Classifying and Analysing Human-Wildlife Conflicts in India using News Articles

Gokhan Egri[1*] , **Xinran Han**[1*] , **Zilin Ma**[1*] and **Sunandan Chakraborty**[2]
[1]Harvard University
[2]Indiana University
{gegri, xinranhan, zilinma}@g.harvard.edu, sunchak@iu.edu

Abstract

Human-wildlife conflict (HWC) is one of the most pressing conservation issues of our time, with incidents leading to human injury and death, crop and property damage and livestock predation. Since acquiring real-time data and performing manual analysis on those incidents are costly, we propose to leverage machine learning techniques to build an automated pipeline to construct an HWC knowledge base from historical news articles. Our unsupervised and active learning methods are not only able to recognize the major causes of HWC such as construction, pollution and farming, but can also classify an unseen news article into its major cause with 90% accuracy. Moreover, our interactive visualizations of the knowledge base illustrate the spatio and temporal trend of human wildlife conflicts across India for index by cities and animals. We hope that our findings can inform the public of HWC hostspots and help future policy making. For more details, please visit our project website at https://egrigokhan.github.io/hwc-article-analysis/.

1 Introduction

India is home to a large variety of wildlife. Its forest land, which covers about 22% of its total geographical area, nurtures more than 500 species of mammals. However, with growing population, deforestation, etc., human communities are moving into habitats previously home to wildlife. With one of the world's largest rural populations living closely to important species such as Bengal tigers, Asian elephants, Indian rhinos, human-wildlife conflicts is a major issue in those areas. Previous studies [Gulati *et al.*, 2021] show that India's cost of human-elephant conflict leads to about 10k to 15k damaged properties, 400 human deaths and 100 dead elephants per year.

While previous studies often focus on the economic cost of human wildlife conflicts through case studies, there are few reports on the overall severeness of the problem and how those conflicts vary spatially and temporally with different animal species. To better inform the public and provide useful information for decision makers in related institutions, we believe it is important to aggregate a common knowledge base of the spatio-temporal patterns of those conflicts and their major causes. Real-time data about human wildlife conflicts is hard to obtain, often requiring dedicated interviews with local residents around the country. It is also costly to manually analyze and update those statistics on a large scale that spans a long time period.

As a result, we take an active learning approach and apply natural language processing techniques to build a scalable and automated pipeline for building the HWC knowledge base. Since many of the human-wildlife conflicts are routinely reported by newspapers such as the Times of India, the most circulated English newspaper in India, articles citing these instances act as viable proxies to real-time statistics for HWC analysis.

In our project, we utilized publicly available news articles from the Times of India website from a 12-year window (2006-2018) to analyze the trend and causes of human wildlife conflicts in India. The contribution of our work is two-fold:

- We build the first automated and scalable pipeline for creating a knowledge base of Human-Wildlife Conflicts from news data
- We provide interactive visualizations that demonstrate spatial hotspots and temporal trends of human-wildlife conflicts in India in the past 12 years

2 Related Work

Wildlife Conservation Researchers and non-profit organizations for wildlife conservation have traditionally focused on topics such as illegal trade and poaching investigation, population monitoring and tracking. The importance of protecting biodiversity and wildlife has also attracted the attention of AI researchers. For instance, [Bain *et al.*, 2019] used vision-based face and body detection approaches to recognize animals in the wild. [Di Minin *et al.*, 2019] proposed to use machine learning to investigate illegal wildlife trade on social media platforms through data mining, filtering and identification for related verbal, visual and audiovisual features.

In recent years, with growing number of cases of human wildlife conflicts, we observe related work addressing this dilemma through measuring the economic cost of

[*] Equal contribution

those conflicts [Gulati *et al.*, 2021] or identifying the causes and preventative measures of those conflicts [Manral *et al.*, 2016][Margulies and Karanth, 2018]. As far as we are aware, our work is first in applying AI techniques for analyzing human wildlife conflicts in India.

