important to review at this stage. These studies normally include a comparison of speeches from two or more leaders looking at the aspects of discourse to study their ideologies (Rababah & Hamdan, 2019; Shah & Alyas, 2019; Sharififar & Rahimi, 2015; Zhu & Wang, 2020). These studies primarily focus on the polarization of "Self" and "Other", power relations, and discourse ideologies. Discourse analysis helps identify intertextualities such as speech preferences and persuasiveness on the speaker's part. It is vital to notice that traditionally the studies in discourse analysis have focused more on the aspect of intertextuality concerning a limited number of leaders who delivered their speeches at the UN General Assembly forum (Aman et al., 2020; Najarzagdegan et al., 2017; Shakoury, 2018). However, an analysis of speeches delivered by a single state at the UN General Assembly is yet to be conducted to investigate a state's preferences and priority settings irrespective of individual personalities who delivered speeches at the UN General Assembly forum.

On the other hand, machine learning approaches provide a robust analysis to analyse the larger chunk of textual data. Topic modeling, in this respect, is a novel text-mining tool for analyzing a large corpus through unsupervised machine learning, automatically discovering thematic information (Aziz & Ahmed, 2024; Khan et al., 2022; Obeidat et al., 2024). Topic modeling is based on probabilistic models that unveil how much a portion of a specific document relates to a particular topic (Blei et al., 2003). It helps in lessening the complexities within the large corpora by assigning specific topics to the whole text (Jacobs & Tschötschel, 2019). Topic modeling has multiple advantages, and firstly, it can deal with a large corpus of text. Secondly, it can deal with a text prolonged over a span of time to understand the process of meaning-making in a larger context, and thirdly, the group texts into 'topics' that are truly thematically coherent with a mixed degree of success, (Brookes & McEnery, 2019). Due to these advantages, topic modeling has been used for discourse analysis to investigate media framing (Ahmed & Khan, 2022; Ahmed, Mubeen, et al., 2022; Heidenreich et al., 2019), mining ethnicity (Viola & Verheul, 2020), climate change (Rabbani & Ahmed, 2025) contrastive perspectives (Stine & Agarwal, 2020), and economic dynamics (Ahmed, Nawaz, et al., 2022; Shahid & Shaikh, 2019). In addition, topic modeling has been equally beneficial in revealing policy analysis (Hagen et al., 2015; Isoaho et al., 2021; Valdez et al., 2018).

Topic modeling, especially Latent Dirichlet Allocations (LDA) uses unsupervised machine learning methods which are near to Grounded Theory in approach (Hannigan et al., 2019). In unsupervised machine learning, the researcher is blind and uses the data-driven approach (Baumer et al., 2017). The data drives research toward the meaning-making process rather than making prior assumptions or hypotheses (Jiang et al., 2021). Grounded Theory has been extended as Computational Grounded Theory (CGT) which is similar to GT in the process of meaning-making but may differ in data collection. Baumer et al. (2017), explaining the similarity between GT and CGT, regard it as "extreme divergence" or "unlikely convergence". CGT involves three main steps; pattern detection (i.e. using computational techniques to reduce complicated messy text into a simpler list of words), pattern refinement (i.e. re-engagement with the data through a computationally guided data), pattern confirmation (i.e. the validity of the inductively identified patterns) (Nelson, 2020). CGT bridges the gap between social sciences and other disciplines for a collaborative step to understand political and social matters through the lens of new and emerging methods, such as topic modeling, to gain an insight into matter through the use of technology (Ophir et al., 2020). In this context, various researchers have highlighted the role of computer-assisted communication concerning the Grounded Theory which opens new avenues of research and data analysis (Carlsen & Ralund, 2022; Pospiech & Felden, 2013).

Method

Topic Modeling

Topic modeling reveals patterns in large datasets by finding the word clusters that are more habitual in a text. These word clusters often occur together, and these occurrences carry traces of meaning (Ylä-Anttila et al., 2018). These word clusters appear in topics modeling as keywords in a group. Hence topic modeling helps in the meaning-making process. In the case of the present study, the meaning-making process focused on the challenges and the priority settings through the UN General Assembly speeches delivered by Indian leaders from 1970 to 2018.

We extracted these speeches from the UN General Debate Corpus (Jankin Mikhaylov et al., 2017). This corpus provides speeches delivered by country leaders. Later, we extracted the speeches delivered by the Indian leaders at the UN General Assembly forum. We thoroughly read some of the speeches to analyze how divergent these speeches were. These speeches consisted of multiple paragraphs. A paragraph reflects a specific topic of the larger discussion, we analyzed each paragraph as a case in topic modeling. A total of 1615 paragraphs were extracted as cases to investigate the prevalent topics from the corpus. Tang et al., (2014) recommend documents of at least 100-200 words in length, while the number of documents needs to be at least 1000. As we have taken paragraphs that vary in size, the number of documents in the form of paragraphs is well enough for the recommended value.

