USING MASS MEDIA DATA TO ANALYSE THE GROWTH IN VARIOUS INDIAN DISTRICTS

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

by

KONARK VERMA

2018MCS2025

Under the guidance of

Dr. AADITESHWAR SETH

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

JULY 2020

Certificate

This is to certify that the thesis titled **USING MASS MEDIA DATA TO ANALYSE THE GROWTH IN VARIOUS INDIAN DISTRICTS** being submitted by **Konark Verma** for the award of **Master of Technology** in **Computer Science & Engineering** is a record of bonafide work carried out by him under our guidance and supervision at the **Department of Computer Science & Engineering**.

The work presented in this thesis has not been submitted elsewhere either in part or full, for the award of any other degree or diploma. The thesis fulfils the requirements and regulations of IIT Delhi and in my opinion meets the necessary standards for submission.

Dr. Aaditeshwar Seth

Department of Computer Science and Engineering

Indian Institute of Technology, Delhi

Acknowledgment

This dissertation would not have been possible without the assistance, help and guidance of several people who have helped me throughout the course of this project.

I am grateful to my project guide **Dr. Aaditeshwar Seth**, for the opportunity to work under him and be part of the ICT4D team. His guidance was instrumental in giving direction to this work.

I would like to thank all my senior groups for laying the foundation of such an interesting work. I also thank my project associates Kumari Rekha, Mehak Gupta and Jasbir Singh for taking time to collaborate with me and help bring out new ideas. They were always approachable and helped me in making progress. I would also like to thank my family and friends for keeping me motivated.

Konark Verma

2018MCS2025

Department of Computer Science and Engineering

Indian Institute of Technology, Delhi

Abstract

We use the mass media data i.e. day to day news articles collected from various Indian news sources to analyse the growth in various Indian districts. For this we first form a total of 5 collections namely Agriculture, Development, Environment, Industrialization and Lifestyle. Each of these collections contains news articles. Then, we use the entity extraction methods, to figure out the locations present in the articles and use these to find the districts that are being talked about in a particular news article. We also use the classification of districts done on the basis of employment type [1] and the classification of districts done on the basis of their pace of growth that we obtained from the ADI predictions of 2019 [2]. Using both these classifications of districts we form a total of 9 subclasses of districts.

Next, we performed a lot of pilot testing where we have tried techniques like Doc-Tag-2-Vec and LDA on the classes of districts based on the employment type. Finally, we ended up using the Doc-Tag-2-Vec method, which helped us obtain the vector embedding for each document.

These vector embeddings for each subclass of each collection were separated and hierarchical clustering was applied on top of this to obtain similar kinds of articles in a cluster. Then, for each subclass of each collection we found their global centroids using median of all the vector embeddings of documents, and the centroid of each cluster using mean of all the documents of that cluster. Now, for each subclass we found out the top 25 clusters that we closest to their global centroid, and from each of these clusters we found out the top 4 articles that were closest to their corresponding cluster centroid using the cosine similarity. Thus, forming a set of 100 articles, which we refer to as the most explainable articles for a particular subclass, for each collection.

We then used TF-IDF on the set of most explainable articles for each subclass of each collection to obtain the explainable keywords which help us further provide insights to these subclasses of districts.

We have then come up with a list of interesting patterns where in each pattern we compare a couple of subclasses of districts to see how similar and dissimilar they are. For both the subclasses in a pattern we compare them using the cosine similarity value between their global centroids, common keywords that we have for both of them and the unique keywords that we have for each subclass.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Motivation and Problem Statement

There has been a lot of research going around the world where people have been trying to analyse and predict the growth of various geographical locations using various kinds of data including Satellite Imagery data, Census data, Wikipedia data, and much more. But, surprisingly not much attention has been paid to day to day news articles.

We by the means of this project wish to analyse the growth in various Indian districts by using mass media data.

Also, we wish to build on some of the work done by the Satellite team and the Census analysis team. The Census data analysis team had done the classification of Indian districts based on the employment type i.e. Unemp, Agri, and Non-agri districts. The satellite team had done predictions using ADI data where they have predicted the change in number of socio-economic variables. We have used their classifications and predictions in our project to do a district level analysis of the mass media data.

We would like to start by building the collections of articles, based on the 5 different categories on which we would like to analyse the mass media data i.e. Agriculture, Development, Environment, Industrialization, and Lifestyle. We would also want to map these articles to the districts that they are about.

Then, we would like to apply some advanced deep learning-based models like LDA and Doc-Tag-2-Vec to extract the most explainable articles and the explainable keywords that help us provide insights to our data.

We would somehow like to compare the datasets of different districts as well as their keywords to see how are these similar or dissimilar to each other.

1.2. Thesis Outline

The remainder of our thesis is organised as follows. Section 2 contains the Related Work. Section 3 contains the information about the datasets. In this section we have also explained about the huge data corpus that we have, the categories of collection we built for analysis, location resolution using entity extraction and classification of the districts. Next in the Section 4, we explain the approaches we used including LDA and Doc-Tag-2-Vec with Hierarchical clustering to extract the most explainable articles and the explainable keywords. In section 5, we talk about the analysis that we did using the keywords extracted using TF-IDF and the global centroid similarity. We have also done a pairwise comparison of districts, whose analysis is shown in this section. At last in the section 6, we conclude this thesis and also talk about some future scope.

2. Related Work

There has been some research work that is related to our work and most of them are from our own lab.

The first work is the work that was done by the Census data analysis team where they performed a clustering of districts to classify them based on the employment. The labels were Unemp, Agri, and Non-agri. We have used their classification in our project to do the district level analysis.

Second work, is the work done by the Satellite team from our lab, where they predicted the change in number of socio-economic variables using the ADI data. The predictions made were for all the districts and had a value between 0-6. We have performed our classification on top of this predictions. We have made the 0-1 value as Slow growing districts, value 2 as Average growing districts, value 3-4 as Fast-growing districts.

Yet another work similar to ours was the work done by the mass media team from our lab, where they formed the collections of news articles for certain policies and used LDA on top of them. We have taken some inspiration from them in applying this technique ourselves.

There were also few other bits and pieces of ideas and results which were taken from people from other projects collaborating with us, and we are very thankful to them.

3. Datasets

3.1. Mass Media Data Corpus

To form the datasets for our project we have used a huge corpus of news articles which has been crawled from various news sources {1} over a period of past 10 years. This huge corpus of documents contains more than 5M news articles, but we need to filter articles based on the categories on which we want to proceed with our analysis.

3.2. Categories for analysis

For the purpose of our analysis we have formed 5 categories, which are as follows:

a. Agriculture
b. Development
c. Environment
d. Industrialization
e. Lifestyle

Now to collect the articles for the above categories from the corpus, we have used Regex based method {2}. Here, we have provided some keywords manually and the articles that contain one or more of these words is collected to the corresponding category.

The fields that we collected for these articles are as follows:

- Article Id
- Article Title
- Article Text
- Article Date
- Article Entities

Now, the article title and article text will be the data that will be used to train our models, and article id will act as a tag for a given article. Also, article entities will be used for extracting the locations from an article.

3.3. Entity Extraction and Location Resolution

The article entities collected for each article in the previous step contains the list of entities that have been extracted using the tool '**OpenCalais'**.

Here, each entity has a value and type associated with it, but we only want to pick those entity values whose type is either '**City**' or '**ProvinceorState**' as these entity types give us the locations. Thus, for each article we have the location names about which that article is talking about.

But this is not enough, as a place can have multiple names, also sometimes the location name belongs to a part to bigger place, i.e. the location name is a village or a subdistrict which is part of a district. What we actually want is that district name, because we wish to do a district level analysis.

So, for this we collected the **Census data** from the Government of India's official website. This data contains a list of all the subdistrict names and villages names belonging to a particular district. Using this, we formed a mapping of location names to their district ids.

Some other names were also added manually to this mapping, as some districts have multiple names, e.g. Varanasi is known by Kasi and Banaras as well.

Also, there were some conflicting names which were resolved. For this, we chose the more popular place over the less popular place and added it to the mapping.

Now, using this mapping we were able to resolve the location names to their corresponding district ids, and this formed our basis for district level analysis.

