IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 February 2023, accepted 2 March 2023, date of publication 13 March 2023, date of current version 21 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3256530

RESEARCH ARTICLE

Mapping Climate Themes From 2008-2021—An Analysis of Business News Using Topic Models

SWARNALAKSHMI UMAMAHESWARAN, VANDITA DAR, ELIZA SHARMA, AND JIKKU SUSAN KURIAN

Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Bengaluru 560100, India

Corresponding author: Vandita Dar (vandita.dar@sibm.edu.in)

ABSTRACT India and other developing economies are receiving more attention in the context of climate change due to their rapid rates of economic expansion and large populations. In terms of absolute emissions, India surpassed China and the U.S. in 2018 to become the third-largest emitter. Solving this wicked problem calls for climate action across the stakeholder spectrum involving governments, business communities, and citizens. While extant literature has focused significantly on the role of governments and individual perceptions, the business sector needs to be more represented. In this study, we consider business news media as a platform that reflects the industry engagement in climate change and as a source of information on climate change for business decision-makers. Hence, understanding the topic and themes in the nexus of climate and business is important to evaluate the business sector's stance towards climate change and how it has evolved. This work explores business news related to climate change using natural language techniques. We first experiment with three topic-modeling techniques, such as LDA, NMF, and BERTopic, on the business news and two more benchmark news datasets. Our test data is derived from digital news archives of 'The Economic Times – India's leading business news daily. We evaluate the performance based on quantitative metrics commonly used for topic models. We choose the algorithm that provides the highest precision for climate-specific information represented by the test dataset. We then apply the algorithm with the best performance, as evaluated by the experiment, to a large corpus of Indian climate news from E.T. spanning from 2008 -2021. We present how different themes, including industry engagement, evolved over the last two decades. The results suggest that climate cooperation has the highest contribution in the corpus, with other themes on resource management, energy and business gaining traction in recent years.

INDEX TERMS Climate change, media, topic models, NLP, computational social sciences, experiment.

I. INTRODUCTION

Climate change is the most critical issue of this millennium. India has high stakes in the global debate since it's the third largest emitter and also high on climate vulnerability [1]. In its most recent update to the Paris pledge, the country has further committed to reducing its emission intensity to 35% of 2005 levels by 2030. Emission reduction of this scale requires solid public support across the stakeholder spectrum, including citizens and enterprises. The role of the commercial sector is particularly relevant given its contribution to country-level emissions.

A practical approach toward understanding how different stakeholders engage in climate change can be through the analysis of news coverage. While all news media reflect contemporary public discourse, business newspapers traditionally report news that is more relevant for business decision-makers. Therefore, mapping the topics pertinent to climate change in business media and how such issues have evolved can help us understand how the commercial sector engages with the issue of climate change.

Automated content analysis of large corpora is an established technique in the social sciences. Specifically, the use of topic models for extracting latent topics from newspapers, discussion forums, and other social media content has been gaining traction in extant literature. Among the different

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

Natural Language Processing (NLP) techniques, topic modeling is a machine-learning task that groups documents and words with similar meanings. All topic models yield topics, each ranked based on the specific terms and their relevance to the topic. The topic emerges automatically without prior annotation or labeling of the raw data. In this sense, topic models form an important category of unsupervised techniques within the larger field of Natural Language Processing. The are many mathematical approaches to topic models and an even higher spread of algorithms that learn the parameters of such model approaches. Conventional topic models range from linear algebraic-based techniques such as LSA to more popular probabilistic models like LDA and PLSA. Metrics decomposition-based methods such as NMF have also been a mainstay of NLP applications. The underlying Bag Of Words (BOW) approach is a distinguishing factor among all these approaches. However, the BoW-based topic model is based on the unrealistic assumption that words are independent of each other and relies on frequencies of word occurrences. A more recent development in NLP, is the representation of text using embeddings. Embeddings in the context of NLP are dense representations of units of the text in the vector space in a way that the semantic similarity between the units of text is captured. Although automated content analysis is gaining importance within the climate discourse literature, there is an overreliance on probabilistic topic models [2], [3], [4], [5]. More importantly, the choice of topic models such as LDA is motivated by popular use and is not backed by experiments to evaluate performance metrics especially in the social sciences [6]. Given this background, the objective of this article is to evaluate the performance of conventional BOW-based models with embeddings-based models. Specifically, we run experiments using LDA, NMF, and BERTopic. We choose the most appropriate model based on the evaluation metrics and use it to map themes within the Economic Times corpus.

The article is organized as follows. Section II introduces the literature on automated content analysis of climate news and discusses how topic models are used in the literature. Section III presents the data and methods. Section IV offers the details of the experiment and its results. Section V presents the results of the final implantation of the chosen model.

II. LITERATURE REVIEW

The adoption of NLP techniques in social sciences has seen significant growth in the recent decade. Language is a social construct and can be considered a proxy for behavior [7]. As expressed by people, language may contain rich latent information that can help identify several dimensions and traits of behavior. Therefore, the rapid growth of NLP methods within social sciences is not surprising. The social science applications of NLP range from understanding political affiliations and voter intentions [8], [9], [10], health care delivery [11], [12], media monitoring [13], and improving public policy implementation [14]. The use of NLP in mining social discourse around climate change is still nascent and broadly categorized based on the analysis of social media platforms, online news, and scientific or technical documents. Several studies analyze societal stance on climate change through the study of Twitter and microblogs. For example, Elgasem et al. combine sentiment and networking analysis to identify climate skeptics and accept communities [15]. Stance detection in climate change tweets emerged as a separate subdomain with the release of the SemEval dataset, which defined detecting climate concerns as a task [8]. Several researchers have employed advanced NLP and machine learning techniques for improving stance detection in climate tweets with the Semeval dataset [16].

While content analysis of climate news is an established norm, the use of NLP in analyzing climate news is relatively new. Compared with the studies on social media, news analysis has received less attention [17]. Keller et al. in [2] applied LDA to ∼18000 climate change articles sourced from two Indian dailies. They extracted 29 themes ranging from climate change impact and politics to climate science. Similarly, Bohr [3] segments 52 US newspapers based on geography, partisan bias, circulation, etc. He then uses structural topic models to discover topics and further analyses the influence of these attributes on the topics [3]. Benites-Lazaro et al. [4] use a corpus of news articles between 2007 -2017 to analyze the climate, food, and energy nexus using topic models. From a methodological perspective, most authors have used topic models, especially the popular Latent Dirichlet Allocation (LDA) [2], [3], [4]. Despite the popularity of LDA, it suffers from certain shortcomings. Notably, the foundational assumption of LDA is that documents are an unordered set, i.e., the words that appear in the text are independent of each other [4]. This assumption is common to all models based on the Bag of Words text representation. This assumption conflicts with linguistic theory, which suggests that words in any human language are always connected and sequential [18]. Another aspect of LDA is that it needs the number of topics extracted as input right at the beginning. Knowing the optimal number of issues at the beginning is only feasible in some contexts and hence introduces an element of subjectivity [17]. Finally, labeling the extracted topics is also left to the users' discretion and can vary considerably between coders. Therefore for credible topic extraction as well as validation, additional information from subject matter experts becomes crucial [19]. In this sense, LDA models do not lend to automation. While some of the shortcomings, such as having to specify a specific number of topics, can be countered using matrix factorization methods such as NMF, the underlying issues related to BoW representation remain the same. The dependency on classical topic models is particularly conspicuous given several algorithmic developments in the domain. For example Topic models based on embeddings have gained traction in recent years. Embeddings are dense text representations such that words with semantic similarity