News Article Analytics Many previous work have used news articles for information retrieval and decision making. In the finance field, [Chang *et al.*, 2016] proposed a tree-structured LSTM model to measure news sentiment and their relation with abnormal market returns. [Hagenau *et al.*, 2013] used an automated feature selection scheme for mining financial news data and predicting stock prices. Other works ([Ming *et al.*, 2014], [Zhang and Skiena, 2010]) share similar goals of mining news text for trading strategies. However, using news data for non-profit social impact purposes has received little attention and we hope that our work can demonstrate the first steps towards utilizing news articles for wildlife conservation and help future policy making.

3 Methodology

3.1 Dataset

Our dataset consists of 758,000 newspaper articles collected from the Times of India (ToI) website[1] between 2006 and 2018. The Times of India is the largest circulated English daily in India and aggregates news from around India.

We make the following observations on the dataset: (1) The dataset is not pre-filtered for Human-Wildlife Conflict research and contains articles on a wide variety of topics. (2) We observe that the dataset also contains a small percentage of duplicate and corrupted entries caused by archive listing errors in the ToI website. (3) Finally, we observe that our dataset does not contain consistent meta-data such as geotags or subject tags.

3.2 Filtering

Based on these observations, we filter our dataset using a four stage filtering pipeline as:

1. Wildlife/city filtering
2. Duplicate/corrupt entry removal
3. Latent Dirichlet allocation (LDA)
4. Human-wildlife conflict filtering

Wildlife/city filtering. We first filter for the wildlife mentions by first filtering for a pre-compiled set of 594 animal names[2]. We modify the pre-compiled list of animals by manually removing domesticated animals which are frequently mentioned but are irrelevant to human-wildlife conflict. After filtering the dataset with the pre-compiled list of wildlife animals, we identify the top fifteen animals by word occurrence and re-filter the dataset accordingly.

We filter for Indian city mentions by observing that the majority of entries in the dataset start with the city name followed by a colon (":") and filter accordingly by comparing with a list of Indian cities.

This preliminary filtering reduces our dataset count to 55K articles.

Duplicate/corrupt entry removal. While multiple instances of the coverage of the same event in a dataset are potentially meaningful and indicative of the significance of the effect, we observe that the causes of duplication in our dataset is caused by errors in the ToI website and are therefore irrelevant. As such we choose to remove these duplicate entries to prevent spurious estimates.

Removing these duplicate and corrupt entries further reduces our dataset count to 50K articles.

Latent Dirichlet allocation (LDA). Latent Dirichlet allocation, or LDA for short, is a topic-modeling algorithm commonly used for clustering unsupervised text datasets.

LDA assigns a set of words $W_i = \{w_{i,0}, w_{i,1}, \ldots, w_{i,L_i}\}$ to each topic t_i, and assigns each document d_j to a distribution of topics $\{t_0, t_1, \ldots, t_N\}$ by calculating the probability $P(t_k|d_j)$ that a given document d_j belongs to the topic t_k.

The algorithm starts by randomly assigning each word w_i in every document d_j to a topic t_k. It then calculates the ratio of words w_i in a document d_j which belong to a topic t_k to estimate the probability $P(t_k|d_j)$. Finally, it calculates the probability $P(w_i|t_k)$ that a word w_i is present in a document belonging to a topic t_k, by aggregating the probabilities $P(t_k|d_j)$ for documents d_j that contain the word w_i.

By alternatingly updating the probabilities $P(w_i|t_k)$ and $P(t_k|d_j)$, the algorithm is able to assign each document to a distribution of topics, which in turn belong to a distribution of words. By inspecting the words belonging to each topic, we are able to identify the primary theme of each topic.

Human-wildlife conflict filtering. Investigating the topics generated by the LDA algorithm, we observe that Topics 3, 4 and 6 relate most closely to human-wildlife conflict based on manual inspection of the corresponding articles and the words assigned to these topics, which are shown in Table 1.

Table 1: LDA-generated topics with assigned words.

