TEXT ANALYSIS ON REVERSE MIGRATION IN INDIA DUE TO COVID-19

Sonali Subbu Rathinam

December 2022

Executive Summary

In order to contain the spread of Covid-19, the Indian government imposed its first country-wide lockdown in March 2020, and it continued till the end of May 2020. The immediate job loss due to the lockdown, and fear and anxiety around the pandemic led to many urban migrant workers returning back to their homes in rural areas. There was significant media coverage on this issue, and it highlighted the plight and miseries that migrant workers endured on a daily basis, and in their travel back home. Therefore, this project aims at inferring sentiments and topics from news articles on this issue by replicating the results produced by Agarwal and Sarkar (2022). The overall neutral sentiment result produced using VADER and TextBlob lexicons are the same as that of the authors. The results from Topic Modelling using LDA are similar to the results produced by the authors, and highlighted transport (particularly railways), food and employment schemes as the main topics. Additionally, topic modelling replication results highlighted the issue of migrant fishermen, particularly from Jharkhand state, being stranded during the first lockdown.

Introduction

Almost all countries and territories in the world instituted lockdowns with varying restrictions across different timelines as one of the primary measures to curb the spread of Covid-19. While this has mostly been effective in reducing the number of covid positive cases in a region, it has had unprecedented social and economic impacts, especially in developing countries. In India, the first lockdown in March 2020 was imposed within 14 hours of its announcement by the national government. Since this led to the immediate closure of factories and markets, millions of Indians in the informal sector lost their jobs, and found themselves struggling with basic necessitates such as food and shelter. This situation was particularly precarious for migrant workers in urban cities, as they were solely dependent on their daily wages. (Chandrashekar, 2020) The lack of alternate employment in urban regions, uncertain futures, and fear around the new health crisis left millions of migrant workers with no other option other than returning to their rural homes. The results of the Periodic Labor Force Survey (2020-2021) by the National Statistical Office revealed that this lockdown forced about 51.6% of men in urban cities in India to move to rural homes (Mishra, 2022). This was a massive non-clinical crisis due to Covid-19, and the biggest domestic migration the country has ever witnessed since its Independence in 1947 (Agarwal & Sarkar, 2022).

Internal migration in India stems from the social and economic inequalities that exist due to caste, gender, tribe, religion and regional divisions. In order to escape these inequalities, migrant workers move to urban areas for informal work that include domestic help, and work in construction sites and factories. According to the 2017 – 2018 Labor Force Survey, there were 28 million rural-to-urban workers in the unorganized informal sector (Chen et al., 2020). The immediate lockdown that strictly restricted movement had particularly adverse effects to this already vulnerable subset of population. Since most of the public transport services were suspended, millions of migrant workers travelled thousands of kilometers on foot and/or on bicycles. Traditional news media and social media were inundated with news and pictures that showed all these issues they faced during their travel, and highlighted that there weren't enough policies and measures to provide economic and social protection to migrant workers. Therefore, Agarwal and Sarkar perform sentiment analysis and topic

modelling of popular English news sources in India on the covid induced reverse migration to understand the issues that migrant workers face in the country, and identify the push/pull factors that resulted in this movement. They executed sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon at the document (news article text) level. Topic Modelling with LDA was performed using the Gensim Python library to construct the dictionary corpus and convert it to Bag-of-words matrix. They also performed agglomerative hierarchical clustering of news articles to compare with the results produced by topic modelling. The replication project, however, focuses on replicating the results produced by sentiment analysis and Topic Modelling with LDA. Since the authors do not perform additional tests to confirm their results (except for topic modelling with clustering), this replication project also conducted sentiment analysis using TextBlob and Topic Modelling with LDA from the sklearn package to create a document-topic matrix using TF-IDF (Term Frequency – Inverse Document Frequency). These additional steps were performed to verify the results obtained from sentiment analysis and topic modelling.

The findings of Agarwal and Sarkar and of this report are particularly significant as there is a dearth of quantitative research on reverse migration in India. Most of the existing research is based on qualitative research, such as Khan & Arokkiaraj's (2021) comparative study based on telephonic survey data from 65 migrants that returned from urban regions, and John and Kapilashrami's(2021) qualitative text analysis of news articles to analyse migrants and refugee health portrayal in Indian media. Data driven research such as Behera et al's (2021) multivariate regression approach to study the effects of reverse migration on labour supply in Odisha doesn't address endogeneity concerns, and heavily relies on domain knowledge of the authors.