Preprocessing

Preprocessing of the data is a crucial step because it influences the final results and interpretability (Isoaho et al., 2021; Vijayarani et al., 2015). We extracted the data from the UN General Debate Corpus (Jankin Mikhaylov et al., 2017). The textual data was divided into paragraphs, as explained earlier. The preprocessing of the data was conducted in steps.

In the first step, all the paralinguistic features were removed. Paralinguistic features include hyperlinks, file headers, markups, and metadata (Schuller & Batliner, 2013). All the text was converted to lowercase. These features make the data noisy, which later affects the results. In the second step, the process of tokenization was conducted. Tokenization is the process of defining the word and sentence boundaries to convert the text into meaningful elements (Kannan et al., 2014; Vijayarani et al., 2015). The third step of preprocessing involved stemming. In stemming, all the words were brought back to their root forms. In addition, the stop words were removed. These words contribute less to meaning-making and are removed before the final analysis (Yao & Ze-wen, 2011). The stop words include prepositions, articles, conjunctions, etc. In addition, the names of people, places, and cities were also added to the list of stop words as they make the data noisy. In the last step, the bigram and trigrams were treated as single terms. The terms that appeared more than 100 times together as a word cluster of two or three words have been treated as a single term. The benefit of using bigrams and trigrams is understanding the local context and avoiding real-world errors (Samanta & Chaudhuri, 2013).

Validation

We applied the Latent Dirichlet Allocations (LDA) method of topic modeling and followed the validation process, including statistical validation, semantic validation, and predictive validation (DiMaggio et al., 2013). Statistical validation involves running the different parameters and finding better values for the number of topics. All this process helps find the right number of topics which is part of statistical validation. In the case of the present study, keeping in view the length and number of documents, we ran the model several times to decide upon the right number of topics. We decided to limit the number of topics to 7 as it seemed more appropriate for the study. Secondly, the semantic validation was conducted keeping in view interpretability and plausibility. The group of keywords was given appropriate themes to interpret a given situation more aptly. This also involved going back to the original data and reading it out thoroughly to analyze a more plausible theme. Grimmer and Stewart (2013) recommend that social sciences researchers review the original text to interpret the matter under discussion better. Finally, the predictive validation process involves confirming the model reflects a relevant collection of external information (Ahmed, 2025). The external events confirm what the topics predict in the model.

Computational Grounded Theory

The present study underpinned Computational Ground Theory for its theoretical framework (Nelson, 2020). It is the extended form of Grounded Theory, "a theory derived from data systematically gathered and analyzed through the research process" (Morse et al., 2016). The grounded theory proposes that the meaning is entirely grounded in the text. It cannot be found from external links or other theories. Data guides the researcher to get to the core of the real problem (Charmaz & Thornberg, 2020). CGT almost follows the principles as explained by the Grounded Theory, but differs from it on the grounds of data collection to meet the contemporary challenges due to the involvement of technology. CGT has

also taken a shift from the qualitative to the quantitative method in dealing with the data. It follows three main steps; pattern detection, pattern refinement, pattern confirmation mainly done through computation (Nelson, 2020).

The present study comes under the umbrella of CGT in the sense that it has accessed the data from the UN General Debate Corpus (Jankin Mikhaylov et al., 2017). Later the data is processed through the computer to change it to a simple list of words for data analysis (i.e. pattern detection). Later on, LDA was applied to generate topics from the text (i.e. pattern refinement). Lastly, the topics have been labelled and confirmed by going back and forth to the original data (i.e. pattern confirmation).

Data Analysis

The topics have been inferred from the given keywords to understand all the keywords' meanings. The results obtained through LDA and the suggested titles of extracted keywords are given below.

Table 1

Extracted Keywords along with Suggested Topics and Explanations

Keywords	Suggested Topics
[(0, '0.048*"terrorism" + 0.032*"international" + 0.030*"must" + 0.027*"peace" + '0.016*"security" + 0.014*"today" + 0.013*"begin" + 0.013*"state" + '0.013*"world" + 0.013*"include"'),	World Peace
(1, '0.031*"nation" + 0.027*"need" + 0.025*"also" + 0.020*"would" + '0.017*"effort" + 0.017*"new" + 0.014*"time" + 0.013*"people" + 0.013*"many" + '0.013*"work"'),	Nation Building
(2, '0.042*"action" + 0.029*"address" + 0.027*"poverty" + 0.022*"place" + '0.022*"india" + 0.022*"adopt" + 0.021*"long" + 0.018*"policy" + '0.017*"launch" + 0.016*"go"'),	Poverty Eradication
(3, '0.071*"country" + 0.051*"develop" + 0.044*"development" + 0.031*"economic" + '0.027*"reform" + 0.017*"sustainable" + 0.016*"process" + '0.016*"financial" + 0.016*"like" + 0.015*"poor"'),	Economic Development
(4, '0.028*"must" + 0.023*"issue" + 0.021*"make" + 0.018*"goal" + '0.018*"programme" + 0.017*"provide" + 0.016*"woman" + 0.015*"nature" + '0.014*"achieve" + 0.012*"technology"'),	Women Empowerment
(5, '0.043*"talk" + 0.031*"fight" + 0.024*"hope" + 0.018*"victim" + '0.017*"agree" + 0.017*"pain" + 0.017*"role" + 0.016*"find" + '0.015*"resolve" + 0.014*"contribute"'),	Dialogue to Resolve Issues
(6, '0.043*"world" + 0.038*"terrorist" + 0.032*"year" + 0.029*"challenge" + '0.024*"become" + 0.020*"change" + 0.019*"global" + 0.014*"ask" + '0.014*"even" + 0.013*"meet"')]	Terrorism as a Global Threat