Topic	Assigned words (LDA)	Topic theme
1	"india", "people", ...	N/A
2	"police", "court", "case", ...	justice
3	**"government", "project", "land", ...**	**legislation/projects**
4	**"road", "railway", "building", ...**	**construction**
5	"bank", "power", "business", ...	finance/business
6	**"forest", "animal", "tiger", ...**	**wildlife**
7	"wildlife", "forest", "reserve", ...	wildlife (spurious)
8	"family", "woman", "child", ...	N/A
9	"party", "minister", "president", ...	politics
10	"student", "school", "college", ...	N/A

We utilize these three topics for further filtering by using the rank of each topic for each article with the filtering form

$$\mathcal{F}(d) = (rank_6 < k) \wedge (\alpha \times rank_4 + \beta \times rank_3 < m)$$

where $rank_i$ refers to the rank of the ith Topic in the article topic decomposition from the LDA model. The filtering equation shown selects an article d where (1) Topic 6 is in

[1] https://timesofindia.indiatimes.com

[2] https://github.com/skjorrface/animals.txt/blob/master/animals.txt

the k-highest ranked topics and (2) the weighted average of Topic 3 (weighted by α) and Topic 4 (weighted by β) is in the m-highest ranked topics.

After tuning the hyper-parameters $\{\alpha, \beta, k, m\}$, we observe that the configuration $\{\alpha, \beta, k, m\} = \{\frac{1}{2}, \frac{1}{2}, 2, 3\}$ works well based on manual inspection of the filtered articles.

This filtering reduces our dataset count to 2,277 articles.

3.3 Clustering

As we are working with an unsupervised dataset of newspaper articles, we use clustering to uncover latent features of the data which we then use to identify the primary modes of conflict in different Indian cities and for different wildlife species.

We experiment with clustering in three forms as:

1. Bag-of-Words + K-Means Clustering

   We create bag-of-word models from the filtered human-wildlife conflict articles and use K-means clustering to identify the primary modes of conflict.

2. Active Learning + K-Means Clustering

   We set up an active learning environment whereby the clustering system queries experts for annotation on representative samples.

3. BERT Classifier on K-Means Clustering

   We fine-tune a BERT-based classifier to predict the primary modes of conflict, using previously generated clusters as pseudo-labels.

Bag-of-Words + K-Means Clustering. We create a bag-of-words model as follows:

1. Pre-processing

   We remove punctuation and common/stop words from articles for easier processing.

2. Stemming

   We reduce words to their stems such that derived words map to their root (eg. “pollutant”, “pollute”, “pollution” all map to “pollut”).

3. Dictionary

   We create a dictionary of words from all processed articles.

4. Embedding

   We generate vector embeddings for each article based on word-presence using words from the generated dictionary from the previous step.

We then apply K-Means clustering with cluster counts $k = \{4, 6, 8, 10, 12, 14, 16\}$.

The recovered conflict causes for $k = 10$ are shown in Table 2.

We observe that using clustering, our system is able to recover a number of the major causes of human-wildlife conflict.

However, notice that the cluster formations display interesting properties. Firstly, observe that Cluster 8 is not assigned a singular semantically meaningful topic as judged by manual inspection of the clustered articles. Furthermore, observe that the system chooses to divide the Pollution topic into two separate clusters, 4 and 5, as Pollution (water, industrial) and Pollution (air). A similar phenomenon occurs with wildlife and wildlife conservation (Clusters 1 and 3).

Table 2: K-Means cluster topics for $k = 10$.

Cluster	Topic
1	wildlife
2	construction
3	wildlife conservation
4	power
5	pollution (water, industrial)
6	pollution (air)
7	disease
8	N/A
9	environmental conservation
10	farming

Active Learning + K-Means Clustering. Based on the properties of the K-Means clustering explained previously, we use active learning to convert the original unsupervised learning problem to a semi-supervised one by introducing expert-annotated pseudo-labels.

Active learning refers to a family of learning models that are allowed to query experts for ground truth annotations on a limited number of representative samples.

Having received expert annotations for a subset of the samples, the clustering algorithm then incorporates the received information as pseudo-labels into the system which results in better clustering performance.

As our dataset is unsupervised, we use an active learning framework where the system queries experts not for labels but on whether or not two samples should be in the same cluster.

The recovered conflict causes using active learning for $k = 10$ are shown in Table 3.

Table 3: Active learning cluster topics for $k = 10$.