Data

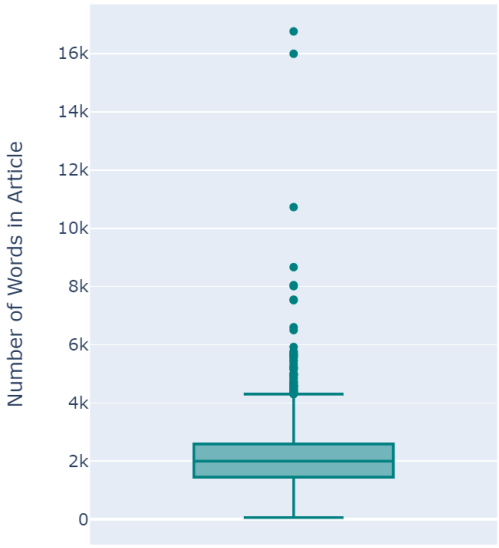
Since this topic aims at inferring sentiments and topics from news articles, text data pertaining to the covid induced reverse migration with each news article as the unit of analysis would be ideal. Sarkar and Agarwal manually collected news articles' links from two mainstream print media in India, The Times of India and The Hindu. The authors chose articles from The Times of India and The Hindu as they are widely popular news sources, and results from Statista also corroborate the same (Basuroy, 2021). 2170 articles from these two sources that contained the terms 'migrants', 'migration', 'lockdown', 'COVID-19', and 'pandemic' between May and June 2020 were collected. The final dataset that the authors used included the news title, the author, the text of the article, the summary, and keywords found from the articles. Since this dataset was publicly available, the same was used for this replication project (Agarwal and Sarkar, 2022).

On exploratory analysis, it is found that the article text variable is the most important of all, as it is the most informative variable in the corpus. Table 1 provides the summary statistics of the variable, and from this, it can be inferred that most articles have about 2100 words, but they also reveal that there are outliers as well with the least and maximum length. This interpretation is validated by Figure 1, which shows that there are a few articles that have a substantially higher number of words as compared to the other articles. However, they were not removed from the analysis as there is a possibility that the longer news articles have a more detailed information about reverse migration and its associated topics, which may provide more insight during text analysis.

Table 1: Summary Statistics of Article Text Length

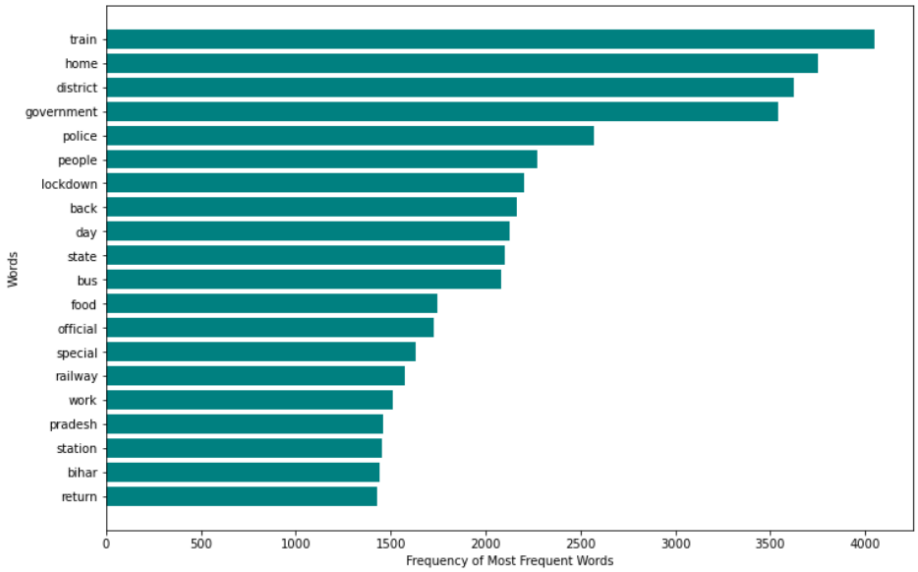
Summary Statistic of Article Text	Value
Count	2170
Mean	2110.04
Standard Deviation	1071.54
Minimum	67
25 %	1449
50 %	1998
75 %	2591
Maximum	16756

Figure 1: Number of Words in Each Article



Since topic modelling and sentiment analysis is performed on the article text variable, Figure 2 was produced to analyse the most frequently occurring words in the variable. It points out that the most frequently occurring words are associated with transportation, the different administrative regions in the country, and authorities that maintain law and order.

Figure 2: Most Frequent Words in Article Text



Although there is no missingness in the dataset, it is not without limitations. The first one is that these are from English news sources, and using news articles that are circulated in regional/local languages may produce different sentiments and topics on analysis. There is a high probability that they include those migration issues that are of more importance to the states/areas where their language is widely-spoken, and include important regional information. However, it will be difficult to perform text analysis on these languages due to translation issues (Khan Jadoon et al., 2017).

Secondly, the dataset only includes articles from the first lockdown in India. India had another lockdown in 2021 due to the increased number of covid cases due to the Delta variant, and many states imposed other separate lockdowns based on the rise of cases in the region. Hence, the scope of this analysis is limited to the migrants who immediately returned to rural regions, and not those who moved back in the next few phases of the pandemic.