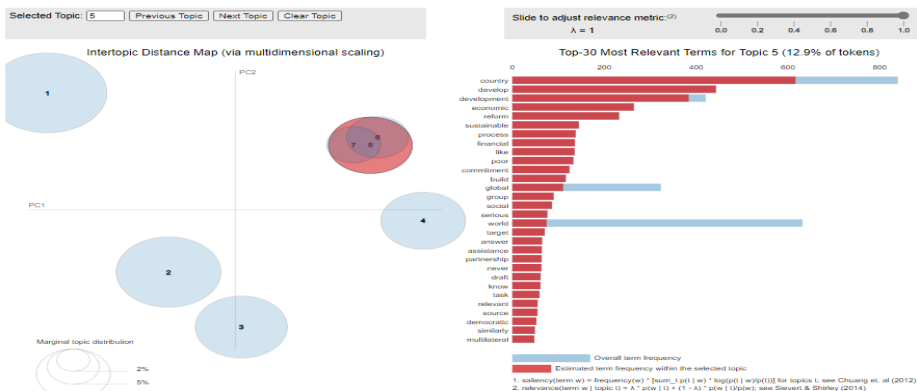


Figure 1: LDA Topic Visualization

The first set of keywords highlights the issue of terrorism (0.048) and the importance of peace (0.027) and security (0.016) in the world. The world (0.013) needs a joint effort to eradicate terrorism and all states (0.013) should be included (0.013) in this campaign. The weight of 'terrorism' is higher than the weights of other keywords in the respective set, therefore, it points towards the need for dialogue to fight against terrorism.

The second set of keywords suggests the need (0.027) to combine effort (0.017) to involve people (0.013) and work (0.013) together for the sake of the nation (0.031). The word 'nation' has the highest weightage of 0.031, which points toward nation-building.

The third set of keywords reflects the issue of poverty (0.027) eradication. The world should take strong action (0.042) to address (0.029) the issue of poverty and to place (0.022) India (0.022) in a better position. India needs to adopt (0.022) long-standing (0.021) policy measures to combat the issue of poverty.

The fourth set of keywords encompasses the economic (0.031) development (0.044) of the country (0.071), which is the backbone of any country. The set of keywords also suggest that there is a need to reform (0.027) the policies and to introduce sustainable (0.017) political process (0.016). The word 'economic development' has the highest weightage of '0.031 and 0.044' respectively. Therefore, it points towards the importance of economic development in controlling poverty.

The fifth set of keywords suggests that a country needs to set a vision to upraise the situation of women (0.016). This is one of the core issues (0.028), so there should be a proper dialogue to resolve these issues (0.023). Women's empowerment can be achieved by providing them the technological (0.012) knowledge. A proper program (0.018) must be launched to make (0.021) women independent and empowered. The word 'goal' has a weightage of 0.018, which shows a tendency to reach the goal of women's empowerment to stabilize a nation.

The sixth set of keywords talks about the need for dialogue (0.043) to resolve (0.015) the national and international issues. The words 'fight' and 'pain' have 0.031 and 0.017 weights respectively, which shows that the pain and suffering of the victims (0.018) can be reduced by giving them hope (0.024) and suggesting measures to have a peaceful country. This goal can be achieved by dialogue. The word 'talk' has the highest weightage of 0.043, which means that the issues can easily be resolved if the nation talks (0.043) about them.

The last set of keywords points toward the global (0.019) threat of terrorism (0.038). It has become (0.024) the biggest hindrance to global development in the last few years. Terrorism is a challenge (0.029) that needs to be ended to bring a positive change (0.020) in the world (0.043). The word 'world' has the highest weightage as compared to other keywords in the set, which shows that the whole world is chained by terrorism.



Figure 2: Word Cloud of Keywords Extracted Through Topic Modeling

The keywords extracted through the process of topic modeling (using LDA) are shown in Figure 2. The font size of the keywords represents the weightage of words that helped assign the topics to the keywords in Figure 2. From the weightage of the keywords suggested in each group, the topics are mostly related to international support, nation's security, terrorism eradication, women empowerment, and eradication of poverty.