Cluster	Topic
1	environmental conservation
2	wildlife/wildlife conservation
3	pollution
4	power
5	farming
6	disease
7	construction
8	natural disaster
9	water shortage
10	proximity to human settlements

We observe that taking an active learning approach resolves a lot of the problems posed by the regular K-Means clustering as all recovered clusters are now semantically meaningful where the spurious bifurcations from the previous method have been successfully removed. Notice also that using active learning, our system was also able to recover the additional causes “natural disaster” and “proximity to human

Table 4: BERT classifier performance on predicting major cause of human wildlife conflicts from news articles

Prediction Performance	Classification Accuracy
Training	92.2%
Validation	90.6%

settlements".

BERT Classifier on K-Means Clustering. Given that active learning provides reasonable labelling of the dataset for causes of HWC, to make our model more generalizable and scalable, we built a classifier that identifies the main cause of HWC from each article.

For the word embeddings, we use the Bidirectional Encoder Representations from Transformers (BERT) model from [Devlin *et al.*, 2018]. Previous studies show that with additional fine-tuning BERT can achieve state-of-the-art performances on various language tasks. For our model, we use the pre-trained BERT encoder to obtain a 768 dimensional vector based on the first 256 words of each article. We then feed the embedding to a multi-layer perceptron with fully connected layers and hyperbolic tangent activation function that compresses the vector into a 10-dimensional probability distribution output.

We train the classifier on a subset of the whole dataset by selecting articles close to the cluster centers and treat the cluster assignments as the ground truth label. This filtering generates about 1200 training samples and ensures that we do not use articles with high uncertainty in clustering for the predictive model. Our classifier performance is shown in Table 4.

The added benefit of the BERT classifier over our clustering mechanism is two-fold: Firstly, using BERT, we obtain a classifier that can be easily adapted to virtually any other end-to-end training pipeline on human-wildlife conflict analysis for fine-tuning. Secondly, by encapsulating our pipeline in a black-box neural network, we obtain a stand-alone model for conflict mode prediction that is more easily generalizable to conflict analysis in different domains and locations.

4 Results

4.1 Exploratory visualizations

For our exploratory analysis, we generated visualizations to identify geological hotpots. We visualized the following metrics:

- Number of articles that mentioned the top animals keywords.
- Ratio of the filtered articles that also mentioned conflict keywords.

A histogram of number of articles broken down by in which cities these articles are reported shows that the number of articles generally correlated with size of the city. In Figure 1, cities such as New Delhi and Mumbai have the largest number of mentions, and each have 27.15 million and 12.48 million population. Therefore, that fact that these cities have a large number of mentions does not indicate that animal human conflicts are more frequent in these cities, as number of news articles correlates with the scale of cities. We subsequently plotted ratio of articles filtered that mentioned human animal conflict, broken down by cities (Figure 3a). We identified cities with the highest ratios of mentions on human-wild life conflict to be Chandrapur, Bhopal and Lucknow. These cities all have natural parks or mingle well with natural habitats as in Figure 1.

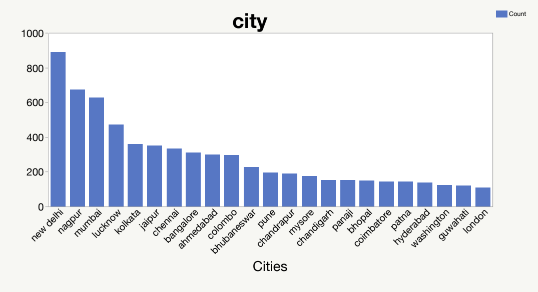
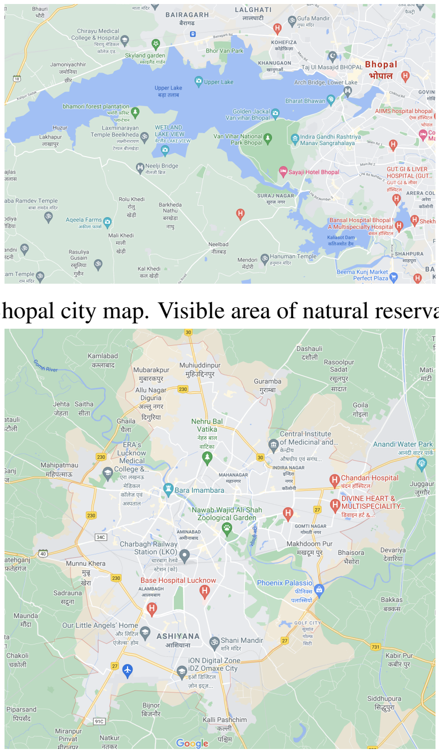


Figure 1: The number of articles that mentioned human-wildlife conflicts.