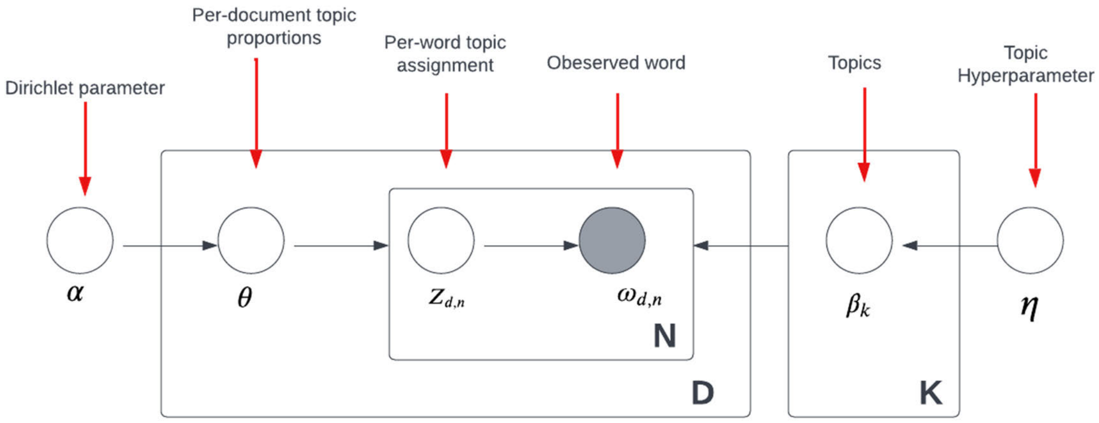


FIGURE 1. Diagrammatic representation of LDA.

are closer in a low dimensional space [20]. Word embeddings capture the relationship between words and convey context better than the word count-based BoW representation of text. In general, embeddings-based models have reported superior performance. However, the application of embeddings based topic models in social sciences has been sparse [21]. Another significant gap in social science based applications is the lack of robust experiments for model selection. Studies typically choose a particular model on a apriori basis based on theoretical assumptions or prior literature contributions [14], [22], [23]

III. METHODS

A. LATENT DIRICHLET ALLOCATION

LDA is the most basic model in probabilistic topic models and is popular in several social science disciplines [5]. LDA typically works on a Bag of Words vector representation of documents of a specific length. The learning algorithm in LDA is unsupervised and is used to discover latent semantic patterns in unstructured documents. The technique assumes that each document is generated by a complex probabilistic generative mechanism.LDA assumes that each document is created one word at a time by selecting a topic from the document's topic distribution and then picking a word from that topic. Therefore, each document is modeled as a mixture of latent topics and each topic is modeled as a multinomial distribution of words. Thus, the learning in LDA is to estimate the parameters of the underlying mechanism that most likely generated the corpus as precisely as possible. Fig 1 provides the outline of the mathematical model

Were

N — Number of words in each document.
k — Number of topics a document belongs to (a fixed number).
D — Corpus, a collection of M documents.
d — Single document.
$z_{d,n}$ —topic for the nth word in the dth document.
θ — Distribution of topics for each document.
β_k — Distribution of words within each topic up to k.
α, η — parameters of prior distributions over θ and β.

The distribution of topics θ derives from a Dirichlet distribution. The use of Dirichlet allows each document to embed topic sparsity, thus mimicking real-word documents. The generative process then can be defined as the joint probability of the documents and topics as represented in

$$p(\theta, z, w \mid \alpha, \beta) \tag{1}$$

where α, and β refer to the hyperparameter related to the Dirichlet distribution. The joint distributions are further computed as the conditional distribution of the hidden topics given the set of documents. The mathematical formulation is given in Eqn 2.

$$p(\theta, z \mid w, \alpha, \beta) = \frac{\boldsymbol{p}(\theta, z, w \mid \alpha, \beta)}{\boldsymbol{p}(w \mid \alpha, \beta)} \tag{2}$$

However, computing the above expression can be difficult since the denominator requires the summation of all possible combinations of topics. The use of approximation methods such as Markov Chain Monte Carlo and variational inference is standard in this context.

B. NON-NEGATIVE MATRIX FACTORIZATION

NMF is a linear algebra-based multivariate technique based on matrix decompositions of a high dimensional vector space. The decomposed non-negative matrices represent hidden structures considered coordinate axes in a transformed low-dimensional vector space [24]. NMF considers every individual document d as a vector of its terms. We then represent the term-document matrix D as follows

$$\mathrm{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n] \in R^{MN} \tag{3}$$

D can be decomposed into low dimensional vector spaces as represented by matrices U & V shown in equation 6. Elements of U and V are non-negative real numbers indicated by the

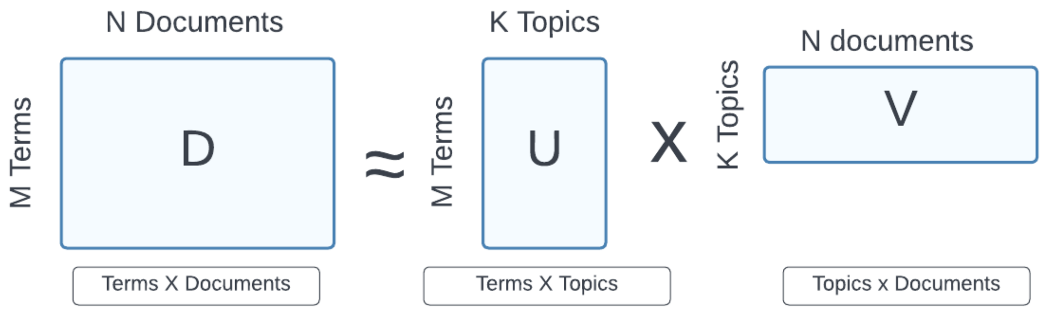


FIGURE 2. Diagrammatic representation of NMF.

matrix R.U has dimensions M and K where M denotes topics and K denotes terms or words

$$\mathbf{U} = [u_1, u_2, \ldots, u_K] \in R^{MN} \tag{4}$$

$$\mathbf{V} = [v_1, v_2, \ldots, v_N] \in R^{KN} \tag{5}$$

$$\mathbf{D} \approx \mathbf{UV} \tag{6}$$

While V represents K topics and N documents. A more intuitive representation of NMF is given in Fig 2. The quality of the formulation in Equation 4 is measured as the squared difference between the document term matrix D and U.V. This can be formally represented below.

$$\text{Min}\, ||\mathbf{D} - \mathbf{UV}||^2 \tag{7}$$

NMF learns U and V iteratively using multiplicative updates such that the error is minimized [25]

C. BERTopic

BERTopic [19] is a combination of several modules. At the top it involves representing text as Embeddings using a pre-trained language model-specifically the Bidirectional Encoder Representations from Transformers(BERT). Such a representation generates a dense matrix where text either in sentences or paragraphs is presented as numeric vectors that capture the semantic similarity. The second layer involves using a dimensionality reduction algorithm such as UMAP. to convert the embeddings into a low-dimensional space. Finally, the reduced embeddings are clustered by HDB-SCAN. or K-means algorithms. At the final layer the topic representations are done such that each cluster is assigned a topic. The topics are then represented using the TF-IDF measure, which combines term frequency and inverse document frequency for rank ordering the importance of words in the text. We define the classical TF-IDF process below

$$\boldsymbol{D_{t,d}} = tf_{t,d} \cdot \log\left(\frac{N}{df_t}\right) \tag{8}$$

where

D - Term document matrix

Tf - Term frequency

Df - Document frequency

BERTopic alters the application of the TF-IDF procedure using a cluster of documents rather than an individual document. This enables the BERTopic to output the topic word distribution for every cluster, thus creating topic representations from clustered embeddings