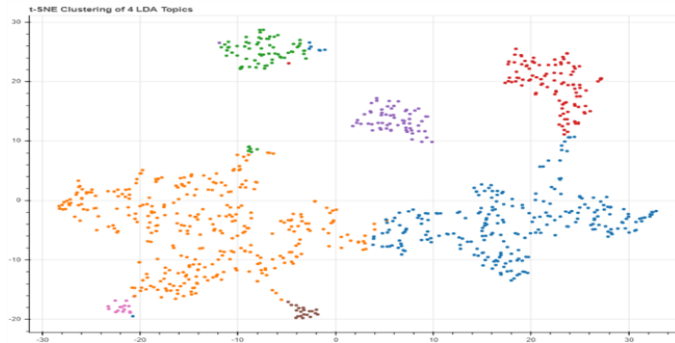


Figure 3: A t-SNE distribution of Topics Conducted through Topic Modeling

The t-SNE distribution of topics clarifies the position of topics in relevance to other topics. Figure 3 shows that the topics are distributed. This distribution shows how the topics are distributed and which of the topics overlap with one another.

Exogenous Validity

Exogenous validity refers to the external factors or the ground realities that endorse the findings of the study. It involves where the study can be generalized or extended to real-world settings (Nowlin, 2016; Quinn et al., 2010). India's UN speeches have been validated according to the topics explained in Table 1. Moreover, the speeches in the UN General Assembly sessions reflect the state's narrative. Hence exogenous validity has been seen through the measures taken by the Indian parliament as it also brings the state's practical approach towards the claims echoed at the UN General Assembly forum.

The range of topics in Table 1 explains that topics 0, 5, and 6 relate to peace and security. Regarding peace and security, various acts have been passed through Indian parliaments, such as the Armed Forces (Special Powers) Act, 1958 (AFSPA), the National Security Act, 1980 (NSA), and the Unlawful Activities (Prevention) Act, 1967 (UAPA) to deal with internal security threats. These measures reflect the government's focus on peace and security in the country. AFSPA was enacted to deal with internal security threats. However, research on AFSPA reveals that AFSPA, in the name of security, emerged as 'draconian' or 'colonial' (Kikon, 2009). "AFSPA operates, the armed forces are alleged to have committed one of the world's least-known abuses of human rights but revel in legal impunity" (Bhattacharyya, 2018). NSA and UAPA also aim to maintain peace and harmony within society and prevent terrorism. However, other views are that 'these laws are least used to serve their intended purpose and have mostly been abused by the governments to silence the dissenters' (Saxena, 2023).

Topics 1, 2, and 3 in Table 1 relate to nation-building, poverty eradication, and economic development in India. In this regard, Mahatma Gandhi National Rural Employment Guarantee Act, 2005 (MGNREGA) and National Food Security Act, 2003 (NFSA) are some measures to improve people's quality and standard of living in India. MGNREGA was introduced in 2005 to uplift the social status of rural areas with a focus on inclusiveness, rural labor markets, and agriculture (Nair et al., 2013; Reddy et al., 2014). Another act regarding basic food essentials is NFSA which is a significant step in fighting against hunger and protecting the rights of the people for food. However, among the main challenges are the increasing population, lack of infrastructure, and operational inefficiencies in implementing this program fully (Sengupta & Mukhopadhyay, 2016; Tanksale & Jha, 2015).

Another topic that emerged through LDA topic modeling belongs to women's empowerment, as explained in Table 1. The Protection of Women from Domestic Violence Act, 2005 (PWDVA), the Prevention of Sexual Harassment at Workplace Act, 2013 (PSHWA), The Criminal Law (Amendment) Act, 2013 (CLAA), The Maternity Benefit (Amendment) Act, 2017 (MBAA), and the Women's Reservation Bill seeks to reserve 33% of seats in the Lok Sabha and state legislative assemblies for women validate the topic of women empowerment in India. PWDVA provided women with the protection to go against domestic violence and empowered them with a legal weapon to fight against it (Bhattacharjee & Pal, 2016; Mukhopadhyay, 2018). However, the ground reality is that the above-counted measures need to be implemented in true letter and spirit (Karanjawala & Chugh, 2009). Similarly, CLAA includes stronger provisions against sexual offences (Bhattacharyya, 2015). Another significant progress is The PSHWA relates to what can be tantamount to sexual harassment and how workplaces need to be proactive about ensuring that they protect the dignity of a woman (Bothra, 2014; Gupta & Garg, 2020). The Maternity Benefit Amendment Act 2017 enhances leave maturity for working

women. Lastly, the Women Reservation bill seeks to reserve 33% of seats in the Lok Sabha and state legislative assemblies for women. All these measures relate to women's empowerment in India which leads to the direction of nation-building (Sharma, 2020).