We also visualized the number of cities that mentioned human-wildlife conflicts with respect to time. The visualization shows that there were increasing amount of cities that mentioned human-wildlife conflict as shown in Figure 4a which indicates an upwards temporal trend.

In addition, through our visualizations, we looked at specific species and attempted to assess the threats on these species. For the animals that we looked at, both the mentions of tigers and elephants showed an upward trend as in Figure 4b and Figure 4c, similar to the overall mentions. The first example is tigers. Through the same analysis, the top Indian cities that mentioned tigers are Bhopal, Nagpur and Chandrapur. These cities, according to a search on the map, contain natural conservation or zoos for tigers. For example, Nagpur hosts the regional office of National Tiger Conservation Authority [Agencies, 2011]. The city is also famed as the "capital of tigers" [Correspondent, 2011]. Various news articles that mentioned conflict with tigers, or tigers entering human residential areas [Naveen, 2018] [Noronha, 2020], [PTI, 2020] occurred in Bhopal and Chandrapur. The second example is elephants. With the same process, we found cities that are potential hot-spots of human-elephant conflicts 4b. News in Coimbatore indicated elephants' unusual death due to human activities [Kaveri and Jayarajan, 2020], encroaching human habitat that collides with elephants' [Correspondence, 2021] and the conflicts due to shrinking habitats [Thomas, 2020]. Chennai is the second city that has a high number of mentions of human-elephant conflict because of tourism. Mentions of conflicts largely come from tourism [Madhav, 2021], illegal trade or ownership [Sureshkumar, 2021] and hurting elephants intentionally [Bureau, 2021].

Our analysis pinpointed potential cities where human-wildlife conflicts occur as well as the most prevalent modes of this conflict. Other users of this dataset can potentially use similar visualizations and techniques to scale down the search of articles that pertain to the animal and type of conflict that they want to survey.



(a) Bhopal city map. Visible area of natural reservation.

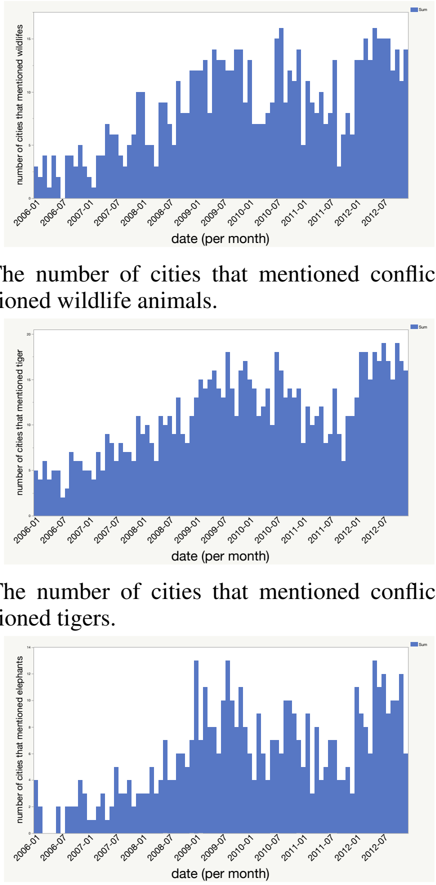
(b) Lucknow city map.

Figure 2

5 Broader Impact

Our work is motivated by the reported increase in the frequency and environmental effect of human-wildlife conflict in India [Gulati *et al.*, 2021] and presents the first large-scale news-based system for human-wildlife conflict analysis in India. We adapt the extensive efforts by [Chang *et al.*, 2016], [Ming *et al.*, 2014], [Zhang and Skiena, 2010] made in financial analysis and market modeling in using news articles as a proxy for statistical data and human sentiment to identify the primary modes of human-wildlife conflict in India, which circumvents the problem of data scarcity commonly observed in AI for Social Impact projects.