Methodology

Sentiment Analysis with VADER and TextBlob

Sentiment Analysis has been performed on the data to identify the overall sentiment in articles related to reverse migration. Sentiments evoked from reading these articles can have significant impact on the economic lives of the migrants, especially because various manufacturing firms and state governments introduced various interventions and rewards to migrants to alleviate their plight, based on the extensive news coverage of the situation during the first lockdown (Pundir, 2020). Sentiment Analysis using VADER and TextBlob are implemented for this replication project. Both these methods use a lexical approach, which implies that it maps words and vocabularies to emotion intensities and previously predetermined scores as positive or negative. These scores are based on pre-trained models by humans. Particularly, VADER outputs a dictionary that informs how positive, negative and neutral a sentence is by indicating the probability for each sentiment classification. Additionally, a compound score, which is calculated by normalizing the aforementioned scores, is also returned (Beri, 2020). TextBlob, on the other hand, returns a polarity and subjectivity score. The Polarity score ranges from (-1, 1) and identifies the most negative scores close to -1 and vice-versa. The subjectivity score ranges from (0, 1) and indicates sentences with higher amount of personal opinion to 1 and vice-versa (Shah, 2020) . These two methods are quite similar as they are both use lexicon-based approach, but one of its main differences is that VADER identifies social media content such as emojis, repititions and punctuations better than TextBlob (Amy, 2022).

Lexical based approaches to sentiment analysis are advantageous as it is easy to implement, and does not require any training data. However, one of its main drawbacks is that it can cannot capture misspellings and grammatical mistakes, even of words with pre-determined scores. Another main drawback is that it cannot recognize sarcasm and irony in sentences well (DeLancey, 2020)

Heiden & Parpinelli performed sentiment analysis using VADER for stock price prediction from newspapers, and their analysis revealed that VADER doesn't need a high volume of data to achieve high accuracy. This bolsters the choice of using VADER for this project, as this corpus only analyses 2170 news articles (Heiden & Parpinelli, 2021). Narang used twitter sentiment analysis towards the Citizenship Amendment Act in India, another issue in India which the government struggled to find immediate solutions to, much like the reverse migration issue. They showed that VADER produced faster and more accurate results as well. (Narang, 2020). Mujahid et al. performed sentiment analysis using TextBlob and other sentiment analysis packages on tweets about online education during Covid, and found that TextBlob showed more accurate results for data annotation when compared to others (Mujahid et al, 2021). Oyebode and Orji used both VADER and TextBlob to identify sentiments from social media posts regarding the 2019 Nigerian Presidential Elections. Their results showed that VADER outperformed all other sentiment analysis models, and attributed this to the fact that VADER particularly focuses on social media texts (Oyebode et al. 2019).

Topic Modelling with LDA using TF-IDF and Bag-of-Words

Topic Modelling helps in uncovering topics that are inherently present in the text data, which are not easily identifiable and understandable through manual inspection of newspaper articles. This unsupervised learning technique detects patterns of tokens and phrases from natural language text and automatically groups them to generate relevant topics. Topic modelling is used instead of topic classification since the data is unlabelled. For the replication project Topic Modelling with LDA is first implemented using LDA in sklearn with TF-IDF, followed by using the Gensim package with Bag-of-Words model.

LDA generates topics based on conditional probability estimates. It is based on the assumption that each document in the text corpus belongs to a mix of topics, and each topic is a mix of words. While using LDA with sklearn, the TF-IDF is used to transform the Document Term Matric to numerical arrays. With the genism package in Python, the Document Term Matrix does not have to be explicitly created. However, it requires that the text corpus to be tokenized.

One of the main advantages of LDA is that it can generalize the model to documents outside the corpora as well (2016) However, one of the main drawbacks is also that a fixed 'number of topics' value must be passed to the LDA model.

Liu et al used topic modelling with LDA on news sources to identify important topics in articles associated with third-hand smoke in China. Their results using LDA revealed that the news sources would have to focus on scientific evidence more, and provide lesser focus on sensational headlines (Liu et al, 2019).Ylä-Anttila et al used topic modelling with LDA to study media debates on climate change in India and USA. Their research highlighted an important advantage of LDA, which is that it does not replace qualitative interpretation, but complements it by enabling a degree of automated discovery, which may highlight unexpected patterns (pg 108, Ylä-Anttila et al., 2021)

Findings

Sentiment Analysis:

This replication project implements sentiment analysis using VADER and TextBlob lexicon. Results from VADER indicated that the most negative probabilities were assigned to news articles that reported on deaths due to covid and mob lynching, accidents, and railway officials and police brutality towards migrants during their travel. The most positive probabilities were assigned to news articles that reported movie celebrities donating money to organizations that helped support migrant workers, announcements about policy interventions taken to manage the logistics associates with migrant travel, and festival celebrations. Based on the compound score, each article was assigned an overall sentiment classification. If the compound score was higher than 0.05, it was labelled positive, if the score was between 0.05 and -0.05, it was labelled neutral, and all others were labelled as negative. Figure 3 provides information on the distribution of sentiment classification according to the aforementioned conditions. From the figure, it can be inferred that most articles were labelled as positive. To confirm these results, sentiment analysis with TextBlob was then implemented. TextBlob scores that were greater than 0 were labelled as positive, 0 as neutral, and all others negative. Figure 4 informs us that the sentiment classification with VADER and TextBlob are quite similar, indicating that TextBlob also assigned a large number of articles as positive. This

result is a clear contradiction to the existing domain knowledge on this issue, as these are primarily news articles that reported on the miseries of migrant workers. Hence, the probability scores for each sentiment classification for all articles was visualized in Figure 5. This clearly shows us that overall, the neutral probability is much higher than that of both the positive and negative probabilities. This leads to the conclusion that most of the words in articles do not have very high positive or negative scores in the VADER lexicon. Figure 6 also confirms the previous result as most articles with TextBlob have their scores around -2.5 and 2.5, again indicating that there aren't very high positive or negative sentiment articles. These overall neutral results obtained are consistent with the results that Agarwal and Sarkar obtained for sentiment analysis as well. This implementation points out the importance of checking and visualization the individual positive and negative sentiment scores for sentiment analysis, instead of solely relying on sentiment classification to form final conclusions. From the individual sentiment scores and classifications from VADER and TextBlob, it can be inferred that both the news sources present in the data reported this issue in a neutral manner and without any party inclinations, as that would have shown higher probabilities for either positive or negative sentiments. It also leads to the conclusion that pictures of the migrants' plights and their travels to rural regions may have played a more significant role than traditional news articles in invoking the overall negative sentiment that the general Indian public for this particular issue. Another important inference to consider is that the negative sentiment from migrants' struggles and the positive sentiments from the various donations and policy interventions may have cancelled each other and contributed to the overall neutral sentiment.