Discussion

The study analyzed the Indian UN General Assembly Speeches of almost five decades (1970-2018) to discover hidden topics using text mining. For this purpose, LDA was used. This method helps extract the topics from a large corpus of data. The main purpose of the LDA analysis was to highlight the issues and their priority setting set by the Indian government. The study followed Computational Grounded Theory which relies on the principle that the data under consideration drives toward the topic's nature (Nelson, 2020). The overall themes extracted from the corpus revolve around peace, security, and nation-building through poverty eradication, women empowerment, and economic development. All the generated themes advance the understanding of the topics through words like "international", "terrorism", "nation", "action", "country", "reform", "talk" and "terrorist".

The extracted speech data is modeled by the LDA application, which frames the topics mentioned in Table 1. Figures and tables reveal a quantitative description of the data. However, a qualitative description must understand the core issues and priority setting from India's lens. Seven topics emerged from the data. These topics can be divided into two major categories depending on their relevance to the nature of problems.

The first category of topics relates to peace and security. Out of the seven suggested topics in Table 1, three of them relate to peace and security. It reveals the stress on peace and the challenges faced by the nation to overcome the menace of terrorism. Terrorism has been a threat in the region for a longer time (Subramaniam, 2012). The previous studies stress revamping policy on terrorism on the part of the Indian state (Mann, 2001; Narain & Rajakumar, 2016). India focuses on world peace and for achieving that, the stress is given to dialogue. A proper dialogue is needed to resolve the issues that are a hindrance to achieving world peace. Terrorism is a threat to national security and driving attention to peace and security is the only way to move forward (Kamath, 2001). As a result, India's priority on peace is evident in speeches at the UN General Assembly forum.

The second category of suggested topics comes under the umbrella of nation-building. India focuses on nation-building to maintain peace and security in the world. A total of four topics point toward nation-building out of the seven extracted topics. Therefore, India is more inclined toward building a nation. India links nation-building with poverty eradication and women empowerment. If women are empowered, the nation will be lifted automatically (Habrigh et al., 2021). Moreover, two-thirds of the Indian population lives in poverty. More than 66.5% of the population lives on approximately \$2 per day (Mishra, 2020). To build a nation, the government must ensure that no one is left behind. People should not be below the average poverty line to maintain a life balance (Singh & Chudasama, 2020). On the other hand, the soft image through nation-building is the primary purpose where the world gets the impression of emerging India (Biswal, 2020). The topic of 'women's empowerment is an evident example of conveying the message of soft image (Mokta, 2014).

It is essential to mention that the extracted two categories overlap with each other. It is evident from the topics that peace and security are linked with nation-building and vice versa. In other words, when women are empowered, and poverty eradicates, the nation can stand on its own. Similarly, peace and security lead to the stability of a country. A nation cannot focus on all the problems simultaneously.

Another indication through the text is essential to discuss that these 49 years of speeches may indicate India's priority setting and policy shift in the future. This priority setting and policy shift may focus on nation-building and improving peace and security as a prime tool in the country. The other reason for raising these points at the UN General Assembly forum is to convince the world of the state's policy and to get moral and diplomatic support on multiple issues related to peace and security.

The study underpinned Computational Grounded Theory which guided the structure of the study. CGT mainly focuses on three main steps; pattern detection, pattern refinement, and pattern confirmation (Nelson, 2020). The pattern detection phase helped in conducting the analysis and identifying the significant latent topics from the corpus. The pattern refinement phase guided the study in refining the themes from topic models. Lastly, the pattern confirmation, phase guided the study in confirming the emerged themes through exogenous validity by confirming the research results from the external sources. In addition, it guided in exploring how data unveils the reality by adopting the data-driven approach where the main point in CGT is that the data drives what the reality. CGT, as compared to traditional Grounded Theory,

provides space to confirm the results from external sources where results and generalizability is observed and text is analyzed concerning context (Boréus & Bergström, 2017).

Conclusion

Based on analysis and discussion, the study concludes that world peace, nation-building, poverty eradication, economic development, women empowerment, dialogue to resolve issues, and terrorism as a global threat are India's leading issues and challenges. The primary focus is on nation-building and economic development that can be achieved by eradicating poverty and terrorism and empowering women. Moreover, India focuses on dialogue to resolve issues and have a peaceful and secure nation. However, through the exogenous validity, it is evident that the steps taken to improve security concerns are criticized on political grounds where the main concern seems to use these laws as a weapon against their opponents and to strengthen their territorial control. The exogeneous validity reveals that research works based on laws passed by the Indian parliament regarding security concerns have been termed as 'draconian' or 'colonial' (Kikon, 2009).

Natural language processing with topic modeling and LDA helps determine the challenges and issues India has discussed in the annual UNGA sessions for two decades. Through topic modeling and word clustering, it becomes clear that India wants to eradicate poverty and terrorism through the joint effort of the whole world. LDA is a robust a viable way to look for a country's preferences and priority settings in the arena of both national and international relations.