D. EVALUATION METRICS

When measuring topic quality, human evaluation of a topic is considered a gold standard. However, it is a resource intensive and expensive strategy to implement. As an alternative, recent literature has focused on the automatic measurement of coherence measures that closely mimic human judgment. Topic coherence generally measures the degree to which the words in a topic are semantically related and make sense together [26] Alternatively, coherence can also be defined as a measure of the internal consistency of a topic and how well it represents the documents it is generated from [27].. For measuring semantic similarity, the most common methods are the Normalized Pointwise Mutual Information (NPMI.). The NPMI. method is based on context vectors $\boldsymbol{v}_{\mathrm{j}}$. $\boldsymbol{v}_{\mathrm{j}}$. It is constructed by extracting co-occurrences and counts within a context window of ±n tokens around $\boldsymbol{w}$ [26]. Therefore mathematically, the j$^{\text{th}}$ element of the context vector $\boldsymbol{v}_{\mathrm{i}}$ of word $\boldsymbol{w}_{\mathrm{i}}$ has an NPMI as represented in equation 9.

$$v_{ij} = \text{NPMI}\left(w_i, w_j\right) = \left(\frac{\log \frac{P(w_i, w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log\left(P\left(w_i, w_j\right)+\epsilon\right)}\right) \tag{9}$$

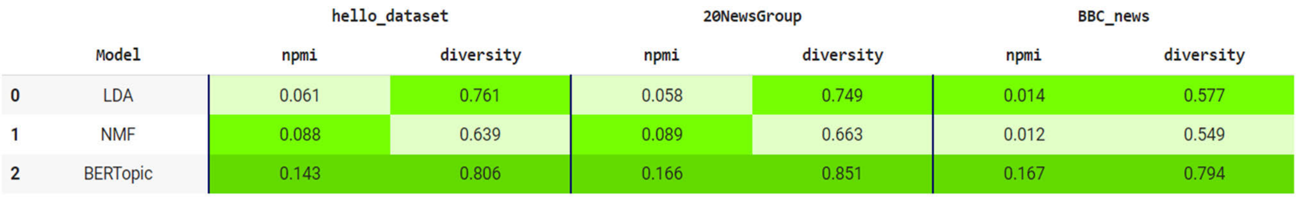
The NPMI metric measures how different in meaning the discovered topics are. A higher score in diversity reduces the redundancy among the topics. A low score suggests that the topics are repetitive and the model's ability to abstract the themes in the corpus is low [28].

It's worth noting that the number of topics chosen for the model can impact topic diversity. Choosing too many topics can lead to similar topics with overlapping top words while choosing too few can result in broad topics that are difficult to interpret. The most common measure for topic diversity is the ratio of unique words among the top 25 words [29].

IV. EXPERIMENTAL SETUP

For the experiment, we used the 20-group news and the BBC news datasets as benchmark datasets. Both these datasets

TABLE 1. Aggregate performance of topic models by dataset.



		hello_dataset		20NewsGroup		BBC_news	
	Model	npmi	diversity	npmi	diversity	npmi	diversity
0	LDA	0.061	0.761	0.058	0.749	0.014	0.577
1	NMF	0.088	0.639	0.089	0.663	0.012	0.549
2	BERTopic	0.143	0.806	0.166	0.851	0.167	0.794

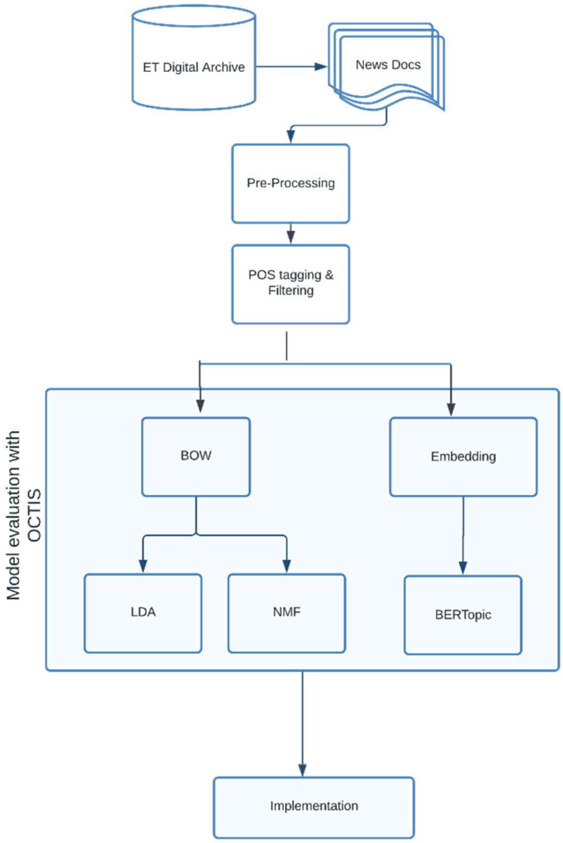


FIGURE 3. Experiment workflow.

were available in the preprocessed form in OCTIS — an open-source framework for training, evaluating, and comparing Topic models. The Economic Times corpus was our custom dataset. We used pretrained transformers for extracting sentence embeddings. Specifically, we used the 'all- mpnet-base-v2, which provides higher and consistent performance [20]. The generic workflow of the experiment is represented in Fig 3. For evaluating and configuring hyperparameters, we used OCTIS. OCTIS provides a unified platform with several topic models, datasets, and evaluation metrics, enabling easy comparison of topic model performance. It adopts a dataset-model-metric approach to provide optimal hyperparameters using a Bayesian Optimization strategy [21]. The OCTIS framework is represented in Fig 4. For the experiment we have chosen topic coherence and topic diversity measures as defined in the earlier section. Topic

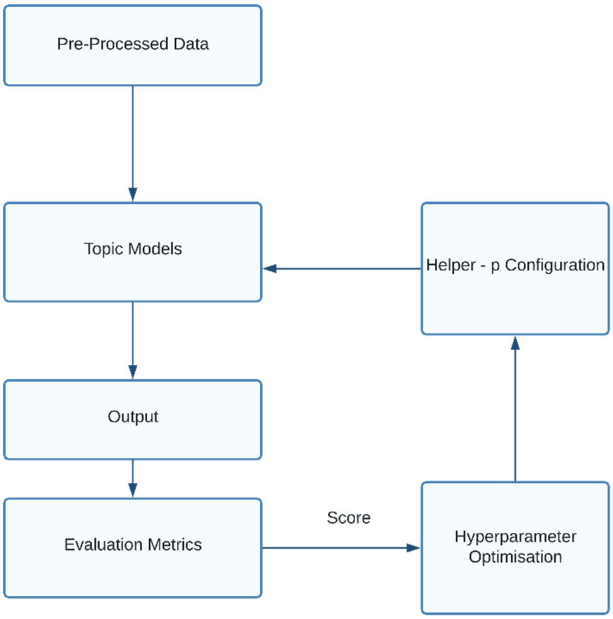


FIGURE 4. OCTIS workflow.

coherence metric is based on NPMI and topic diversity is based on unique words among top 10 words. The metrics are common for all the three models and is available as part of the OCTIS library.