3.4. Classification of Districts

3.4.1. Based on Employment type

As mentioned earlier, we collected a classification of districts which was done using the employment data from the census data. According to this classification a total of 593 the districts were divided into the following three classes:

a. Unemployment districts
b. Agriculture districts
c. Non-agri districts

Unemployment districts: The districts with very little employment belongs to this class.

Agriculture districts: The districts which had moderate amount of employment and of this majority is in agriculture domain belongs to this class.

Non-agriculture districts: The districts with comparatively high employment belong to this class.

3.4.2. Based on Pace of Growth

The 2019 predictions based on the ADI data, had values between 0-4.

So, we classified:

The districts with values 0-1 as slow growing districts,

Districts with value 2 as average growing districts,

Districts with value 3-4 as fast-growing districts.

Thus, we formed a classification based on the pace of growth.

3.4.3. Subclassification of the districts

Since, we now have 2 different classification of districts, one based on the employment type and other based on the pace of growth, we can combine both these classifications to form a subclassification of districts. This subclassification contains 9 different subclasses of districts as is shown in the table below:

	SLOW	AVG	FAST
UNEMP	124	75	29
AGRI	145	72	12
N-AGRI	99	32	5

Table-1: Number of districts in each subclass

A value in the above table represents the number of districts in that particular subclass, and all these values sum up to a total of 593 districts.

3.5. Final Collections

Now, we using these subclasses of districts, mapping of locations and regex-based method, we finally map the news articles to their corresponding category and the corresponding subclass.

A couple of points to note are as follows:

- First, an article can belong to multiple subclass and multiple categories, as it can be both about say development and agriculture, also it can have multiple locations in it.
- Second, an article in the corpus need not belong to any of these categories.

The count of the articles is shown in the following 2 tables:

Collections	Total Number of Articles	Unemp districts		
		Slow (124)	Avg (75)	Fast (29)
Agriculture	80221	7506	5155	4884
Development	15984	1426	848	1034
Environment	100038	8720	5704	4711
Industrialization	111291	6466	4017	3898
Lifestyle	234158	18829	12904	15173

Table-2: Total number of articles and articles in Unemp districts

Collections	Agri districts			Non Agri districts		
	Slow (145)	Avg (72)	Fast (12)	Slow (99)	Avg (32)	Fast (5)
Agriculture	**14513**	**5429**	**735**	**53840**	**8955**	**443**
Development	**2077**	**642**	**124**	**11936**	**1571**	**137**
Environment	**12141**	**3885**	**650**	**75165**	**9444**	**1261**
Industrialization	**8379**	**3002**	**325**	**94602**	**9001**	**1126**
Lifestyle	**19616**	**6198**	**925**	**187286**	**19622**	**2017**

Table-3: Articles count in Agri districts and Non-agri districts

4. Methodology

Now, in the following section we will explain the entire methodology for our project.

4.1. Methodology Pipeline

Since, the data we have is unlabelled, we wanted to perform some sort of clustering, therefore the first idea that crossed our mind was to perform clustering using LDA and then analyse the results. But, with this approach the problem that we faced was that there was no way to achieve a ranking of the documents. Also, when we got the results from other method i.e. Doc-Tag-2-Vec the results from LDA were not as appealing as we would have wanted them to be. Therefore, we dropped this approach altogether.

Then, we tried a completely different approach using Doc-Tag-2-Vec. By using this technique, we were able to obtain a 50-dimensional vector embedding for each document, words and tags.

After this, we clustered the documents using hierarchical clustering, and picked top 25 clusters with 4 articles each. These articles formed our set of most explainable articles for each subclass for each collection. We then use TF-IDF based keyword extraction method to get the top 100 explainable keywords. These keywords help us get a better insight.

Also, we performed manual analysis of these most explainable articles to get a deeper look, to see if the results are making sense, which we found to be positive.

Now, to evaluate our method we then came up with a list of interesting patterns, where we did a pairwise comparison of subclasses on all the collections to obtain some exciting results, which are shown in the analysis section.

4.2. LDA

4.2.1. LDA as a technique

LDA’s approach to topic modelling is that it considers each document as a collection of topics in a certain proportion. Here, each topic as a collection of keywords, again, in a certain proportion.

We have merged two layers of LDA. First is an unsupervised LDA while second is a semi-supervised LDA. In this approach after cleaning the data, we compute the coherence value, which is used for assessing the quality of the learned topics. This Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. After selecting the best coherence value, we finalize the number of topics and train our LDA model.

Then, we cluster the whole corpus and extract top 50 keywords from each topic. After getting the keywords we use semi-supervised LDA in order to guide topic models to learn topics of specific interest to a user. After that we recluster the corpus. Finally, we extract the top 100 documents by taking (100/topics) topics form each cluster and analyse them.

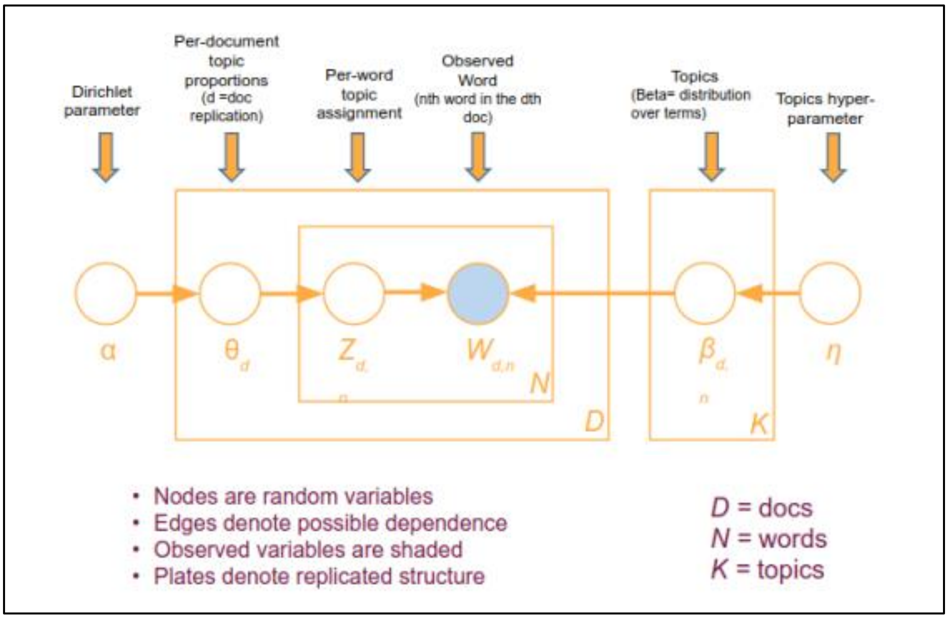

Dirichlet parameter
Per-document topic proportions (d =doc replication)
Per-word topic assignment
Observed Word (nth word in the dth doc)
Topics (Beta= distribution over terms)
Topics hyper-parameter
α
θ d
Z d, n
W d,n
N
D
β d, n
η
K
• Nodes are random variables
• Edges denote possible dependence
• Observed variables are shaded
• Plates denote replicated structure
D = docs
N = words
K = topics

Figure-1: Representation of our LDA Model

In the above figure, α is the per-document topic distributions, η is the per-topic word distribution, θ is the topic distribution for document m, β is the word distribution for topic K, Z is the topic for the nth word in document d, and W is a specific word.

4.2.2. Pilot Testing

To see how this approach was performing we did the following:

a. Manually read the articles to see if it yields a good set of news articles.
b. Performed TF-IDF keyword extraction to see if it yields a good set of keywords.
c. Evalua1ted the model on the basis of the following matrices:

	Agriculture	Development	Environment	Industrialize
Jaccard's Similarity	0.337601	0.272914	0.280433	0.339291
Entropy	7.882912	8.38541	8.359092	7.992541

Table-4: Metrices for pilot testing of LDA Model

4.2.3. Problems with this approach

There are few problems that we faced with this approach moving forward:

a. We can not come up with a direct method for ranking the documents.
b. On comparing this approach with Doc-Tag-2-Vec approach we saw that other approach performed better not only in terms of keywords and documents yielded but also on entropy measures.

Based on these problems we decided to drop this approach completely while moving forward.