The study has been delimited to the UNGA sessions with respect to the focus of a single country to investigate the nations' priority settings. The second delimitation relates to the timeframe from 1970 to 2018. This timeframe provided an overall analysis of the country's priority settings regardless of a specific government agenda. The study has been delimited to the revealed topics. There may be some other topics which have not been touched upon due to the limitation of the study as the study only analyzed the data based on speech corpus.

The study possesses implications for policymakers in determining the right course of future directions. It may be helpful for text analyzers to understand how text can guide the possible understanding of a specific phenomenon. In addition, from the perspective of linguistics, it may be helpful for researchers to understand the relation between text and context as text without context may not be able to bring the right interpretation. Lastly, the study has also implications for text analysts as it provides a guideline on how to deal with text and confirm the results from inside the corpus and from external resources.

Last but not least, text and context in combination provide valuable insights for researchers and policymakers to understand how to understand a nation's priorities. Text, devoid of context, may not be able to bring to the fore an in-depth understanding of the phenomenon under investigation. From this point of view, the study proposes to take into account the contextual factors to validate the textual insights emerging from rhetoric.

References

- Ahmed, F. (2025). Formalizing the Social Aspects of Topic Modeling: Focus on the Social Positioning of Researchers. In *Text Mining in Educational Research: Topic Modeling and Latent Dirichlet Allocation* (pp. 97-106). Springer.
- Ahmed, F., & Khan, A. (2022). Topic Modeling as a Tool to Analyze Child Abuse from the Corpus of English Newspapers in Pakistan. *Social Science Computer Review*, 08944393221132637.
- Ahmed, F., Mubeen, M., & Nawaz, M. (2022). Framing South Asian politics: An analysis of Indian and Pakistani English print media discourses regarding Kartarpur corridor. *PloS one*, 17(2), e0264115.
- Ahmed, F., Nawaz, M., & Jadoon, A. (2022). Topic Modeling of the Pakistani Economy in English Newspapers via Latent Dirichlet Allocation (LDA). *SAGE Open*, 12(1), 21582440221079931.
- Aman, A., Imtiaz, A., & Shahzad, A. K. (2020). Framing in Pakistani and Indian Discourse at the United Nations General Assembly: A Political Discourse Analysis.
- Arnold, A. (1985). *Afghanistan: The Soviet Invasion in Perspective* (Vol. 321). Hoover Press.
- Ashraf, F. (2007). India-Afghanistan Relations: Post-9/11. *Strategic Studies*, 27(2), 90-102.
- Aziz, A., & Ahmed, F. (2024). Ecological Modelling: Applying Computational Linguistic Analysis to the UN Secretary-General's Speeches on Climate Change (2018–2022). *Digital Studies/Le champ numérique*, 14(1).
- Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, 4(2), 2053168017712821.
- Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6), 1397-1410.

- Bhattacharjee, K., & Pal, S. (2016). Protection of women from 'domestic violence' Act 2005: statistics reveal what Society conceals. *Bhattacharyya, R. (2015). Understanding the spatialities of sexual assault against Indian women in India. Gender, Place & Culture, 22(9), 1340-1356.*
- Bhattacharyya, R. (2018). Living with armed forces special powers act (AFSPA) as everyday life. *GeoJournal, 83(1), 31-48.*
- Biswal, S. (2020). Soft power: The Modi way.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3, 993-1022.*
- Bonyadi, A. (2019). Discourse analysis and language pedagogy: A review. *Journal of Teacher Education for Sustainability, 21(1), 128-136.*
- Boréus, K., & Bergström, G. (2017). *Analyzing text and discourse: Eight approaches for the social sciences.* Sage.
- Bose, S. (2005). Anatomy of violence: Analysis of civil war in east Pakistan in 1971. *Economic and Political Weekly, 4463-4471.*
- Bothra, N. (2014). The Sexual Harassment of Women at Workplace (Prevention, Prohibition and Redressal) Act, 2013. Available at SSRN 2498990.
- Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies, 21(1), 3-21.*
- Brown, G., Brown, G. D., Brown, G. R., Yule, G., & Gillian, B. (1983). *Discourse analysis.* Cambridge university press.
- Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society, 9(1), 20539517221080146.* <https://doi.org/10.1177/20539517221080146>
- Charmaz, K., & Thornberg, R. (2020). The pursuit of quality in grounded theory. *Qualitative Research in Psychology, 1-23.*
- Chen, W. (2018). A Critical Discourse Analysis of Donald Trump's Inaugural Speech from the Perspective of Systemic Functional Grammar. *Theory and Practice in Language Studies, 8(8), 966-972.*
- Crane, R. I. (1966). THE STRATEGIC SIGNIFICANCE OF THE ASIAN SUBCONTINENT. *Naval War College Review, 19(1), 8-19.* <http://www.jstor.org/stable/44635081>
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41(6), 570-606.*
- Gordon, S. (1995). South Asia after the cold war: winners and losers. *Asian Survey, 35(10), 879-895.*
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis, 21(3), 267-297.*
- Gupta, D., & Garg, J. (2020). Sexual harassment at workplace. In *International Journal of Legal Science and Innovation.*
- Gurciullo, S., & Mikhaylov, S. J. (2017). Detecting policy preferences and dynamics in the un general debate with neural word embeddings. 2017 International Conference on the Frontiers and Advances in Data Science (FADS),
- Habrigh, V., Bobek, V., & Horvat, T. (2021). The Influence of Economic Activity of Women in Malaysia and Guatemala on National Development. In *Emerging Markets.* IntechOpen.
- Hagen, L., Uzuner, Ö., Kotfila, C., Harrison, T. M., & Lamanna, D. (2015). Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. 2015 48th Hawaii International Conference on System Sciences,
- Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals, 13(2), 586-632.*
- Heidenreich, T., Lind, F., Eberl, J.-M., & Boomgaarden, H. G. (2019). Media framing dynamics of the 'European refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies, 32(Special_Issue_1), i172-i182.*
- Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal, 49(1), 300-324.*
- Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology, 22(5), 469-485.*
- Jankin Mikhaylov, S., Baturo, A., & Dasandi, N. (2017). *United Nations General Debate Corpus Version V6* [text corpus collection of speeches]. Harvard Dataverse. <https://doi.org/doi:10.7910/DVN/0TJX8Y>
- Jiang, J., Adam, M., & Benlian, A. (2021). Algoactivistic Practices in Ridesharing-A Topic Modeling & Grounded Theory Approach.
- Kamath, P. (2001). Terrorism in India: Impact on national security. *Strategic Analysis, 25(9), 1081-1987.*
- Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks, 5(1), 7-16.*
- Kapur, A. (2010). *India and the South Asian strategic triangle.* Routledge.
- Karanjawala, T., & Chugh, S. (2009). The legal battle against domestic violence in India: Evolution and analysis. *International Journal of Law, Policy and the Family, 23(3), 289-308.*