We posit that the potential impact our work is two-fold: Firstly, our system presents an interpretable method for constructing an aggregate representation of news articles relating to a certain topic, which can then be used effectively to identify relevant spatial and temporal trends. As (1) news articles are a ubiquitous source of information that can be collected for virtually any topic reflected in media and (2) our method makes no implicit assumption regarding the thematic nature of the collected data, we expect our system to be easily adaptable for different social problems which can act either as a primary or auxiliary mode of analysis for a wide variety of AI for Social Impact projects. Additionally, we believe that our method is also capable of accommodating different types of news data simultaneously such as vernacular press or social media coverage which might be helpful to uncover better local structure for human-wildlife conflict in an area.



(a) The number of cities that mentioned conflict and mentioned wildlife animals.

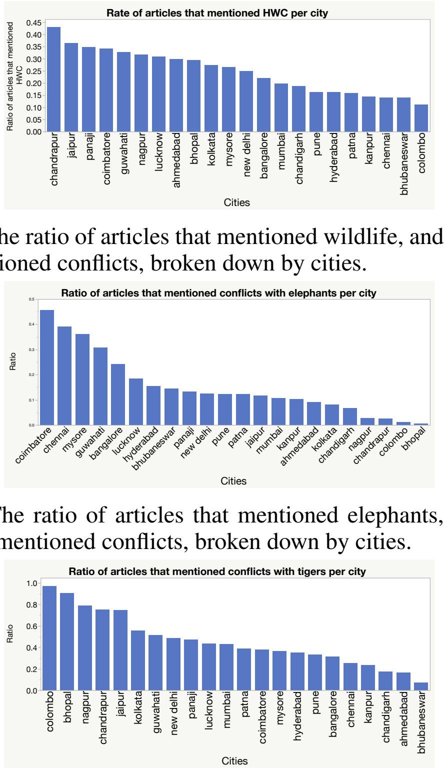
(b) The number of cities that mentioned conflict and mentioned tigers.

(c) The number of cities that mentioned conflict and mentioned elephants.

Figure 3

Secondly, our clustering model is able to not only identify the primary modes of human-wildlife conflict in India but also simultaneously reveal the temporal and spatial patterns of conflict throughout the country. This allows us to identify the hotspots of human-wildlife conflict as well as the day-to-day trends indexed on cities or animals. We believe that this provides valuable insight for policy makers who would benefit from taking into consideration the local structure of human-wildlife conflict in their domain. As a secondary use case, we also believe that our system can be used to create awareness among the local population whereby people could access all previous records of human-wildlife conflict in their area through a simple interactive web portal.

In publishing this work, we are aware of the potential misuses of our system. Primarily, we believe that the interpretable nature of our method makes our results susceptible to accidental or deliberate misinterpretation, the first of which we actively guard ourselves against by making our results clear and explicit and the second of which we ask against by imploring proper decorum from all policy makers. We argue that a significant misuse of our work would be to undermine the environmental impact of certain policies by citing our results that similar policies have not been ranked highly in simi-



(a) The ratio of articles that mentioned wildlife, and also mentioned conflicts, broken down by cities.

(b) The ratio of articles that mentioned elephants, and also mentioned conflicts, broken down by cities.

(c) The ratio of all articles that mentioned tigers, and also mentioned conflicts, broken down by cities.

Figure 4

lar contexts to incite human-wildlife conflict. For example, in our specific case study, elephant conflicts in Coimbatore was frequent and severe. Tiger conflicts in the same city were not frequently mentioned in the news outlet. This result does not justify the development in the area that can potentially hurt the livelihood of tigers. We state firmly that our results are not meant to replace the diligent legal work that is required of all environmental policies, but to provide a simple and accessible means to uncover human-wildlife conflict trends that can be used to better inform the policy makers.

6 Conclusion

We propose a method for utilizing news articles to identify the major causes of human-wildlife conflict in India. Our system uses a dataset consisting of 758,000 news articles collected from the Times of India website between 2006 and 2018 as a proxy to real-time human-wildlife conflict data that is difficult to collect for a long period of time, which we then process to identify and localize major conflict causes and hot-spots for different cities and animals. We first pre-process our dataset using a four-stage filtering pipeline to identify articles relating to human-wildlife conflict, which are then clustered to identify the primary modes of human-wildlife conflict in India. With the temporal visualizations, we observed that HWC is on the rise in the recent years. With the clustered keywords, we were able to identify specific cities where HWC occured frequently. We verified our methods with specific wild animals, and found that the hotspots identified did have frequent HWC through discoveries on local news. We hope our methods and findings can inform public of HWC hostspots and help future policy making. Finally, we believe our method is extendable to different news datasets to uncover hotspots and patterns in different conflict domains.