A. DATA

We have used three datasets in the articles - The primary dataset actively curated by researchers and two other standard news datasets for benchmarking the performance of the various topic models.

1) THE ECONOMIC TIMES NEWS CORPUS

The experiment dataset for the article comes from the digital archives of the Economic Times, which is an established English business daily in India. We curated the dataset from news articles between 2008 and 2021. The news articles were extracted based on the keyword search 'climate change'. After removing shot news stubs and news items with videos, the final dataset had 9774 documents. The total number of terms in the corpus approximates 5 million.

2) BENCHMARK DATASETS

We used the 20 newsgroups and the B.B.C. news dataset as benchmark datasets for evaluating and comparing the

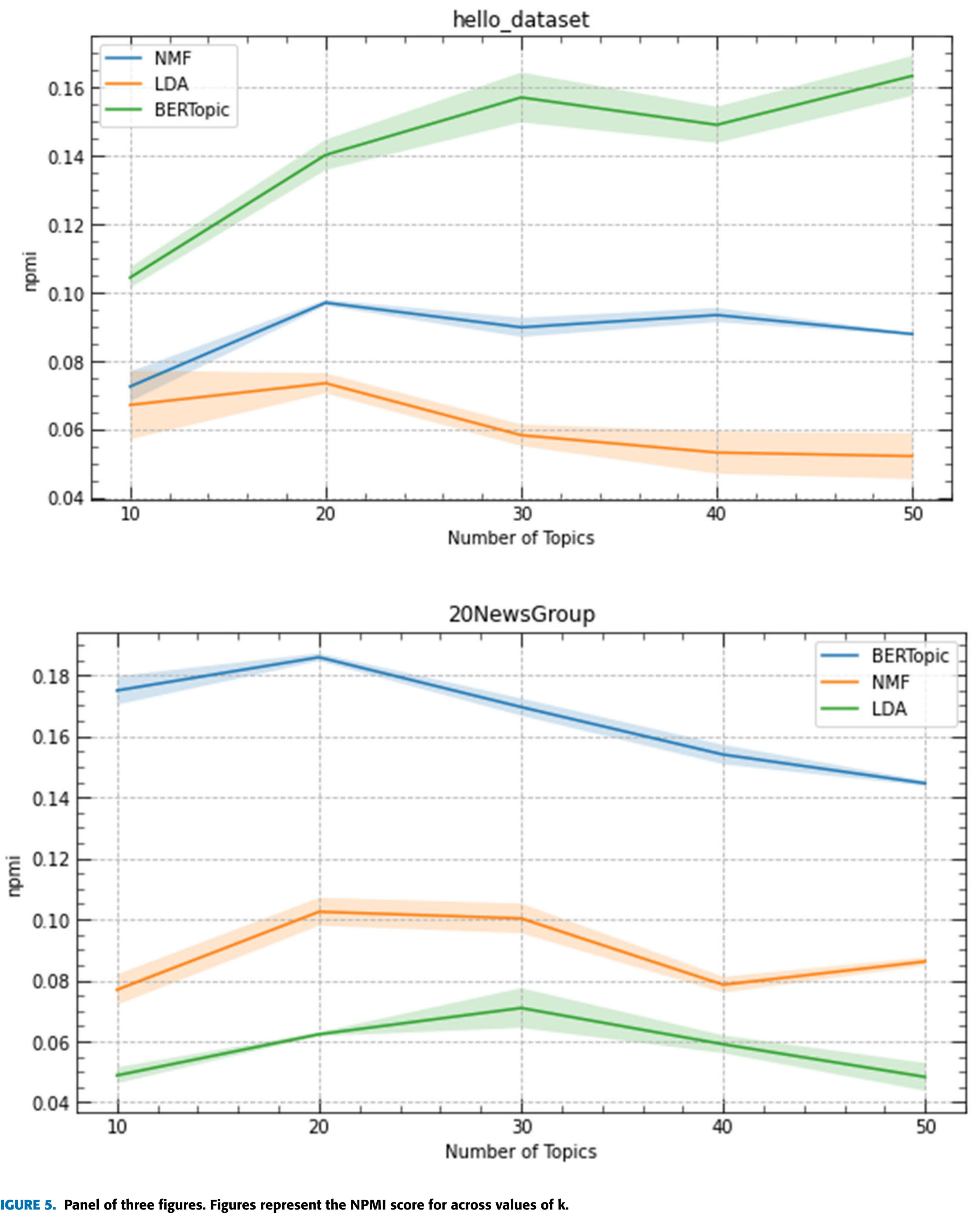


FIGURE 5. Panel of three figures. Figures represent the NPMI score for across values of k.

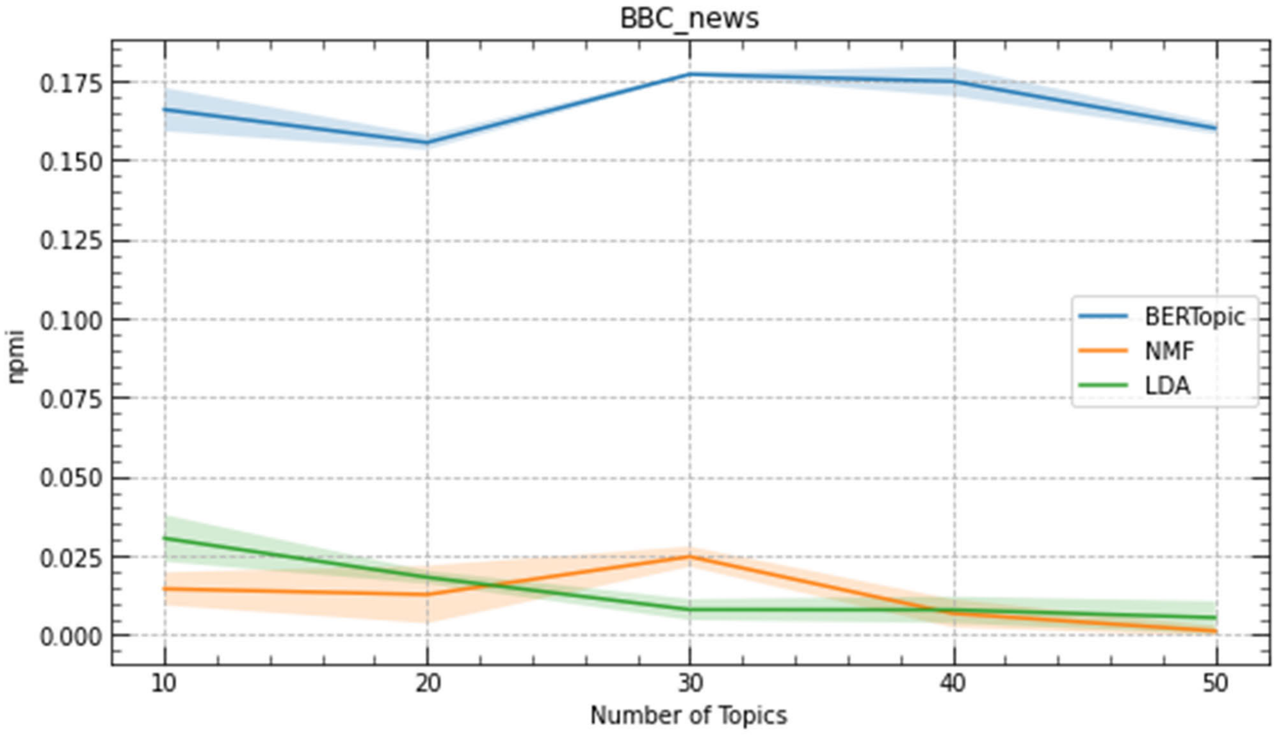


FIGURE 5. *(Continued.)* Panel of three figures. Figures represent the NPMI score for across values of k.

performance of topic modes. The 20 newsgroups dataset is popularly used in NLP experiments as a benchmarking dataset consisting of 20000 news articles equally distributed across 20 categories. The categories include technology, politics, religion, auto, sports, etc. Similarly, the B.B.C. news dataset has 2225 news articles reported between 2004-2005 across five topical areas. The categories include Business, Sports, Technology, Entertainment, and Politics.