- Khan, S., Ahmed, F., & Mubeen, M. (2022). A Text-Mining Research Based on LDA Topic Modelling: A Corpus-Based Analysis of Pakistan's UN Assembly Speeches (1970–2018). *International Journal of Humanities and Arts Computing*, 16(2), 214-229.
- Kikon, D. (2009). The predicament of justice: fifty years of Armed Forces Special Powers Act in India. *Contemporary South Asia*, 17(3), 271-282.
- Lamb, R. D., Hameed, S., & Halterman, J. (2014). Regional dynamics and strategic concerns in South Asia. In: Center for Strategic and International Studies Washington, DC.
- Latupeirissa, D. S., Laksana, I. K. D., Artawa, K., & Sosiowati, I. G. A. G. (2019). On Political Language Ideology: Critical View of Indonesian President Speech. *Journal of Language Teaching and Research*, 10(4), 843-850.
- Malhotra, N. (2011). South Asia-Political and Economic Region. *The Association for Geographical Studies*, 11(2), 1-33.
- Mann, P. (2001). Fighting terrorism: India and central Asia. *Strategic Analysis*, 24(11), 2035-2054.
- Mishra, S. (2020). Poverty Reduction under Economic Reforms. *Journal of Retail Marketing & Distribution Management*, 1(1), 22-25.
- Mohammadi, M., & Javadi, J. (2017). A critical discourse analysis of Donald Trump's language use in US presidential campaign, 2016. *International Journal of Applied Linguistics and English Literature*, 6(5), 1-10.
- Mokta, M. (2014). Empowerment of women in India: A critical analysis. *Indian Journal of public administration*, 60(3), 473-488.
- Morse, J. M., Bowers, B., Stern, P. N., Corbin, J., Charmaz, K., & Clarke, A. E. (2016). *Developing grounded theory: The second generation* (Vol. 3). Routledge.
- Mukhopadhyay, A. (2018). Found and Lost in Translation: Exploring the Legal Protection of Women from the Domestic Violence Act 2005 Through the Social Public Space of Kolkata. *Social & Legal Studies*, 28(3), 349-369. <https://doi.org/10.1177/0964663918769473>
- Nair, M., Ariana, P., Ohuma, E. O., Gray, R., De Stavola, B., & Webster, P. (2013). Effect of the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) on malnutrition of infants in Rajasthan, India: a mixed methods study. *PloS one*, 8(9), e75089.
- Najarzadegan, S., Dabaghi, A., & Eslami-Rasekh, A. (2017). A Critical Discourse Analysis of Iran and US Presidential Speeches at the UN: The Sociopragmatic Functions. *Theory & Practice in Language Studies*, 7(9).
- Narain, A., & Rajakumar, V. (2016). Revamping India's Counter-Terrorism Approach. *Singapore*. Retrieved December 5th.
- Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3-42.
- Nowlin, M. C. (2016). Modeling issue definitions using quantitative text analysis. *Policy Studies Journal*, 44(3), 309-331.
- Obeidat, M. M., Haider, A. S., Tair, S. A., & Sahari, Y. (2024). Analyzing the Performance of Gemini, ChatGPT, and Google Translate in Rendering English Idioms into Arabic. *FWU Journal of Social Sciences*, 18(4).
- Ophir, Y., Walter, D., & Marchant, E. R. (2020). A collaborative way of knowing: Bridging computational communication research and grounded theory ethnography. *Journal of Communication*, 70(3), 447-472.
- Peterson, M. (2006). *The UN General Assembly*. Routledge.
- Pospiech, M., & Felden, C. (2013). A descriptive big data model using grounded theory. 2013 IEEE 16th International Conference on Computational Science and Engineering,
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209-228.
- Rababah, A. G., & Hamdan, J. M. (2019). A Contrastive Critical Discourse Analysis of Netanyahu's and Abbas's Speeches on the Gaza War (2014). *Journal of Language Teaching and Research*, 10(1), 178.
- Rabbani, S., & Ahmed, F. (2025). Ecological modelling: A computational analysis of air pollution discourses in English print media of India and Pakistan. *PloS one*, 20(2), e0315087.
- Reddy, D. N., Reddy, A. A., & Bantilan, M. (2014). The impact of Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) on rural labor markets and agriculture. *India Review*, 13(3), 251-273.
- Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89(3), 327-356.
- Sajad, M. B. (2022). Invasion, war and Politics: Study of Afghanistan's Political system and Political Parties. Available at SSRN 4073214.
- Samanta, P., & Chaudhuri, B. B. (2013). A simple real-word error detection and correction using local word bigram and trigram. Proceedings of the 25th conference on computational linguistics and speech processing (ROCLING 2013),
- Saxena, S. (2023). Crushing Dissent Under the Veil of National Security: A Critical Analysis of the Provisions of the UAPA and the NSA. *System*.
- Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.
- Sengupta, P., & Mukhopadhyay, K. (2016). Economic and environmental impact of national food security act of India. *Agricultural and Food economics*, 4, 1-23.