4.3. Doc-Tag-2-Vec Model

4.3.1. Doc-Tag-2-Vec as a technique

Another model that we have used is Doc-Tag-2-Vec Model. This model is an unsupervised learning model and an extension to word-2-vec, doc-2-vec kinds of model. As an input we provide it with our tagged documents, here our processed data of news articles and their corresponding tags, from which it not only learns the word and document embeddings but also learns the tag embeddings.

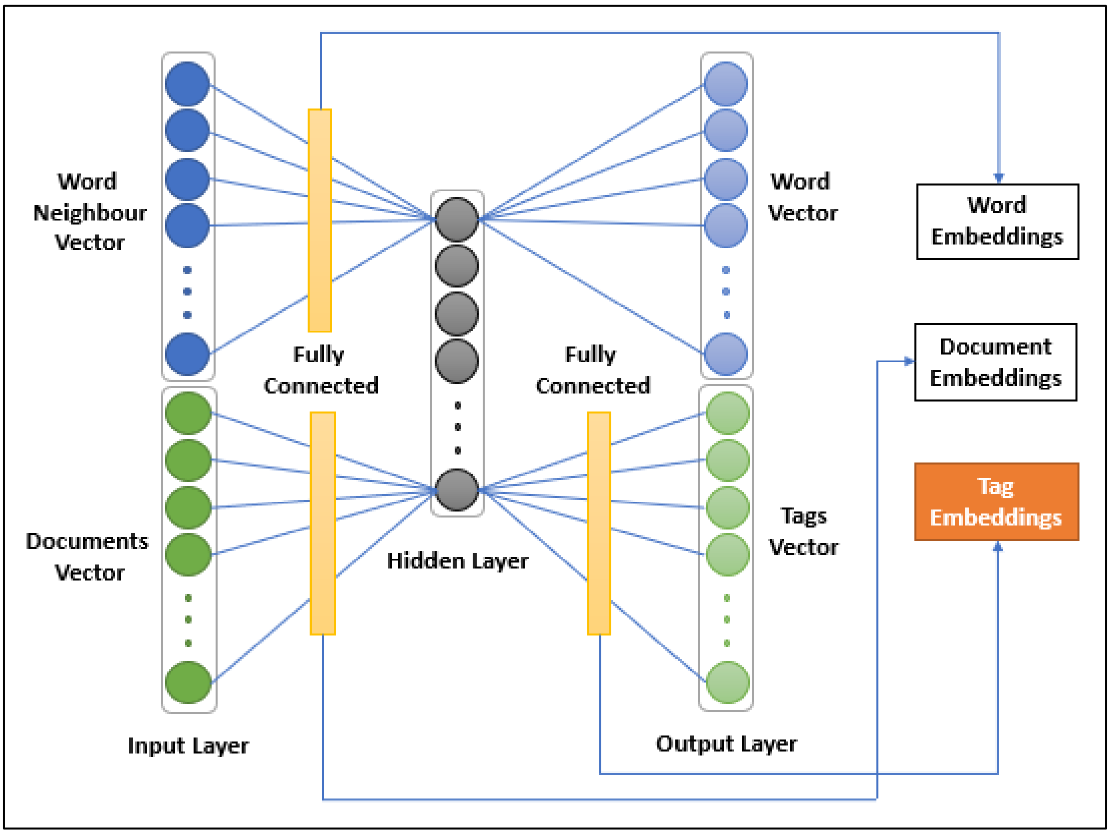


Figure-2: Representation of Doc-Tag-2-Vec Model

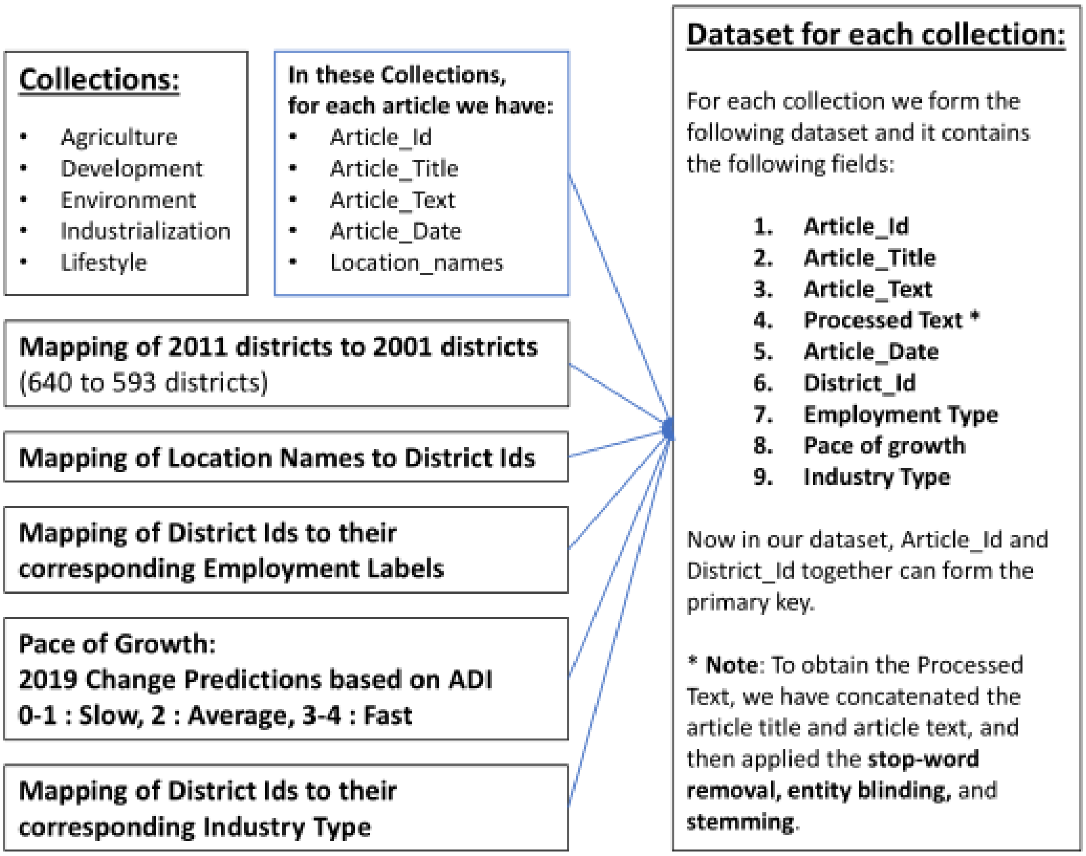
In Doc-Tag-2-Vec Model the dimension of the vector embeddings of words, documents and tags is equal to the number of units in our hidden layer. Also, another important point to note is that the word and document embeddings are learned from one side of the hidden layer while the tag embeddings are learned from the other side.

These document embeddings play a very important role in our project as they not only help us built a ranking method on top of them, but also allows us to cluster the documents and help remove some of the outliers.

4.3.2. Forming the datasets for DT2V Model

First, we have combined the mapping of location names to districts ids, employment labels for districts, and pace of growth for districts with the collection of articles to form our dataset.

Here, an important thing to note is that we have done processing of our data, which contains both the article title and article text. To do this we have removed all the stop words and have blinded entities along with performing stemming.



Collections:

- Agriculture
- Development
- Environment
- Industrialization
- Lifestyle

In these Collections, for each article we have:

- Article_Id
- Article_Title
- Article_Text
- Article_Date
- Location_names

Mapping of 2011 districts to 2001 districts (640 to 593 districts)

Mapping of Location Names to District Ids

Mapping of District Ids to their corresponding Employment Labels

Pace of Growth:
2019 Change Predictions based on ADI
0-1 : Slow, 2 : Average, 3-4 : Fast

Mapping of District Ids to their corresponding Industry Type

Dataset for each collection:

For each collection we form the following dataset and it contains the following fields:

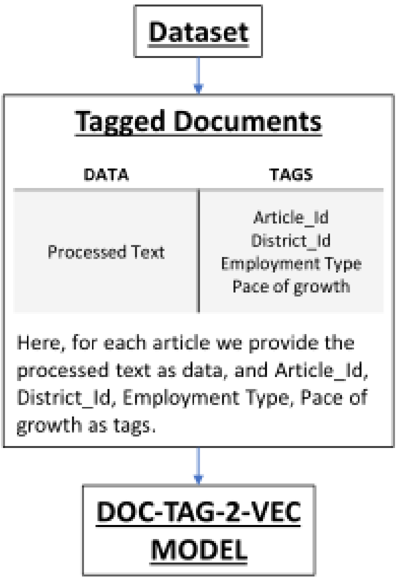
1. **Article_Id**
2. **Article_Title**
3. **Article_Text**
4. **Processed Text ***
5. **Article_Date**
6. **District_Id**
7. **Employment Type**
8. **Pace of growth**
9. **Industry Type**

Now in our dataset, Article_Id and District_Id together can form the primary key.