3) RESULTS

Every topic model chosen for the experiment was run on each dataset three times. In each run was defined by the choice of a random seed. In every run we selected five OCTIS values for k in multiples of 10 until 50. In other words, each topic model was run 15 times on a dataset. The performance of the model can be analyzed in two ways. First, both NPMI. and topic diversity can be aggregated at a dataset level. We average the NPMI and topic diversity scores over the three iterations and summarize it at a dataset level as shown in Table 1. BERTopic performs better than LDA and NMF across all three datasets and on both the evaluation metrics. Between LDA and NMF, we also see that LDA performs better than NMF in terms of topic diversity, while NMF does better in terms of coherence. Second, we drill down into the performance of the three models at a dataset level where we calculate NPMI. with respect to **k**. Figure 5 is a panel of three graphs that reports the NPMI. performance. From the display, we observe that BERtopic outperforms the two other models across all three datasets. We find that NPMI. reached its maximum for different values of **k** for each of the dataset. In the case of the Economic Times dataset, we find that NPMI. value increases steadily till k = 30, declines and increases again until it reaches its high at k = 50. For the other two datasets, maximum NPMI. is achieved at k = 20 and k = 40, respectively.

Besides coherence, the experiment also yielded topic diversity scores for each model. The results are displayed in Figure 6. Once again, we observe that BERTopic performs better than the other models for the experiment and benchmark datasets. For the Economic Times corpus we see that topic diversity is maximized at k = 20 and remains stable till k = 50. For the benchmark datasets on the other hand, we see that topic diversity declines continuously for higher values of k. Based on the results for the Economic Times dataset, we set the k value at 50, where we achieved the highest NPMI. and highest topic diversity as well.

V. MAPPING THEMES USING BERTOPIC

Based on the model selection experiment in the last section, we chose BERTopic as the most suitable among all the three models. We then applied it to the E.T. corpus with k = 50. One of the advantages of BERTopic is that it generates topic labels along with the topics. The labels are generated as a combination of the topic number and dominant keywords. A partial list is provided below

- *1_water_said_pollution_air*
- *2_countries_climate_developing_agreement*
- *4_energy_climate_india_coal*
- *5_ice_climate_said_warming*

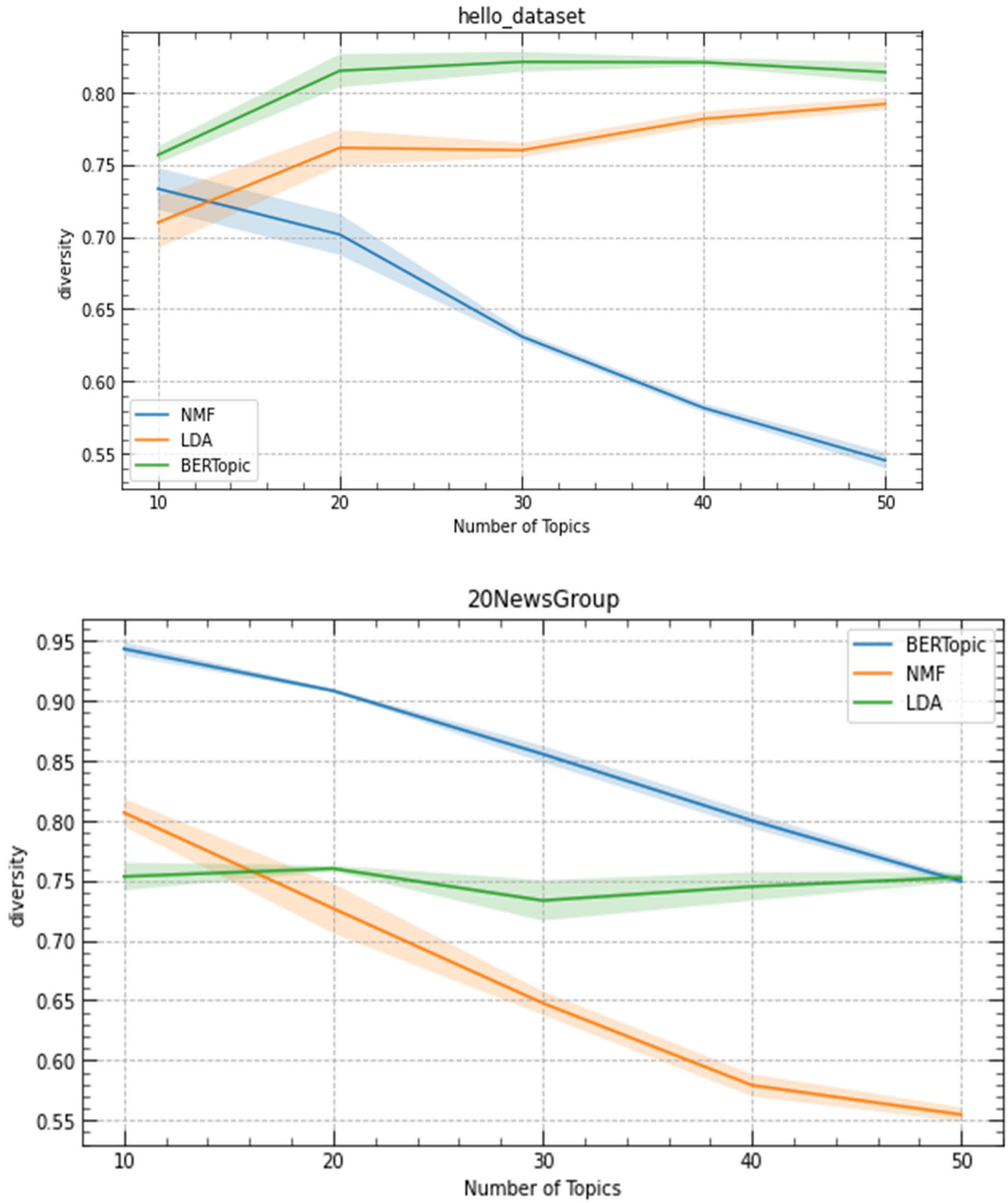


FIGURE 6. Panel of three figures. Figures represent the diversity score for across values of k.