- Shah, M. I., & Alyas, R. (2019). A Critical Discourse Analysis of Imran Khan's Speech at Global Peace and Unity Forum. *language, Shahid, S., & Shaikh, M. A.* (2019). Impact of " WhatsApp Chaupal" on the Academic Performance of Graduate Students of Karachi—A Case Study. *FWU Journal of Social Sciences, 13*(2), 94-107.
- Shakoury, K. (2018). *Critical discourse analysis of Iranian Presidents' Addresses to the United Nations General Assembly (2007-2016)* University of Saskatchewan].
- Sharififar, M., & Rahimi, E. (2015). Critical discourse analysis of political speeches: A case study of Obama's and Rouhani's speeches at UN. *Theory and Practice in Language Studies, 5*(2), 343.
- Sharma, E. (2020). Women and politics: A case study of political empowerment of Indian women. *International Journal of Sociology and Social Policy, 40*(7/8), 607-626.
- Singh, P. K., & Chudasama, H. (2020). Evaluating poverty alleviation strategies in a developing country. *PloS one, 15*(1), e0227176.
- Stine, Z. K., & Agarwal, N. (2020). Comparative discourse analysis using topic models: Contrasting perspectives on China from Reddit. International Conference on Social Media and Society,
- Stoica, A., Pu, K. Q., & Davoudi, H. (2020). NLP Relational Queries and Its Application. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI),
- Subramaniam, A. (2012). Challenges of protecting India from terrorism. *Terrorism and Political Violence, 24*(3), 396-414.
- Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. International Conference on Machine Learning,
- Tanksale, A., & Jha, J. (2015). Implementing national food security act in India: issues and challenges. *British Food Journal, 117*(4), 1315-1335.
- Valdez, D., Pickett, A. C., & Goodson, P. (2018). Topic modeling: latent semantic analysis for the social sciences. *Social Science Quarterly, 99*(5), 1665-1679.
- Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16.
- Viola, L., & Verheul, J. (2020). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. *Digital Scholarship in the Humanities, 35*(4), 921-943.
- Wagner, C. (2016). The Role of India and China in South Asia. *Strategic Analysis, 40*(4), 307-320.
- Widdowson, H. G. (2007). *Discourse analysis* (Vol. 133). Oxford University Press Oxford.
- Yao, Z., & Ze-wen, C. (2011). Research on the construction and filter method of stop-word list in text preprocessing. 2011 Fourth International Conference on Intelligent Computation Technology and Automation,
- Ylä-Anttila, T., Eranti, V., & Kukkonen, A. (2018). Topic modeling as a method for frame analysis: Data mining the climate change debate in india and the usa.
- Zhu, L., & Wang, W. (2020). A Critical Discourse Analysis of the US and China Political Speeches-Based on the Two Speeches Respectively by Trump and Wang Yi in the General Debate of the 72nd Session of UN Assembly. *Journal of Language Teaching and Research, 11*(3), 435-445.