Figure 3: Distribution of Sentiment Classification with VADER

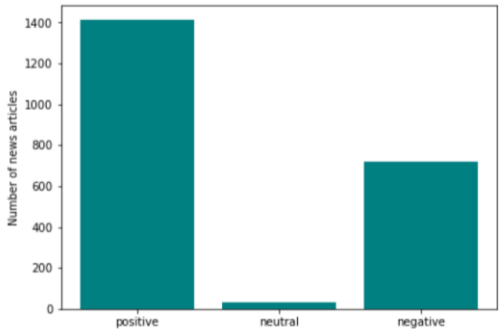


Figure 4: Sentiment Classfication Matching across VADER and TextBlob

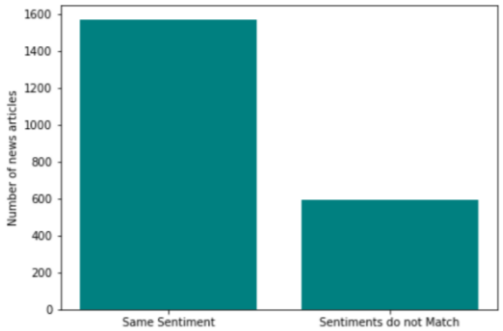


Figure 5: Sentiment Probabilites with VADER

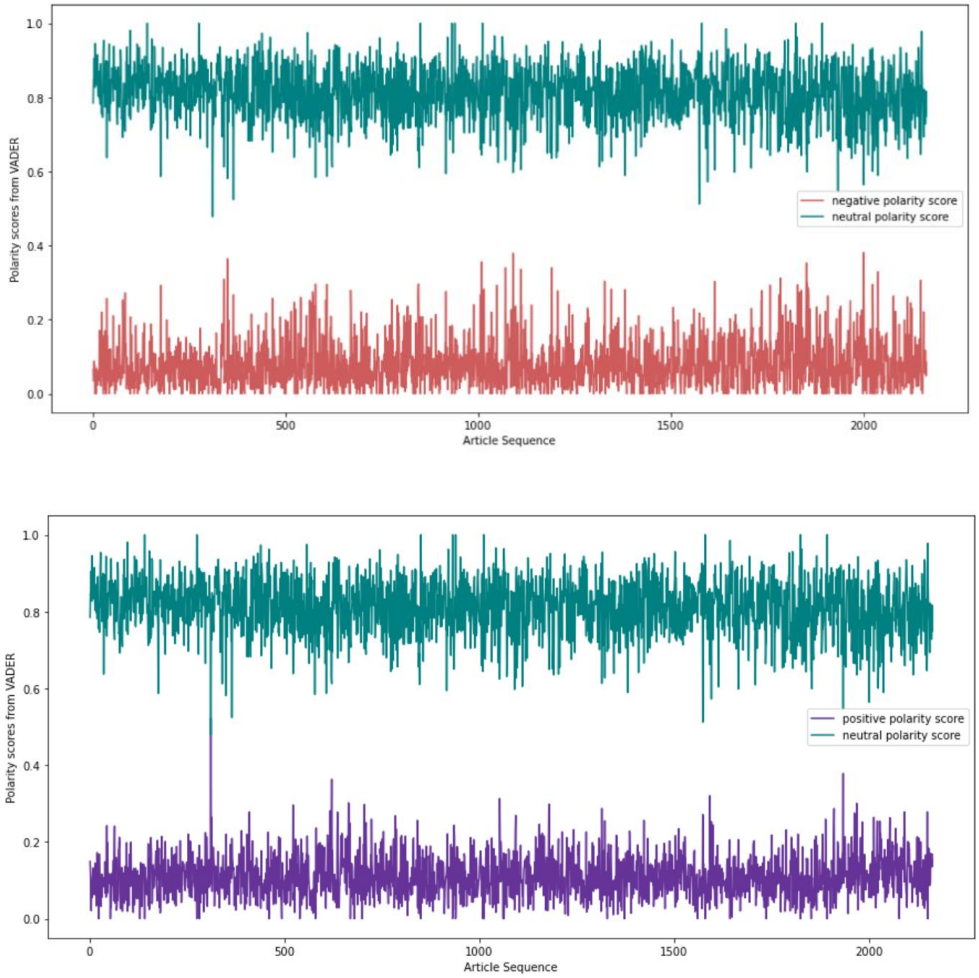
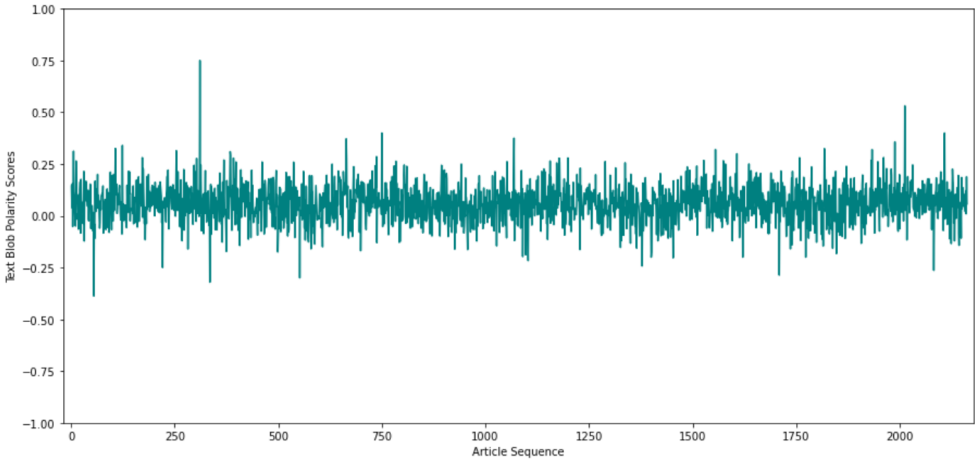


Figure 6: Sentiment Probabilites with TextBlob



Topic Modelling with LDA:

In this replication project, topic modelling with LDA from both sklearn and Gensim packages have been implemented. For both the LDA models, we use 5 as the number of topics as it produced the best topic modelling results when compared to other values (in the (2, 10) range) for this parameter. In order to implement LDA from sklearn, Figure 7 was produced to identify the ideal minimum document frequency to be used with TF-IDF. The figure shows that 5 would be an appropriate minimum document frequency, hence it was used in TD-IDF to convert the document term matrix to numerical arrays for LDA. Table 2 shows the results of topic modelling with LDA from sklearn. From the Table, it can be inferred that topics have overlapping words (such as government and killing), and not all provide clear interpretation. However, it can be inferred that Topic 1 is about governments and courts discussing about employment related issues and schemes, particularly MGNREGA (Mahatma Gandhi National Rural Employment Guarantee Act 2005), which is India's largest social security measure that guarantees minimum 100 days of work to all citizens (2022). The most frequent words in Topic 2 match with the results produced in Figure 2. It can be observed that it is associated with transportation, particularly railways and the special 'shramik' trains that carried stranded migrants. Interestingly, this topic also includes food, and the state Bihar, indicating that most of the migrants were originally from Bihar and were returning back there during the lockdown. Results from Topic 3 do not fall into a general topic, but indicate that it is associated with accidents and murders due to words such as murders, homicide, and NHRC (National Human Rights Commission of India, responsible for inquiring into human rights violations by governments or public servants). Topic 4 words also do not provide much interpretation, but indicate that it is associated with grassroot and village (or panchayati) level discussions and political parties (BSP, Bahujan Samajwadi Party) that are popular in villages. Topic 5 produced the most interesting result, as it highlighted the issue that fishermen particularly faced during the first lockdown. Many fishermen, originally from the state of Jharkhand (as some of the words in this topic are districts within this state) were stranded at ports and harbours during the first lockdown due to restricted movement.