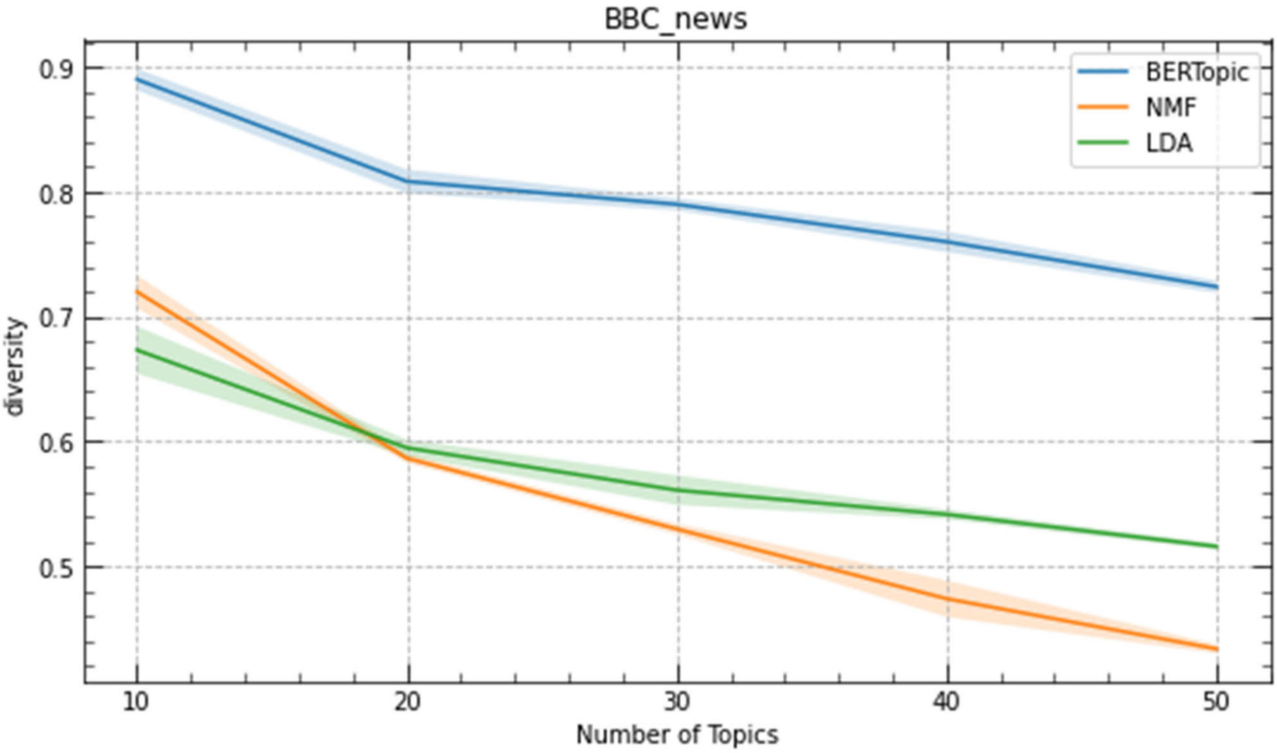


FIGURE 6. *(Continued.)* Panel of three figures. Figures represent the diversity score for across values of k.

- *6_climate_emissions_countries_change'*
- *8_energy_solar_renewable_power*
- *12_monsoon_rainfall_climate_said*
- *15_climate_trump_paris_said*
- *17_species_tiger_conservation_wildlife*
- *39_energy_india_renewable_clean*
- *47_adani_mine_queensland_coal*
- *49_Yoga_Resolution_day*

While topics might be coherent semantically, not all of them may be directly relevant to climate change. For example, in topic 49 in the above-covered news about yoga day events, speakers also mentioned climate change. Alternatively, topics can be merged to represent a larger theme, such as in the case of topics 2 & 6. Both topics relate to climate change negotiations with obvious differences in their point of view. Topic 2 covers news articles with a 'developing country' perspective, while topic six focuses on news articles which discuss country-specific emissions within the summit context.

To further identify opportunities for merging similar topics we performed the following steps. First we examined keywords within each topic to identify topics that can be merged or need to be discarded. Second, we manually sampled the full articles within each topic to infer the context. Finally, we investigated the relationship between topics through clustering based on the cosine distance between topic embeddings. The extracted themes, and their relationship is presented in Fig 7. At the end of the process we were able to identify topics that were peripheral to climate change and discarded them. The other topics, we grouped them into eight overarching themes. The themes were 'Climate cooperation' and 'Climate agreements & 'Energy & Emissions', 'clean energy', Resource management, Country emissions and Business engagement 'country emissions.' 'Climate cooperation' combines topics 2,13,21,27,35,42, 43,46, represents all news articles on bilateral and multilateral visits, trade relationships, and regional summits (ex: SAARC, Quad), etc., where climate change is on the agenda. 'Resource management and ecosystems' consists of news articles about natural resources management, including air, land, water, wildlife, food, etc. News stories often discuss the impact of climate change on natural resources. The theme groups topics 4,9,17,33 and ranks second in terms of overall contribution to the corpus. 'Climate agreements & negotiations' contains news articles that discuss COP summits, primarily including Paris, Glasgow, Copenhagen, and Lima. Following this we have two themes on energy' Energy & Emissions' and 'clean energy. While they cover energy at large, the key distinctions come from the focus of the news stories. The 'Energy & emissions' topic consists of articles that report the role of energy in emissions and emission reduction. On the other hand 'clean energy' exclusively focuses on domestic action in improving the share of renewables in electricity regeneration. On a similar note, we find the