* **Note**: To obtain the Processed Text, we have concatenated the article title and article text, and then applied the **stop-word removal, entity blinding,** and **stemming**.

Figure-3: Representing the datasets formation

Now, to obtain the final DT2V dataset for each collection we have done the following:



Here, for each article we provide the processed text as data, and Article_Id, District_Id, Employment Type, Pace of growth as tags.

Figure-4: Representing the formation of DT2V datasets.

One thing to note is that we form a total of 5 datasets, a DT2V dataset corresponding to each collection.

4.3.3. Training the Doc-Tag-2-Vec Model

Now, to train this model, there are certain parameters that we have used. The parameters used for training along with some other points are as follows:

1. Each article is considered as a separate document, along with appropriate tags.
2. Tags used: Article Id, District Id, Employment type, Pace of growth.
3. Here, we are getting the both the direct document embeddings as well as tag embeddings for each document. We are ignoring the direct document embeddings and are using the **tag embeddings** for each document.

4. **Parameters** for training the model:
    - Epochs: 100
    - Learning rate: 0.1
    - Minimum learning rate: 0.001
    - Vector size: 50Min occurrences for a word: 3
    - Distributed bag of words (D-BOW)
    - Window size: 3
    - In our model, the learning rate linearly falls down till the value of min-learning-rate.

4.3.4. Pilot Testing

To see if this approach was performing better than the LDA approach we did the following:

d. Manually read the articles to see which approach yield better set of news articles.

e. Performed TF-IDF keyword extraction and manually compared the keywords to see which approach yields better set of keywords.

f. Evalua1ted the model on the basis of the following matrices:

	Agriculture	Development	Environment	Industrialize	Lifestyle
Jaccard's Similarity	0.368768	0.302348	0.310856	0.333287	0.3513144
Entropy	2.90719	2.98013	2.905920	2.871511	2.928785

Table-5: Metrices for pilot testing of DT2V Model

Here, this method clearly performs better than the LDA method.

4.4. Hierarchical Clustering

We have obtained the vector embeddings for all the documents of each subclass for each collection.

Now, we know that there are many similar articles in a dataset of a subclass. To avoid getting very similar articles we perform a clustering on top of the Doc-Tag-2-Vec Model.

We perform this clustering in such a way that the **intra-cluster distance is minimized and the inter-cluster distance is maximized**.

By doing this we observed that the number of clusters formed for most of the subclasses were around 50-80.

We have also calculated the median of all the vector embeddings for a subclass and the resultant vector is called the **Global Centroid**.

4.5. Obtaining the most explainable articles

For each of the cluster a cluster centroid is calculated by taking the mean of all the documents belonging to that cluster. Then, for each subclass we find the closest 25 cluster centroid to their global centroid.

The reason for taking 25 is that even if there are a lot of outlier clusters, we can ignore them as such clusters will not be more than 50 percent.

Now, from each cluster we pick 4 articles which are closest to their corresponding cluster centroid, and thus we form a set of 100 articles.

We refer these articles as the most explainable articles for a subclass, and these articles along with their global centroids are used for further analysis.

5. Analysis

5.1. Analysis Pipeline

We do the following things for each collection:

Step-1: We obtain the global centroid for each subclass of each collection and do a pairwise comparison.

Step-2: We obtain the most explainable articles for each subclass of each collection. Here for each article we have their article title, article text and processed text.

Step-3: We apply TF-IDF on the processed text to obtain the explainable keywords that help us gain more insights.

Step-4: We prepare a list of interesting patterns by taking 2 subclasses at a time and comparing their global centroid similarity, keywords and their titles.

Step-5: To do a keywords-based comparison we obtain the following set of keywords:

a. Obtain those words that occur for all the subclasses of a collection.

b. Obtain those words that are unique to a particular subclass (i.e. the words that occur in a particular subclass but not in any other subclass).

c. For a pattern (subclass-A vs subclass-B) we find out those words that occur for both these subclasses (i.e. both A, B).

d. For a pattern (subclass-A vs subclass-B) we find out those words that occur only for subclass A, and only for subclass B.

Step-6: Next we prepare tables that shows these comparisons clearly.

5.2. Similarity Matrices

5.2.1. Agriculture Collection

We have taken the 50-dimensional global centroid for each subclass for the agriculture collection and have found out the cosine similarity between them. This is shown in the table below:

		Unemp districts			**Agri districts**			**Non-Agri districts**		
		Slow	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**
Unemp	**Slow**	1	0.83	0.63	0.4	0.29	0.42	-0.63	0.06	0.41
	Avg	0.83	1	0.75	0.16	0.06	0.29	-0.49	-0.01	0.45
	Fast	0.63	0.75	1	-0.2	-0.23	0.07	-0.13	-0.2	0.45
Agri	**Slow**	0.4	0.16	-0.2	1	0.56	0.35	-0.72	0.07	-0.34
	Avg	0.29	0.06	-0.23	0.56	1	0.85	-0.51	0.45	-0.27
	Fast	0.42	0.29	0.07	0.35	0.85	1	-0.47	0.42	0
Non-Agri	**Slow**	-0.63	-0.49	-0.13	-0.72	-0.51	-0.47	1	-0.34	0.05
	Avg	0.06	-0.01	-0.2	0.07	0.45	0.42	-0.34	1	0.18
	Fast	0.41	0.45	0.45	-0.34	-0.27	0	0.05	0.18	1

Table-6: Similarity Matrix for Agriculture Collection

5.2.2. Development Collection

We have taken the 50-dimensional global centroid for each subclass for the development collection and have found out the cosine similarity between them. This is shown in the table below:

		Unemp districts			Agri districts			Non-Agri districts		
		Slow	Avg	Fast	Slow	Avg	Fast	Slow	Avg	Fast
Unemp	Slow	1	0.83	0.65	0.63	0.59	0.45	-0.71	0.35	0.07
	Avg	0.83	1	0.78	0.45	0.47	0.4	-0.62	0.28	-0.04
	Fast	0.65	0.78	1	0.09	0.14	0.12	-0.35	0.08	-0.05
Agri	Slow	0.63	0.45	0.09	1	0.88	0.22	-0.83	0.38	-0.21
	Avg	0.59	0.47	0.14	0.88	1	0.4	-0.77	0.53	-0.2
	Fast	0.45	0.4	0.12	0.22	0.4	1	-0.18	0.49	0.37
Non-Agri	Slow	-0.71	-0.62	-0.35	-0.83	-0.77	-0.18	1	-0.47	0.14
	Avg	0.35	0.28	0.08	0.38	0.53	0.49	-0.47	1	0.42
	Fast	0.07	-0.04	-0.05	-0.21	-0.2	0.37	0.14	0.42	1

Table-7: Similarity Matrix for Development Collection

5.2.3. Environment Collection

We have taken the 50-dimensional global centroid for each subclass for the environment collection and have found out the cosine similarity between them. This is shown in the table below:

		Unemp districts			Agri districts			Non-Agri districts		
		Slow	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**
Unemp	**Slow**	1	0.69	0.41	0.65	0.54	0.47	-0.84	-0.03	-0.05
	Avg	0.69	1	0.81	0.19	0.21	0.18	-0.58	0.14	0.09
	Fast	0.41	0.81	1	-0.1	-0.06	0.07	-0.31	0.25	0.13
Agri	**Slow**	0.65	0.19	-0.1	1	0.77	0.51	-0.75	-0.01	-0.29
	Avg	0.54	0.21	-0.06	0.77	1	0.83	-0.62	0.15	-0.13
	Fast	0.47	0.18	0.07	0.51	0.83	1	-0.48	0.12	-0.09
Non-Agri	**Slow**	-0.84	-0.58	-0.31	-0.75	-0.62	-0.48	1	-0.16	0.14
	Avg	-0.03	0.14	0.25	-0.01	0.15	0.12	-0.16	1	0.53
	Fast	-0.05	0.09	0.13	-0.29	-0.13	-0.09	0.14	0.53	1