Topic Modelling with Gensim requires the words to be tokenized, and creates a dictionary based on these tokens. The dictionary created was filtered so that words that appear in more than 70% documents are removed. Particularly, the Bag of Words model along with LDA was used. Table 3 shows the topic modelling results for this model. Although the topic numbers differ, the results from both the packages are fairly similar. Topic 1 from LDA with TF-IDF and Topic 4 from LDA with Bag of Words are both associated with employment schemes in the rural regions. Figure 8 provides a better interpretation of employment related words in Topic 4 from the Bag of Words model. The results of Topic 2 from LDA with TF-IDF are comparable with the results in both Topic 1 and Topic 2 in LDA with Bag of Words model. Although Topic 3 is not quite interpretable, results from Topic 5 indicate that it refers to health and covid regulations.

Together, both these results that the most important topics for the covid induced revere migration crisis in India were transportation (particularly court and government discussions on the special 'shramik' trains), employment schemes and food. These results are consistent with the results that Agarwal and Sarkar obtained. Additionally, topic modelling result in this project found the stranded fishermen issue as an important topic during the first lockdown as well.

Figure 7: Number of features for each document frequency in TF-IDF

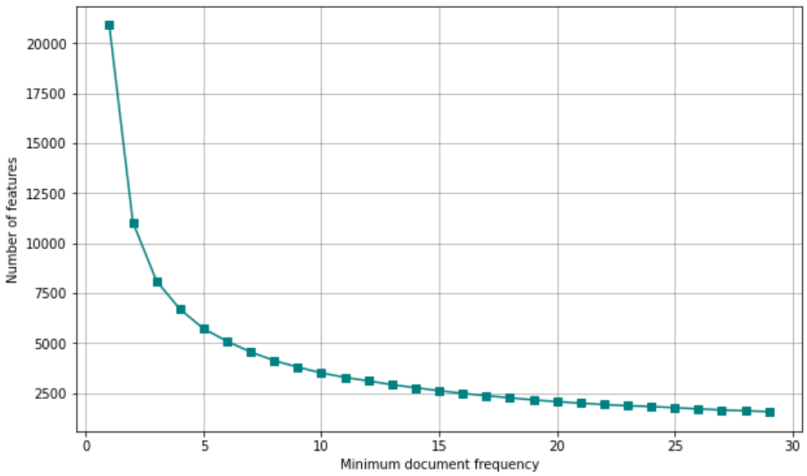


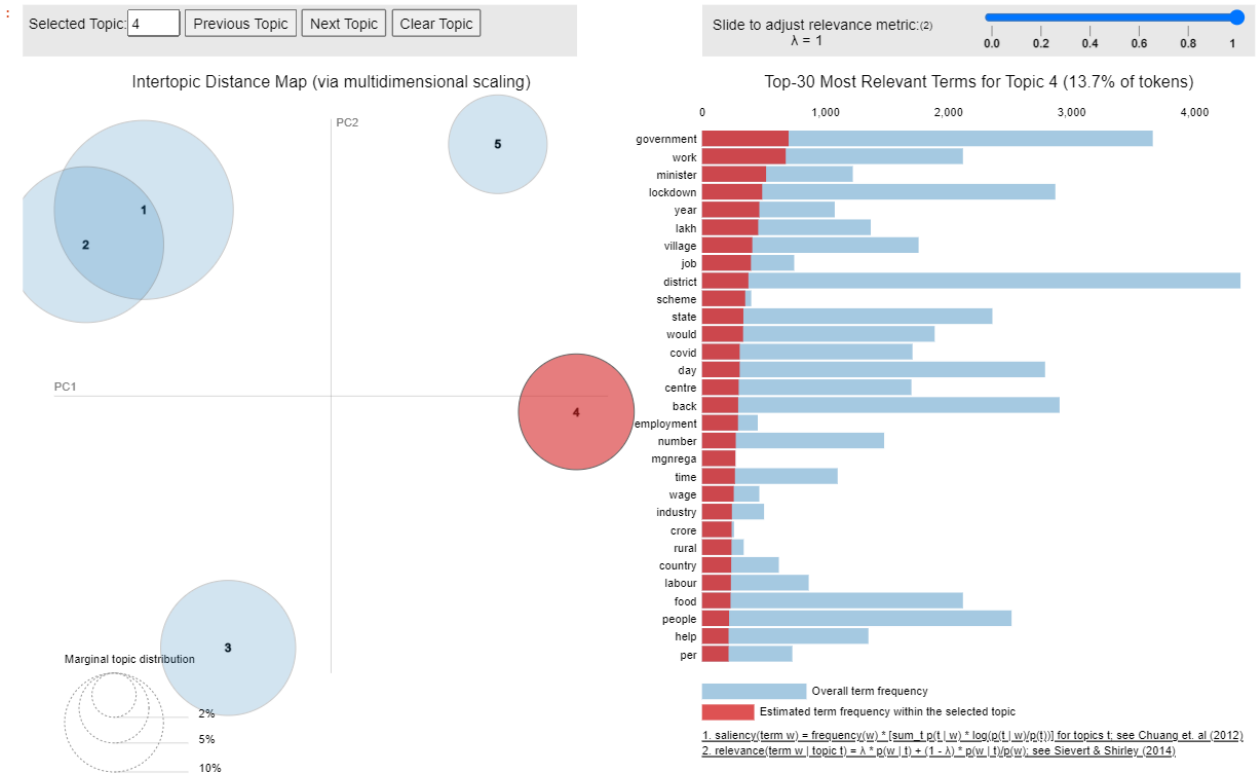
Table 2: Results of Topic Modelling with LDA using TF-IDF