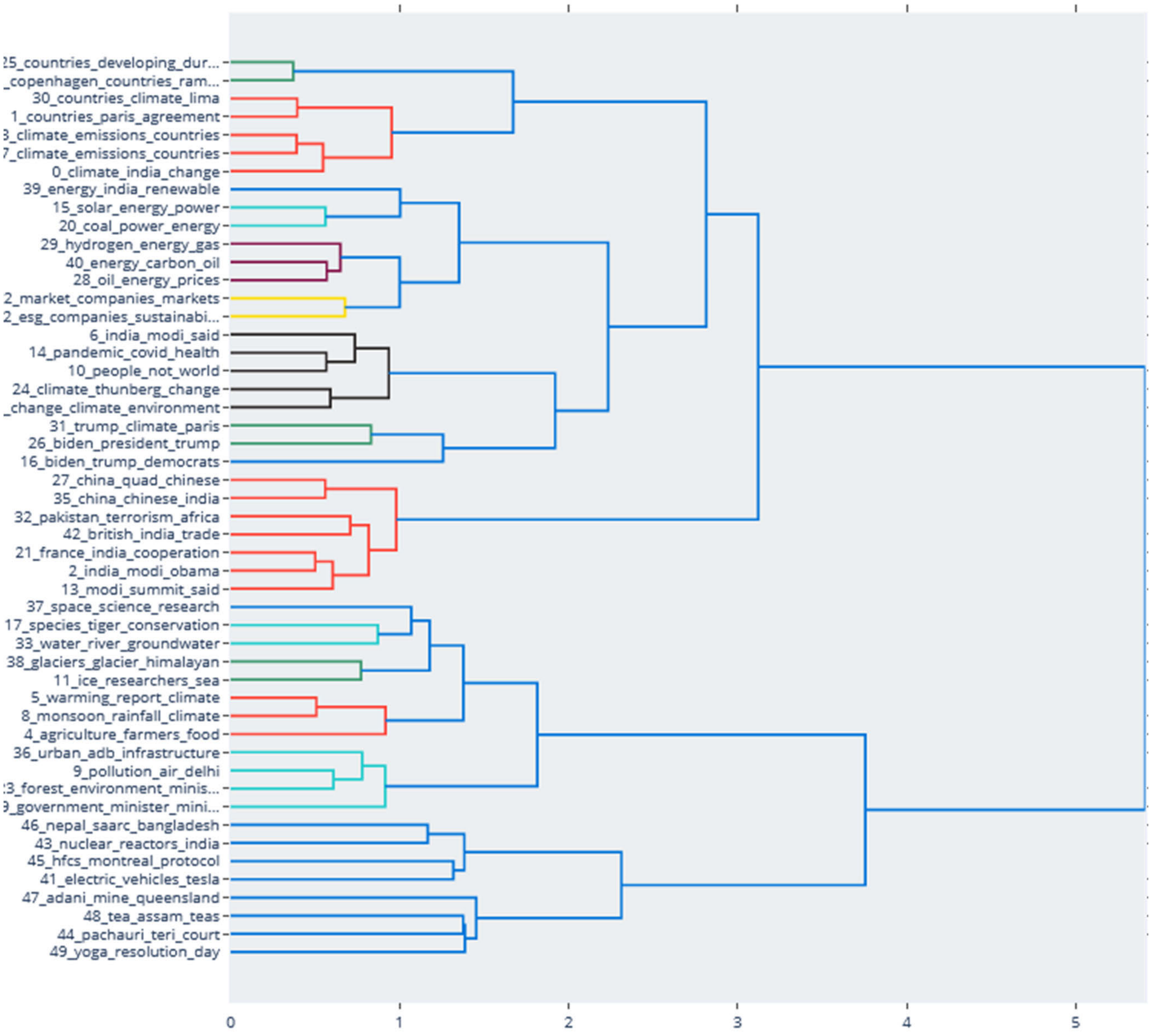


FIGURE 7. Hierarchial topic cluster of the corpus.

'country emissions' focuses on new stories that present national emission targets and goals, and performance of different countries. Specifically, several stories discuss India and China and their role as prominent contributors in terms of absolute emissions.

Finally, we have the "Business and climate" themes, which subsume three topic categories- 12,22 and 49. The news stories are a collection of company agendas for reducing climate footprints, improving sustainability practices, CEO. speeches on climate change, impact on the stock, product markets, and new emerging industries such as Electric Vehicles & Batteries. We further track the aggregated themes over 2008 -2020, as shown in Fig 8. It is interesting to note that news across all categories spiked sharply in 2015 and 2021. These spikes may be attributed to the greater attention given to climate change around the Paris COP and the Glasgow COP.

There are also interesting insights that emerge from how different themes trend. Before 2014 news on climate summits and negotiations dominated other categories. Post 2014, we see that news on climate cooperation has the highest contribution to the overall corpus. In tandem with climate cooperation, we also observe the increase in reporting on the impact of climate change on resources and ecosystems. We also observe that both the categories of energy news (emissions & clean energy) move together. Finally we, see that news around business themes has gained traction since 2012 but has increased significantly from 2017 onwards.

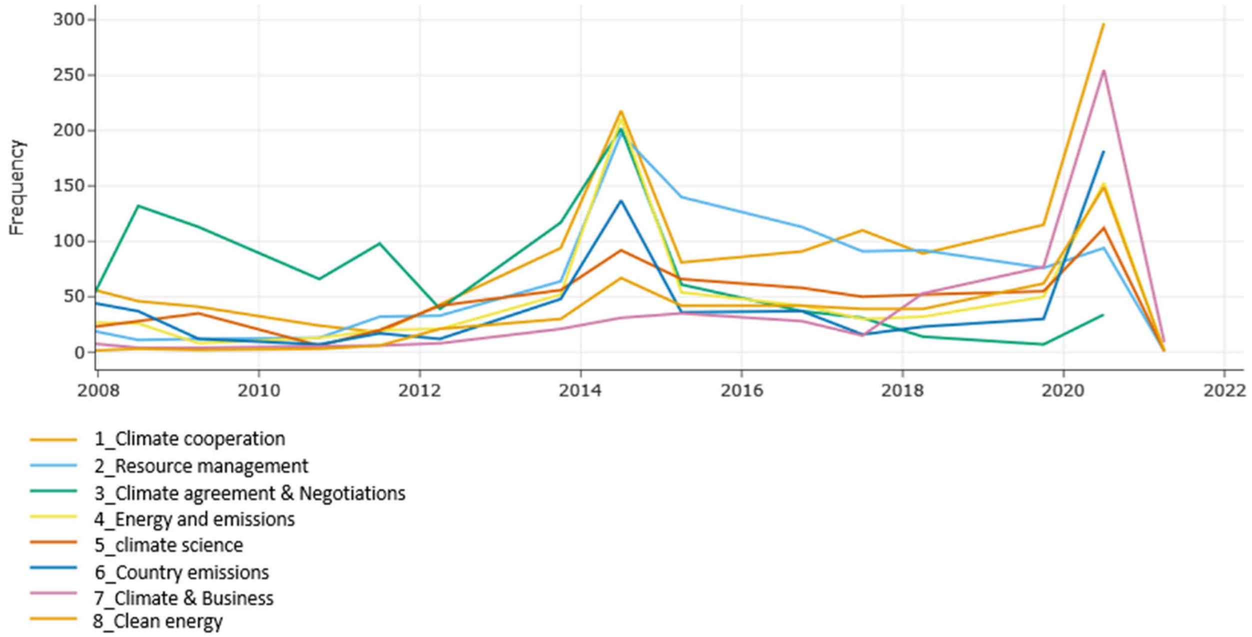


FIGURE 8. Trends in climate change news within themes.

VI. DISCUSSION AND CONCLUSION

The scientific consensus on climate change is unassailable and there is an urgent need for stronger climate action globally. However, countries continue to grapple with socio-political challenges which have emerged as a deterrent to effective climate action. The social and political discourse around climate change involves the public, the state, and the scientific community, with media as the common platform. Understanding how media reports on climate change gives a glimpse into this diverse stakeholder system's perspectives, debates, and responses.

However, the tools used to understand the climate change discourse have been derived mainly from traditional quantitative and qualitative research. Natural Language based techniques can augment the existing toolbox to unfold and extract insights from the climate change debate. This study is a step towards achieving integration of NLP in social science research and, more specifically, within the domain of climate change discourse [17], [30]. Another important contribution of the study is the demonstration and use of embeddings based topic models BERTopic as an alternative to LDA and NMF. Additionally, it also furthers the use of experimental frameworks such as OCTIS for evaluating the performance of topic models. Finally, through the longitudinal analysis of climate news the study provides a developing country perspective to the literature on climate change discourse

The analysis in our study includes two components. First an experimental study on topic models to determine relative performance.Second is the actual mapping of the climate news based on the model selected in the experiment Our experiments show that embedding models perform significantly better than the other topic models. The results are consistent with other recent comparative studies [31], [32]. We further demonstrate the application of BERTopic by training it on the Economic Times corpus and discovering news frames and their evolution. Results suggest that news frames around climate negotiation have been dominant before 2014 while climate cooperation gains importance post 2014. We also found that climate frames related to domestic action also has gained traction in recent years.

Stakeholder consensus is an essential aspect of climate action. Given the multiple stakeholders in climate change discourse, media analysis can play a crucial role in gauging the engagement of citizens, the state, and the commercial sector. Using NLP techniques in this context can scale up information processing and augment insights derived from other research methods, thus helping make informed policy decisions for countering climate change.

REFERENCES

[1] D. Eckstein, *Global Climate Risk Index 2020*. Bonn, Germany: Germanwatch, 2019.

[2] T. R. Keller, V. Hase, J. Thaker, D. Mahl, and M. S. Schäfer, "News media coverage of climate change in India 1997–2016: Using automated content analysis to assess themes and topics," *Environ. Commun.*, vol. 14, no. 2, pp. 219–235, Feb. 2020.

[3] J. Bohr, "Reporting on climate change: A computational analysis of U.S. newspapers and sources of bias, 1997–2017," *Global Environ. Change*, vol. 61, Mar. 2020, Art. no. 102038.