Table-8: Similarity Matrix for Environment Collection

5.2.4. Industrialization Collection

We have taken the 50-dimensional global centroid for each subclass for the industrialization collection and have found out the cosine similarity between them. This is shown in the table below:

		Unemp districts			Agri districts			Non-Agri districts		
		Slow	Avg	Fast	Slow	Avg	Fast	Slow	Avg	Fast
Unemp	Slow	1	0.89	0.77	0.73	0.75	0.79	-0.83	0.51	0.19
	Avg	0.89	1	0.86	0.52	0.63	0.73	-0.66	0.39	0.13
	Fast	0.77	0.86	1	0.41	0.42	0.55	-0.56	0.52	0.19
Agri	Slow	0.73	0.52	0.41	1	0.82	0.57	-0.74	0.38	0.01
	Avg	0.75	0.63	0.42	0.82	1	0.77	-0.76	0.32	0.05
	Fast	0.79	0.73	0.55	0.57	0.77	1	-0.67	0.44	0.28
Non-Agri	Slow	-0.83	-0.66	-0.56	-0.74	-0.76	-0.67	1	-0.59	-0.09
	Avg	0.51	0.39	0.52	0.38	0.32	0.44	-0.59	1	0.46
	Fast	0.19	0.13	0.19	0.01	0.05	0.28	-0.09	0.46	1

Table-9: Similarity Matrix for Industrialization Collection

5.2.5. Lifestyle Collection

We have taken the 50-dimensional global centroid for each subclass for the lifestyle collection and have found out the cosine similarity between them. This is shown in the table below:

		Unemp districts			Agri districts			Non-Agri districts		
		Slow	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**	**Slow**	**Avg**	**Fast**
Unemp	**Slow**	1	0.88	0.57	0.46	0.53	0.56	-0.68	0.13	0.06
	Avg	0.88	1	0.79	0.39	0.4	0.53	-0.69	0.19	0.11
	Fast	0.57	0.79	1	0.13	0.12	0.39	-0.55	0.11	0.12
Agri	**Slow**	0.46	0.39	0.13	1	0.78	0.53	-0.62	0.31	-0.12
	Avg	0.53	0.4	0.12	0.78	1	0.83	-0.61	0.49	-0.1
	Fast	0.56	0.53	0.39	0.53	0.83	1	-0.51	0.47	0.04
Non-Agri	**Slow**	-0.68	-0.69	-0.55	-0.62	-0.61	-0.51	1	-0.42	-0.07
	Avg	0.13	0.19	0.11	0.31	0.49	0.47	-0.42	1	0.48
	Fast	0.06	0.11	0.12	-0.12	-0.1	0.04	-0.07	0.48	1

Table-10: Similarity Matrix for Lifestyle Collection

5.3. Some General Analysis

While doing the keywords-based analysis, we have observed that the following:

- For **Agriculture collection**,

  the following words**: 'monsoon', 'rain', 'water', 'farmer', 'village', 'land', 'rainfall', 'crop'** do come up on top for all the subclasses.

  On reading some of the articles manually, we see that there are a lot of talks regarding the weather, crops and farmers.

- For **Development collection**,

  the following words**: 'project', 'scheme', 'develop', 'central', 'program'** do come up on top for all the subclasses.

  On reading some of the articles manually, we see that there are a lot of articles about the central government policies, schemes and projects.

- For **Industrialization collection**,

  the following words**: 'project', 'develop', 'power'** does come up on top for all the subclasses.

  On reading some of the articles manually, we see that there are a quite some articles about the electricity, and ways of generating electricity from different sources.

- For **Lifestyle collection**

  the following words**: 'health', 'hospital', 'policies'** do come up on top for all the subclasses.

  On reading some of the articles manually, we see that there are many articles about the hospitals and human-health.

5.4. Initial List of Interesting Patterns

Now we want to do a pairwise comparison of subclasses and since there are nine subclasses, we can form a total of 36 patterns. But we have decided to first stick to some interesting patterns, which are as follows:

Pattern-1: Unemployment districts that are slow growing
(vs) Unemployment districts that are fast growing.

Pattern-2: Agriculture districts that are slow growing
(vs) Agriculture districts that are fast growing.

Pattern-3: Non Agri districts that are slow growing
(vs) Non Agri districts that are fast growing.

Pattern-4: Unemployment districts that are slow growing
(vs) Agriculture districts that are fast growing.

Pattern-5: Agriculture districts that are slow growing
(vs) Non-Agri districts that are fast growing.

Pattern-6: Unemployment districts that are slow growing
(vs) Non-Agri districts that are fast growing.

Now, we will use these patterns to do a do the further analysis.

We will do a comparison on the basis of keywords as well as will show how the titles belonging to the most explainable articles are different for both the subclasses of a pattern.

5.5. Analysis of Patterns

Pattern-1: Unemp districts that are slow growing (vs) Unemp districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing unemp districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing unemp districts
AGRICULTURE	loan, rainfall, suicide, well, agriculture, dam, rain, farmer, protest, monsoon, crop, hit, season, price, sow, collector, cultivation, water, drought, mm, bank, loss	kill, sugar, rabi, sugarcane, insurance, shower, waiver, factories, debt, tanker, storage	reservoir, tenant, field, tank, power, remain, river, canal, market, cattle, inspector, cotton, engine, kharif, onion
DEVELOPMENT	worker, scheme, construction, agriculture, irrigation, drinking water, fodder, given, mgnrega, farmer, job, land, crop, panchayat, employ, policy, welfare, implement, farming, wage, water, drought, bank, labour	pond, fish, aadhaar, municipality, waste, relief, sabha, subsidy, forest, accord, account, current, set, poultry	nrega, school, death, tank, well, borewell, cattle, milk, fund, hospital, electricity, road, women, inspector, revenue, migration, target

Table-11: Keyword comparison table for Pattern-1

Pattern-1: Unemp districts that are slow growing (vs) Unemp districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing unemp districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing unemp districts
ENVIRONMENT	kill, death, poach, irrigation, reservoir, farmer, tribal, court, maoist, arrest, forest, policy, conserve, project, spot, wildlife, tiger, water, investigation, elephant, village	illegal, tree, cub, field, grant, export, power, well, trap, dam, wild, rain, chinkara, ore, honey, encroach, peacock, mine, bail, carcass, tourist, tourism, record, badger, sanctuary, skin	lake, herd, fish, skeleton, naxal, monkey, camp, waterfall, visitor, murder, tusker, women, attack, rape, smuggler, tribe, seize, gun, past, smuggle, adivasi
INDUSTRIALIZATION	illegal, worker, power, scheme, river, construct, toilet, process, farmer, protest, land, panchayat, court, forest, mine, water, sand, miner, acre, coal, transport	handloom, die, cave, ore, mill, textile, weaver, arrest, women, dead, campaign, investigation	agriculture, reservoir, rain, fertile, sabha, urea, bank, health
LIFESTYLE	medic, death, river, die, treatment, farmer, hospital, doctor, protest, crime, children, women, patient, water, health	complaint, illegal, worker, dam, bull, rain, tribal, flu, crop, murder, court, jallikattu, dengue, forest, quarry, rape, mine, victim, farm, disease	school, student, college, culture, pollution, youth, idol, booth, voter, boat, life, festival, education

Pattern-1: Unemp districts that are slow growing (vs) Unemp districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about rainfall, climate changes, farmer protests, suicides and bank loans.
- In slow-growing Unemp districts there are much more articles about cultivation of sugarcane, sugar, rabi crops, tanker, storage facility, and loan waivers.
- In fast-growing Unemp districts there are more articles about canals, reservoirs, electricity, cotton, and kharif crops cultivation.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about schemes regarding farmers, workers, agriculture, irrigation, wages.
- In slow-growing Unemp districts, articles are more about relief funds, subsidies, poultry farming and fishing.
- In fast-growing Unemp districts, articles talk more about cattle, women, schools, migration, electricity, road, and hospitals.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk about irrigation, tribal and maoist activities, and some illegal activities.
- In slow-growing Unemp districts there are much more articles about mining, honey, death of wildlife animals, tourism in wildlife sanctuaries.
- In fast-growing Unemp districts there are more articles about Adivasi, naxal activities, rape, and other crimes.