Topic Number	Words
1	court, government, scheme, congress, gandhi, employment, lakh, industry, job, justice, crore, mgnrega
2	train, district, police, government, bus, railway, food, lockdown, stranded, bihar, shramik, quarantine
3	homicide, culpable, nhrc, thrashed, killing, truth, widow, murder, accidental, dismissed, complaint
4	bsp, refuse, collided, unfortunate, attacked, immunity, killing, fallout, panchayati, grassroots, nda,
5	fishermen, fishing, vessel, hazaribag, ranchi, andaman island, singhbum, palamu, boat, harbor

Table 3: Results of Topic Modelling with LDA using Bag-of-Words

Topic Number	Words
1	train, special, railway, government, shramik, station, home, bus, state, official
2	home, police, lockdown, food, work, day, bus, village
3	court, government, bench, train, lockdown
4	government, work, minister, lockdown, lakh, scheme,
5	district, case, quarantine, covid, health, centre, official,

Figure 8: Frequently occurring words for Topic 4 in Topic Modelling with LDA using Bag-of-Words



Conclusion

This project aimed at identifying sentiment and topics in the reverse migration crisis in India from news articles, as there was extensive news coverage on this issue. This was implemented to help understand the issues that urban migrant workers in the country face, and in the push/pull factors that resulted in this mass movement. Our sentiment analysis results indicated an overall neutral sentiment for this issue. This can be ascribed to the inference that the positive and negative sentiments are distributed throughout the article text, thus resulting in an overall neutral sentiment result. The topic modelling results indicated that the most important issue was employment, and that migrants returned back home due to lack of alternate sources of income. This led to MGNREGA being an important term in topics related to employment as well. This highlights that migrants moved to rural areas not only because it was their home town, but because the rural scheme guaranteed 100 days of work. Such economic protection schemes are absent at the urban level. Another important topic that emerged was transportation. It highlights the unpreparedness of the government at the national and local level to deal with mass movements, and required the involvement of courts as well. Other important topics that emerged were food and fishermen being stranded. All these topics indicate that there isn't enough economic and social protection for migrant workers, and highlights their dependence on daily wages for their livelihood. These results show that employment guarantee schemes at the urban level, and alternate or special transport services are essential in India due to the increased presence of internal migration in the country.

The main limitation of this project is that it is limited to English news article from the Times of India and the Hindu. News articles from regional languages would provide more detailed information about the issues at the local level, but they are computationally difficult to implement due to translation issues. Another limitation is that it is only limited to news articles from the first lockdown. Future considerations of this project would include news articles from subsequent lockdowns and from other popular English news media sources in India as well. Additionally, future

implementations of topic modelling for this project would include hyper-parameter tuning to find the ideal number of topics for topic modelling with LDA, instead of using brute-force method to find the same value.

Bibliography

Agarwal, S., & Sarkar, S. (2022). Topical analysis of migration coverage during lockdown in India by mainstream print media. PLOS ONE, 17(2). https://doi.org/10.1371/journal.pone.0263787

Agarwal, Swati; Sarkar, Sayantani (2022), "Topical analysis of migration coverage during lockdown in India by mainstream print media", Mendeley Data, V1, doi: 10.17632/s7j9x7c9bk.1

Amy. (2022, November 18). *Textblob vs vader for sentiment analysis using Python*. Grab N Go Info. Retrieved December 14, 2022, from https://grabngoinfo.com/textblob-vs-vader-for-sentiment-analysis-using-python/

Basuroy, T. (2021, March 19). India - most read English publications 2019. Statista. https://www.statista.com/statistics/885417/india-most-read-english-publications/