[4] L. L. Benites-Lazaro, L. Giatti, and A. Giarolla, "Topic modeling method for analyzing social actor discourses on climate change, energy and food security," *Energy Res. Social Sci.*, vol. 45, pp. 318–330, Nov. 2018.

[5] S. Umamaheswaran, V. Dar, and J. Thaker, "The evolution of climate change reporting in business media: Longitudinal analysis of a business newspaper," *Sustainability*, vol. 14, no. 22, Nov. 2022, Art. no. 15214.

[6] J. Foulds, "Mixed membership word embeddings for computational social science," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 86–95.

[7] D. Hovy and S. L. Spruit, "The social impact of natural language processing," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2016, pp. 591–598.

[8] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, ''SemEval-2016 task 6: Detecting stance in tweets,'' in *Proc. 10th Int. Workshop Semantic Eval. (SemEval-)*, 2016, pp. 31–41.

[9] K. Johnson, I.-T. Lee, and D. Goldwasser, ''Ideological phrase indicators for classification of political discourse framing on Twitter,'' in *Proc. 2nd Workshop NLP Comput. Social Sci.*, 2017, pp. 90–99.

[10] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, ''Sentiment, emotion, purpose, and style in electoral tweets,'' *Inf. Process. Manage.*, vol. 51, no. 4, pp. 480–499, 2015.

[11] R. Kashyap and A. Nahapetian, ''Tweet analysis for user health monitoring,'' in *Proc. 4th Int. Conf. Wireless Mobile Commun. Healthcare Transforming Healthcare Through Innov. Mobile Wireless Technolog.*, Nov. 2014, pp. 348–351.

[12] P. Lavanya and E. Sasikala, ''Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: A comprehensive survey,'' in *Proc. 3rd Int. Conf. Signal Process. Commun. (ICPSC)*, May 2021, pp. 603–609.

[13] A. Farzindar and D. Inkpen, ''Natural language processing for social media,'' *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 2, pp. 1–166, 2015.

[14] L. Hagen, O. Uzuner, C. Kotfila, T. M. Harrison, and D. Lamanna, ''Understanding Citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach,'' in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, Jan. 2015, pp. 2134–2143.

[15] D. Elgesem and L. N. Steskal Diakopoulos, ''Structure and content of the discourse on climate change in the blogosphere: The big picture,'' in *Climate Change Communication and the Internet*. Evanston, IL, USA: Routledge, 2019, pp. 21–40.

[16] H. Elfardy and M. Diab, ''CU-GWU perspective at SemEval-2016 task 6: Ideological stance detection in informal text,'' in *Proc. 10th Int. Workshop Semantic Eval. (SemEval-)*, 2016, pp. 434–439.

[17] P. Swarnakar and A. Modi, ''NLP for climate policy: Creating a knowledge platform for holistic and effective climate action,'' 2021, *arXiv:2105.05621*.

[18] B. Nerlich, R. Forsyth, and D. Clarke, ''Climate in the news: How differences in media discourse between the U.S. and U.K. reflect national priorities,'' *Environ. Commun.*, vol. 6, no. 1, pp. 44–63, Mar. 2012.

[19] P. DiMaggio, M. Nag, and D. Blei, ''Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding,'' *Poetics*, vol. 41, no. 6, pp. 570–606, Dec. 2013.

[20] Y. Bengio, R. Ducharme, and P. Vincent, ''A neural probabilistic language model,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1137–1155.

[21] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, ''Topic modeling in embedding spaces,'' *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 439–453, Dec. 2020.

[22] P. Ghasiya and K. Okamura, ''Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach,'' *IEEE Access*, vol. 9, pp. 36645–36656, 2021.

[23] J.-R. Lin, Z.-Z. Hu, J.-L. Li, and L.-M. Chen, ''Understanding on-site inspection of construction projects based on keyword extraction and topic modeling,'' *IEEE Access*, vol. 8, pp. 198503–198517, 2020.

[24] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, ''Experimental explorations on short text topic mining between LDA and NMF based schemes,'' *Knowl.-Based Syst.*, vol. 163, pp. 1–13, Jan. 2019.

[25] D. Lee and H. S. Seung, ''Algorithms for non-negative matrix factorization,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1–7.

[26] M. Röder, A. Both, and A. Hinneburg, ''Exploring the space of topic coherence measures,'' in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408.

[27] D. Mimno, ''Optimizing semantic coherence in topic models,'' in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 1–11.

[28] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, ''Topic modeling algorithms and applications: A survey,'' *Inf. Syst.*, vol. 112, Feb. 2023, Art. no. 102131.

[29] S. Limwattana and S. Prom-On, ''Topic modeling enhancement using word embeddings,'' in *Proc. 18th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jun. 2021, pp. 1–5.

[30] M. Stede and R. Patz, ''The climate change debate and natural language processing,'' in *Proc. 1st Workshop NLP Positive Impact*, Aug. 2021, pp. 8–18.

[31] M. J. Sánchez-Franco and M. Rey-Moreno, ''Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings,'' *Psychol. Marketing*, vol. 39, no. 2, pp. 441–459, 2022.

[32] R. Egger and J. Yu, ''A topic modeling comparison between LDA, NMF, Top2 Vec, and BERTopic to demystify Twitter posts,'' *Frontiers Sociol.*, vol. 7, May 2022, Art. no. 886498.

SWARNALAKSHMI UMAMAHESWARAN received the Ph.D. degree in public policy from the TERI School of Advanced Studies, New Delhi. Her doctoral work was on policy analysis for low-carbon growth development in the Indian context. Prior to her doctoral studies, she was an analytics consultant across several MNCs in the banking and marketing sector. As part of her corporate career, she has built credit scoring and risk-profiling models using machine learning and data science tools. She is an economist by education and a data scientist by practice. She is currently a Business Analytics Faculty Member with the Symbiosis Institute of Business Management, Bengaluru. Her current research interest includes computational social sciences, with a special focus on climate and sustainability.

VANDITA DAR received the Ph.D. degree in economics from Mumbai University. She is a passionate economist and a researcher working as a Faculty Member with the Symbiosis Institute of Business Management, Bengaluru. She has diverse work experience in academics, research, and corporate for more than 20 years. In her last corporate assignment, she held the profile of an economist with a global financial services company. She has conducted capacity-building workshops for senior government officers spanning the areas of macroeconomic policy and business environment, and was the head of the training division with an apex state training institute. Her research interests include sustainability, climate issues, macroeconomic policy, and development economics. She is a reviewer of reputed international journals.

ELIZA SHARMA received the Ph.D. degree in management and finance from the Jaypee Institute of Information Technology, Noida. She was with the Indian Institute of Management, Ahmedabad, India, as a Research Assistant for the Government of India project to develop the Integrity Index for public sector organizations. She is currently a Finance Faculty Member with the Symbiosis Institute of Business Management, Bengaluru. She has published many research papers in national and international journals and conference proceedings. Her research interest includes exploring corporate social responsibility practices in the industry.

JIKKU SUSAN KURIAN received the Ph.D. degree in the core area of organizational behavior. She has 17 years of experience, of which eight years were with leading corporates and the rest of seven years with KL University Business School, Vijayawada. As a faculty member, she handles organizational behavior and HR-related subjects. She is currently an Assistant Professor with the Symbiosis Institute of Business Management, Bengaluru (SIBM-B). Her current research interests include researching sustainable HRM and technological transitions.