Pattern-1: Unemp districts that are slow growing (vs) Unemp districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about mining and farmer protests.
- In slow-growing Unemp districts there are much more articles about textile activities.
- In fast-growing Unemp districts there are more articles about agricultural activities.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about medical requirements, deaths, farmer protests.
- In slow-growing Unemp districts there are much more articles about diseases, rape, mining and farming.
- In fast-growing Unemp districts there are more articles about education, culture, voting, pollution and festival celebration.

Pattern-2: Agri districts that are slow growing (vs) Agri districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing agri districts	Keywords that occur much more often in slow-growing agri districts	Keywords that occur much more often in fast-growing agri districts
AGRICULTURE	loan, rainfall, agriculture, rain, farmer, protest, monsoon, water, loan, crop, farm, village, demand	system, variety, suicide, plant, irrigation, grape, dam, wheat, hectare, hailstorm, season, flood, cultivate, paddy	forecast, field, problem, power, consumption, tariff, protest, waiver, electricity, growth, corrupt, sector, develop
DEVELOPMENT	air, scheme, connect, market, route, airport, train, skill, employ, invest, industry, project, flight, technology, develop	msme, assist, worker, manage, power, farmer, agriculture, irrigation, land, crop, cultivate, plant	policy, student, business, fund, service, job, company, investor, heath, infrastructure, digital, data, research

Table-12: Keyword comparison table for Pattern-2

Pattern-2: Agri districts that are slow growing (vs) Agri districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing agri districts	Keywords that occur much more often in slow-growing agri districts	Keywords that occur much more often in fast-growing agri districts
ENVIRONMENT	air, system, student, school, manage, river, waste, land, court, pollution, environment, industry, plant, pollution, develop	medic, tree, survey, agriculture, farmer, conserve, activist, water, village, drain, tourism, forest, proposal	illegal, consult, suicide, construct, poor, electricity, vehicle, clearance, public, education, transport
INDUSTRIALIZATION	manage, power, real, estate, market, business, opportunity, invest, industry, technology, develop	passenger, air, survey, river, hike, waste, land, grass, pollution, water, city, village, demand, steel, sanitisation	policy, term, fuel, airline, domestic, investor, money, spectrum, bank, growth, firm, global
LIFESTYLE	medic, air, college, service, pollution, tourist, water, environment, develop, health	worker, care, problem, student, traffic, agriculture, vehicle, organisation, business, doctor, festival	theatre, show, cbi, culture, market, travel, violence, protest, social, film, industry, invest, terror, crime

Pattern-2: Agri districts that are slow growing (vs) Agri districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about rainfall, climate changes, farmer protests, bank loans.
- In slow-growing Agri districts there are much more articles about cultivation of crops, floods, irrigation projects.
- In fast-growing Agri districts there are more articles about loan and electricity waivers.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about schemes regarding airports and technology related developments.
- In slow-growing Unemp districts, articles are more about agricultural activities and related issues.
- In fast-growing Unemp districts, articles talk more about health, investments and digital development.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk about pollution, waste and other development activities.
- In slow-growing Agri districts there are much more articles about agricultural activities and forest conservation.
- In fast-growing Agri districts there are more articles about forest clearances and transportation.

Pattern-2: Agri districts that are slow growing (vs) Agri districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about business and industry development.
- In slow-growing Agri districts there are much more articles about pollution and waste related problems.
- In fast-growing Agri districts there are more articles about investment and monetary policies.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about protests, tourism, hospital services.
- In slow-growing Agri districts there are much more articles about festivals, facilities, policies.
- In fast-growing Agri districts there are more articles about violence, protests, crimes.

Pattern-3: Non-Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing non-agri districts	Keywords that occur much more often in slow-growing non-agri districts	Keywords that occur much more often in fast-growing non-agri districts
AGRICULTURE	loan, march, rainfall, kill, field, normal, death, suicide, agriculture, compensation, rain, farmer, flood, harvest, monsoon, protest, paddy, village, forest	Fire, school, student, die, condition, scheme, seed, hailstorm, season, cultivate, education	sanction, power, construct, irrigation, dam, wheat, delay, rape, expressway, weather, life, develop
DEVELOPMENT	Complaint, worker, person, scheme, connect, water, wage, mnrega, health, develop, programme, farmer, job, connect	School, problem, sanction, airport, bank, household, claim, inquiry, law, labour	Fire, dam, grant, poor, budget, scam, protest, elect, corrupt, arrest, dalits, infrastructure

Table-13: Keyword comparison table for Pattern-3

Pattern-3: Non-Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing and fast-growing non-agri districts	Keywords that occur much more often in slow-growing non-agri districts	Keywords that occur much more often in fast-growing non-agri districts
ENVIRONMENT	fire, kill, death, herd, river, die, poacher, wildlife, wild, rain, rescue, tourism, forest, attack, animal, develop	Poach, crocodile, reserve, farmer, land, crop, blackbuck, tourist, protest, bauxite, mine	Ganga, march, student, suicide, leopard, drive, pollution, clean, troop, zoo, protect, city
INDUSTRIALIZATION	Illegal, worker, school, kill, student, power, river, bandh, recover, train, rescue, area, plant, coal, develop	Fire, estate, vehicle, support, toilet, rain, tribal, protest, flood, maoist, weaver, attack, rape, victim, protest	Yatra, pond, death, farmer, criminal, mafia, miner, labour, heath, city, health, industry
LIFESTYLE	Medic, fire, night, kill, death, power, violence, hospital, doctor, protest, murder, crime, flood, rape, attack	Die, vehicle, military, rain, highway, curfew, strike, youth, dispute	Worker, politician, loot, relief, criminal, terror, cop, women, rally, demand, team

Pattern-3: Non-Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about rainfall, floods and farmer suicides.
- In slow-growing Non-Agri districts there are much more articles about hailstorms and crop cultivation.
- In fast-growing Non-Agri districts there are more articles about irrigation, dams and sowing crops.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about schemes regarding agriculture and technology related developments.
- In slow-growing Non-Agri districts, articles are more about sanction of funds and law related claims.
- In fast-growing Non-Agri districts, articles talk more about scams, protests and corruption.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk about poaching and rescue of wildlife.
- In slow-growing Non-Agri districts there are much more articles about agriculture and mining activities and tourism.
- In fast-growing Non-Agri districts there are more articles about pollution and protection of wildlife.

Pattern-3: Non-Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about worker issues and mining activities.
- In slow-growing Non-Agri districts there are much more articles about pollution and protests.
- In fast-growing Non-Agri districts there are more articles about criminal activities and healthcare.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about rapes, attacks, murders and other crimes.
- In slow-growing Non-Agri districts there are much more articles about disputes and defence activities.
- In fast-growing Non-Agri districts there are more articles about violence, crime and terrorism.

Pattern-4: Unemp districts that are slow growing (vs) Agri districts that are fast growing

	Keywords that occur in both slow-growing unemp and fast-growing agri districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing-agri districts
AGRICULTURE	loan, rainfall, supply, well, agriculture, rain, farmer, protest, monsoon, land, crop, waiver, center, house, project, farm, water, bank, committee, village, demand	sugar, suicide, meter, rabi, sugarcane, drink, fodder, dam, shower, relief, factory, season, debt, sow, tanker, damage, drought, loss, amount, storage	forecast, home, power, consume, tariff, elect, company, secure, court, draft, food, metro, invest, charge, law, park, growth, corrupt, opposite, finance, increase
DEVELOPMENT	scheme, aadhaar, job, employ, children, center, house, project, wage, product, bank, program, develop	sanction, fish, mgnrega, farmer, relief, land, crop, season, forest, village, worker, power, agriculture, poultry, construct, waste, sabha, subsidy, accord, welfare, drought, panchayat, water, labour	policy, air, ticket, support, budget, skill, elect, invest, investor, mission, health, airport, research, manufacture, smart, technology, infrastructure, connect, market, fund, company, railway, industry, flight, school, student, train, court

Table-14: Keyword comparison table for Pattern-4

Pattern-4: Unemp districts that are slow growing (vs) Agri districts that are fast growing