Behera, M., Mishra, S., & Behera, A. R. (2021). The COVID-19-led reverse migration on labour supply in Rural Economy: Challenges, opportunities and road ahead in Odisha. The Indian Economic Journal, 69(3), 392–409. https://doi.org/10.1177/00194662211013216

Beri, A. (2020, May 27). *Sentimental analysis using vader*. Medium. Retrieved December 14, 2022, from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

CHANDRASHEKHAR, V. (2020, March 31). *1.3 billion people. A 21-day lockdown. can India curb the coronavirus?* Science. Retrieved December 14, 2022, from https://www.science.org/content/article/13-billion-people-21-day-lockdown-can-india-curb-coronavirus

Chen, M., By, Chen, M., Sinha, S., Sinha, S., Narayan, M., Majithia, A., Talbott, T. C., Talbott, T. C., Chandran, P., Allen, C., Narayan, L., Boampong, O., Pillay, V., Pillay, V., Bett, E., Alfers, L., Alfers, L., Braham, C., … Wiego. (n.d.). To die from hunger or the virus: An all too real dilemma for the poor in India (and elsewhere), https://www.wiego.org/blog/die-hunger-or-virus-all-too-real-dilemma-poor-india-andelsewhere

DeLancey, J. (2020, May 29). *Pros and cons of NLTK sentiment analysis with vader*. CodeProject. Retrieved December 14, 2022, from https://www.codeproject.com/Articles/5269447/Pros-and-Cons-of-NLTK-Sentiment-Analysis-with-VADE

John, E. A., & Kapilashrami, A. (2021). Victims, villains and The rare hero: Analysis of migrant and refugee health portrayals in the Indian print media. Indian Journal of Medical Ethics, 145–155. https://doi.org/10.20529/ijme.2020.131

Khan Jadoon, N., Anwar, W., Bajwa, U. I., & Ahmad, F. (2017). Statistical machine translation of Indian languages: A survey. Neural Computing and Applications, 31(7), 2455–2467. https://doi.org/10.1007/s00521-017-3206-2

Khan, A., & Arokkiaraj, H. (2021). Challenges of reverse migration in India: A comparative study of internal and international migrant workers in the post-covid economy. Comparative Migration Studies, 9(1). https://doi.org/10.1186/s40878-021-00260-2

Liu, Q., Chen, Q., Shen, J., Wu, H., Sun, Y., & Ming, W.-K. (2019). Data Analysis and visualization of newspaper articles on thirdhand smoke: A topic modeling approach. JMIR Medical Informatics, 7(1). https://doi.org/10.2196/12414

Mishra, A. R. (2022, June 15). *Covid may have forced return of 51.6% urban men to villages: PLF survey*. Business Standard News. Retrieved December 14, 2022, from https://www.business-standard.com/article/economy-policy/covid-may-have-forced-return-of-51-6-urban-men-to-villages-plf-survey-122061501191_1.html

Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Applied Sciences*, *11*(18), 8438. https://doi.org/10.3390/app11188438

O. Oyebode and R. Orji, "Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019," *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 0140-0146, doi: 10.1109/IEMCON.2019.8936139.

Pundir, P. (2020, June 11). India's lockdown forced millions out of jobs. now their employers want them back. India's Lockdown Forced Millions Out Of Jobs. Now Their Employers Want Them Back. https://www.vice.com/en/article/n7wzbx/india-lockdowncoronavirus-migrants-right

Shah, P. (2020, November 6). *My absolute go-to for sentiment analysis - textblob.* Medium. Retrieved December 14, 2022, from https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

Topic modeling with Latent Dirichlet allocation. Topic Modeling with Latent Dirichlet Allocation - Trusted Analytics Platform 0.6.0 documentation. (2016, January 8). Retrieved December 14, 2022, from https://pythonhosted.org/trustedanalytics/LdaNewPlugin_Summary.html

Wikimedia Foundation. (2022, November 26). *National Rural Employment Guarantee Act, 2005*. Wikipedia. Retrieved December 14, 2022, from https://en.wikipedia.org/wiki/National_Rural_Employment_Guarantee_Act,_2005

Ylä-Anttila, T., Eranti, V., & Kukkonen, A. (2021). Topic modeling for Frame Analysis: A Study of media debates on climate change in India and USA. Global Media and Communication, 18(1), 91–112. https://doi.org/10.1177/17427665211023984

Implementation Appendix

This appendix provides information about the data pre-processing steps implemented in this project.

While checking for punctuations, Hindi diacritics were found. This was an indication that Hindi words were present in the dataset. Hence, all non-alpha characters from the data was removed to eliminate Hindi words.

Random sampling of the dataset showed that there were 8 articles which only contained the letter head from Times of India. Since this provides no benefit for text analysis on reverse migration, and because it affects sentiment analysis results, these articles were removed from the dataset.

For the article text variable, all stop words were removed and Lemmatization was performed using WordNetLemmatizer from the NLTK module.