	Keywords that occur in both slow-growing unemp and fast-growing agri districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing-agri districts
ENVIRONMENT	illegal, power, project, land, company, reside, court, well	kill, death, farmer, maoist, arrest, encroach, forest, conserve, tourist, dam, rain, protect, bail, sanctuary, demand, cbi, export, leopard, carcass, tourism, badger, ore, tribal, honey, mine	air, river, electricity, pollute, metro, road, oil, **ngt**, children, invest, quality, ban, education, system, consult, problem, traffic, global, construct, poor, vehicle, municipal, women, industry, city, school, student
INDUSTRIALIZATION	cbi, fund, high, power, project, company, house, industry, court, coal	river, temple, farmer, electricity, mill, textile, forest, miner, activist, driver, worker, protest, construct, weaver, transport, illegal, handloom, bench, process, ore, mine	sale, invest, quality, data, airline, estate, call, institute, global, manufacture, sell, technology, fuel, market, kyc, telecom, gst, bank, student, import, wto, trade, spectrum, service, tax, foreign, point, drug, dollar, aircraft
LIFESTYLE	medic, hospital, protest, crime, accuse, arrest, women, rape, police, project, water, tourism, health	worker, river, swine, dam, bull, farmer, rain, flu, treatment, tribal, children, jallikattu, dengue, forest, quarry, mine, disease, woman, screen, program	theatre, air, college, political, culture, poll, market, travel, violence, pollute, gender, leader, agency, life, interest, tourist, brand, industry, terror, drug, law, movie, film

Pattern-4: Unemp districts that are slow growing (vs) Agri districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about rainfall, waivers and agriculture.
- In slow-growing Unemp districts there are much more articles about farmer suicides and crop cultivation.
- In fast-growing Agri districts there are more articles about tariff, metro and laws.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about schemes regarding employment, aadhaar and wages.
- In slow-growing Unemp districts, articles are more about agriculture and poultry related activities.
- In fast-growing Agri districts, articles talk more about technology, infrastructure and transportation.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk illegal activities.
- In slow-growing Unemp districts there are much more articles about farmers, tourism, conservation of wildlife.
- In fast-growing Agri districts there are more articles about pollution, metro, education and transportation.

Pattern-4: Unemp districts that are slow growing (vs) Agri districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about electricity, coal and funding activities.
- In slow-growing Unemp districts there are much more articles about workers, protests, transportation.
- In fast-growing Agri districts there are more articles about real estate, telecommunication, manufacturing, banking and finance activities.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about rapes, women, medical facilities and also tourism.
- In slow-growing Unemp districts there are much more articles about mining and tribal activities.
- In fast-growing Agri districts there are more articles about theatre, travel, violence, and tourism.

Pattern-5: Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing agri and fast-growing non-agri districts	Keywords that occur much more often in slow-growing agri districts	Keywords that occur much more often in fast-growing non-agri districts
AGRICULTURE	loan, rainfall, suicide, agriculture, dam, rain, farmer, monsoon, crop, wheat, price, flood, forest, farm, water, bank, loss, weather, harvest, village, demand	supply, variety, pond, scheme, market, stock, grape, shower, hailstorm, season, strike, winery, sow, horticulture, tanker, temperature, onion,	kill, field, sanction, death, power, college, river, gang, construct, relief, hospital, protest, court, attack, rape, expressway, life, police, tiger, product, claim, cabinet, develop
DEVELOPMENT	worker, scheme, construct, connect, airport, farmer, tribal, highway, train, land, employ, women, project, implement, water, program, village, develop	air, power, college, estate, agriculture, prison, market, subsidy, skill, crop, enterprise, textile, bird, institute, youth, invest, horticulture, industry, flight, park, smart, entrepreneurship, storage, city, plant, technology, entrepreneur	complaint, school, sanction, student, die, girl, fund, yojana, job, court, road, bridge, mnrega, maoist, child, commission, police, beti, wage, pmgsy, campaign, engine, household, law, bank, monitor, labour, health, demand

Table-15: Keyword comparison table for Pattern-5

Pattern-5: Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing agri and fast-growing non-agri districts	Keywords that occur much more often in slow-growing agri districts	Keywords that occur much more often in fast-growing non-agri districts
ENVIRONMENT	river, farmer, land, court, bird, area, forest, police, project, water, tourism, reside, village, develop	medic, air, plastic, hospital, conserve, quality, activist, mayor, clean, noise, plant, green, municipal, waste, commission, industry, drain, school, student, NGO, treatment, puja, dispose, cover, swachh	fire, kill, death, poach, electricity, blackbuck, relief, crop, maoist, arrest, guard, tourist, rain, camp, recover, protest, flood, attack, tiger, sanctuary, college, crocodile, poacher, wild, park, herd, tribal, wildlife
INDUSTRIALIZATION	school, student, power, river, estate, construct, well, toilet, train, land, company, sanity, mine, water, reside, village, plant	air, survey, connect, hike, shop, market, airport, waste, pollute, tax, textile, garbage, institute, children, commission, invest, industry, product, engine, manufacture, wine, technology, demand, steel	illegal, fire, worker, kill, vehicle, rain, tribal, protest, murder, rescue, raid, flood, weaver, arrest, forest, dengue, attack, rape, agency, police, victim, rebel, coal
LIFESTYLE	medic, vehicle, service, hospital, doctor, flood, police, patient, water, village, health, develop	air, kumbh, farmer, pollute, children, tourist, worker, problem, traffic, agriculture, dam, institute, research, disease, product, college, toilet, municipal, mela, women, tourism, screen,	fire, kill, death, die, violence, land, elect, strike, arrest, power, poll, rain, highway, protest, spot, law, force, reside, murder, rape, vote, dispute, secure, curfew,

Pattern-5: Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about rainfall, floods and farmer suicides.
- In slow-growing Agri districts there are much more articles about different cultivation practices, schemes to resolve farmers issues, strikes of farmers, situation of market, effect of temperature.
- In fast-growing non-agri districts there are much more articles about sanction of funds, power, construction, hospital, field of crops.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about schemes regarding village development schemes and technology related developments.
- In slow-growing Agri districts, articles are more about horticulture, production of crops, implementation of technology, skills of farmers, facilities to farmers.
- In fast-growing non-agri districts, articles talk more about students, IT jobs, air pollution, bank and health facilities.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk about forest clearance wildlife and tourism.
- In slow-growing Agri districts there are much more articles about programmes for wildlife, Forest clearance, plantation, industries,
- In fast-growing non-agri districts there are more articles about relief factors, wildlife, mining, tourism, agriculture fire.

Pattern-5: Agri districts that are slow growing (vs) Non-Agri districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about illegal mining and villagers’ issues
- In slow-growing Agri districts there are much more articles about industry, textile, studies, pollution, investment.
- In fast-growing Non-Agri districts there are more articles about forest, illegal mining, rape, cop duties, temperature.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about health awareness and crimes like rape, violence.
- In slow-growing Agri districts there are much more articles about pollution, tourism, woman and children, treatment, health program, dam.
- In fast-growing Non-Agri districts there are more articles about strikes, minorities success, murder, sexual harassment, raids.

Pattern-6: Unemp districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing unemp and fast-growing non-agri districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing non-agri districts
AGRICULTURE	loan, rainfall, kill, suicide, agriculture, dam, rain, farmer, relief, protest, monsoon, crop, policy, farm, water, mm, loss, part, village, demand	sugar, rabi, sugarcane, fodder, shower, waiver, factory, season, debt, sow, collector, cultivation, taluka, drought, storage, committee	sanction, power, river, gang, college, construct, hospital, wheat, flood, arrest, forest, express-way, rape, product, cabinet, weather, harvest, develop
DEVELOPMENT	worker, sanction, scheme, construct, mgnrega, farmer, job, panchayat, children, area, police, implement, water, wage, bank, committee, labour, program, sever, demand, develop	benefit, pond, power, fish, agriculture, fodder, aadhaar, municipal, waste, relief, subsidy, crop, season, rate, zilla, forest, welfare, farmer, product, drought, poultry, progress, yield, distribute	school, student, connect, airport, fund, yojana, highway, train, court, road, bridge, commission, payment, women, beti, **pmgsy**, campaign, engine, household, law, health

Table-16: Keyword comparison table for Pattern-6

Pattern-6: Unemp districts that are slow growing (vs) Non-Agri districts that are fast growing

	Keywords that occur in both slow-growing unemp and fast-growing non-agri districts	Keywords that occur much more often in slow-growing unemp districts	Keywords that occur much more often in fast-growing non-agri districts
ENVIRONMENT	kill, death, wild, rain, farmer, tribal, court, maoist, arrest, forest, mine, police, spot, wildlife, tiger, tourist, water, tourism, sanctuary, village, demand	illegal, cub, field, grant, export, power, leopard, trap, dam, chinkara, company, cost, honey, encroach, peacock, protect, conserve, carcass, badger, skin, leas	fire, college, river, seat, die, herd, poacher, blackbuck, camp, relief, protest, crop, murder, secure, train, rescue, flood, bird, morn, attack, guard, park, encounter
INDUSTRIALIZATION	illegal, worker, power, river, construct, toilet, protest, elect, land, company, panchayat, court, arrest, area, forest, attack, mine, police, project, water, sand, coal	justice, handloom, die, temple, trap, fund, mine, ore, farmer, mill, textile, women, ban, dead, iron, campaign, miner, park, activist, file, driver, transport, leas	fire, school, kill, student, estate, vehicle, rain, tribal, train, rescue, murder, raid, flood, maoist, dengue, agency, rape, victim, stone, record, seize, temperature, plant, rebel
LIFESTYLE	medic, fire, death, die, communication, rain, service, hospital, doctor, protest, land, elect, murder, court, crime, station, arrest, rape, police, patient, victim, water, village, sever, health	illegal, worker, scheme, river, swine, dam, bull, farmer, flu, treatment, high, tribal, crop, children, jallikattu, dengue, forest, quarry, mine, farm, ban, disease, tourism, record, woman, screen	kill, power, gang, girl, vehicle, military, minor, violence, highway, curfew, strike, flood, maoist, youth, life, army, restrict, vote, law, force, dispute, surrender

Pattern-6: Unemp districts that are slow growing (vs) Non-Agri districts that are fast growing

From Agriculture Collection we see the following:

- In both subclasses, there are a lot of talk about monsoon, projects, farmers issues and facilities
- In slow-growing unemp districts there are much more articles about schemes, sugarcane, season, cultivation, drought.
- In fast-growing non-agri districts there are much more articles about sanction of funds, paddy crop, weather, colleges.

From Development Collection we see the following:

- In both subclasses, there are a lot of talk about technology development, bank facilities, schemes and programs for farmers and workers.
- In slow-growing unemp districts, articles are more about Fisheries, welfare, drought, production of crop, studies.
- In fast-growing non-agri districts, articles talk more about schools ranking, airport facilities, bank projects, internet connectivity.

From Environment Collection we see the following:

- In both subclasses, there are a lot of talk about illegal mining, wildlife conservation.
- In slow-growing unemp districts there are much more articles about Illegal activities, conservation of wildlife, carcass, relocation of animals.
- In fast-growing non-agri districts there are more articles about murder security, crop, rescue, relief.

Pattern-6: Unemp districts that are slow growing (vs) Non-Agri districts that are fast growing

From Industrialization Collection we see the following:

- In both subclasses, there are a lot of talk about illegal mining, coal sector, worker’s issues.
- In slow-growing unemp districts there are much more articles about farmers help, demand of facilities, environment, mining, woman issues
- In fast-growing Non-Agri districts there are more articles about real estate, raids, defence, residents’ facilities.

From Lifestyle Collection we see the following:

- In both subclasses, there are a lot of talk about health issues, crime, rape.
- In slow-growing unemp districts there are much more articles about welfare scheme, health and care, woman, illegal activities, facilities.
- In fast-growing Non-Agri districts there are more articles about youth success, strike, army, safety and security, disputes, law.

6. Conclusions and Future Scope

6.1. Conclusions

We have come a long way in this project. First, we started by building the regex-based methods to form the collections of articles, based on the 5 different categories i.e. Agriculture, Development, Environment, Industrialization, and Lifestyle. We then moved to building the entity extraction and location resolution method for identifying the locations that are being talked in the news articles. We did take a look on classification of our 593 districts based on employment type as well as the pace of growth. Then, we combine all of these to finally form our collections.

On these collections we ran two separate approaches, one using LDA and other using Doc-Tag-2-Vec and Hierarchical Clustering. We evaluated both these approaches to see which yields better explainable articles as well as better keywords, to later drop the LDA approach and continue forward with the DT2V approach.

Then we came up with a list of interesting patterns where we did the pairwise comparison of 2 subclasses of districts. We compared the keywords that were generated using this, as well as their global centroid similarity and article titles.

Finally, we would like to conclude that this has been a really amazing project, where we not only took a problem and built a solution for it, but also learned a lot along the way.

6.2. Future Scope

Though, a lot of work has been done in this project, but this in no way seems over. There are multiple things that we can do further:

1. There are a lot more patterns, whose analysis can be done.
2. Time based analysis can be done, which can yield some good analysis.
3. A lot more data can be collected with time and we can see how our method applies to them, and what interesting results it could yield.

Appendix

{1} Various news sources:

Hindustan Times
Times of India
The Hindu
The Indian Express
Deccan Herald

{2} Seed keywords used for Regex based classification of articles

Agriculture

agri, pesticide, insecticide, kharif crop, kharif-crop, rabi crop, rabi-crop, crops, monsoon, irrigate, farmer (includes farmer protest, farmers rally farmers distress), loan waiver, bhartiya kisan sangathan, bks, pradhan mantri fasal bima, pm fasal bima,national agriculture market , enam, pmksy, pkvy, pradhan mantri krishi sinchayee, pm krishi sinchayee, paramparagat krishi vikas, pm kisan yojna, pradhan mantri kisan yojna

Development

development scheme, development program, pradhan mantri gram sadak, national rural employment guarantee, mgnrega, nrega, pmgsy, make in india, jan dhan yojna, beti bachao beti padhao, digital india, stand up india, prime minister ujjwala plan,pm ujjwala plan gramoday se bharat uday, shramew jayate , ujjwala scheme, udan, regional connectivity scheme, smart cities mission, skill india mission, national career service, egovernance, egov, aadhaar, pds, ration, nutrition, malnutrition, sanitation, hygiene, immunization, vaccines, ayusman bharat, rsby

Environment

forest, eco, environment, deforestation, wildlife, pollution, swachh bharat mission, swachchh bharat mission, swachhgram, clean india mission, pmfby, fra, integrated conservation and development, icdp, jfm, poaching, ntfp, tiger, leopard, zool

Industrialization

coal, lignite,steel product, industry, leather product,crude petroleum, metal product, textile, fertilizer, pesticide, enterprise,prime minister employment generation programme,estate, pmegp,credit guarantee trust fund for micro & small enterprises,cgt sme,mine, mining, stock market, equity market, share market, factory

Lifestyle

lifestyle, life-style, fashion, art, art and culture, health tips, tourism, culture, travel, tech, spirituality, astrology, celebrity, riot, movie, crime, violence, communal, hatred, fake news, misinformation, migration, suicide

References

[1] Dibyajyoti Goswami, Shyam Bihari Tripathi, Sansiddh Jain, Shivam Pathak, and Aaditeshwar Seth. 2019. Towards Building a District Development Model for India Using Census Data. In ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) http://www.cse.iitd.ernet.in/~aseth/district-development-model.pdf

[2] Chahat Bansal, Arpit Jain, Phaneesh Barwaria, Anuj Choudhary, Anupam Singh, Ayush Gupta, and Aaditeshwar Seth. 2020. Temporal Prediction of Socio-economic Indicators Using Satellite Imagery. In 7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3371158.3371167

[3] DocTag2Vec: An Embedding Based Multi-label Learning Approach for Document Tagging, Sheng Chen, Akshay Soni, Aasish Pappu , Yashar Mehdad, University of Minnesota-Twin Cities, Minneapolis, MN 55455, USA Yahoo Research, Sunnyvale, CA 94089, USA and New York, NY 10036, USA Airbnb, San Francisco, CA 94103, USA https://arxiv.org/pdf/1707.04596.pdf