References

[Agencies, 2011] Agencies. Govt approves ig posts for ntca hq, regional offices: Nagpur news - times of india, Apr 2011.

[Bain *et al.*, 2019] Max Bain, Arsha Nagrani, Daniel Schofield, and Andrew Zisserman. Count, crop and recognise: Fine-grained recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[Bureau, 2021] FPJ Bureau. Police seal tamil nadu resort where an elephant was set on fire, Jan 2021.

[Chang *et al.*, 2016] Ching Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. Measuring the information content of financial news. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225, 2016.

[Correspondence, 2021] Correspondence. Wild elephants terrorise residents of sirumugai: Coimbatore news - times of india, Apr 2021.

[Correspondent, 2011] dna Correspondent. Nagpur to be country's tiger capital, Apr 2011.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Di Minin *et al.*, 2019] Enrico Di Minin, Christoph Fink, Tuomo Hiippala, and Henrikki Tenkanen. A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, 33(1):210, 2019.

[Gulati *et al.*, 2021] Sumeet Gulati, Krithi K Karanth, Nguyet Anh Le, and Frederik Noack. Human casualties are the dominant cost of human–wildlife conflict in india. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

[Hagenau *et al.*, 2013] Michael Hagenau, Michael Liebmann, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697, 2013.

[Kaveri and Jayarajan, 2020] Megha Kaveri and Sreedevi Jayarajan. 14 elephants die in coimbatore forests in 2020: Experts point to several factors, Jul 2020.

[Madhav, 2021] Pramod Madhav. Heartbreaking video shows chained elephant struggling to walk, forced to limp in tamil nadu, Jan 2021.

[Manral *et al.*, 2016] Upma Manral, Shruti Sengupta, Syed Ainul Hussain, Sakshi Rana, and Ruchi Badola. Human wildlife conflict in india: A review of economic implication of loss and preventive measures. *Indian Forester*, 142(10):928–940, 2016.

[Margulies and Karanth, 2018] Jared D Margulies and Krithi K Karanth. The production of human-wildlife conflict: A political animal geography of encounter. *Geoforum*, 95:153–164, 2018.

[Ming *et al.*, 2014] Felix Ming, Fai Wong, Zhenming Liu, and Mung Chiang. Stock market prediction from wsj: text mining via sparse matrix factorization. In *2014 IEEE International Conference on Data Mining*, pages 430–439. IEEE, 2014.

[Naveen, 2018] P Naveen. Madhya pradesh's tiger released in odisha's satkosia reserve: Bhopal news - times of india, Jun 2018.

[Noronha, 2020] Rahul Noronha. Saving bhopal's famed urban tigers, Aug 2020.

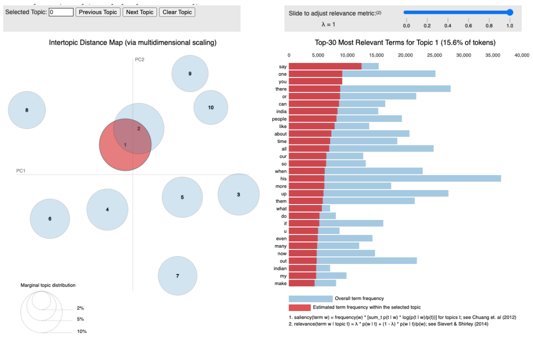
[PTI, 2020] PTI. 'tiger state' madhya pradesh lost 290 big cats in 19 years: Official: Bhopal news - times of india, Aug 2020.

[Sureshkumar, 2021] Sureshkumar. Consider prohibition of private ownership of elephants by temples and individuals, madras high court suggests: Chennai news - times of india, Feb 2021.

[Thomas, 2020] Wilson Thomas. Elephant tramples woman to death in coimbatore dist., Nov 2020.

[Zhang and Skiena, 2010] Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

A LDA outputs


Selected Topic: 0 Previous Topic Next Topic Clear Topic

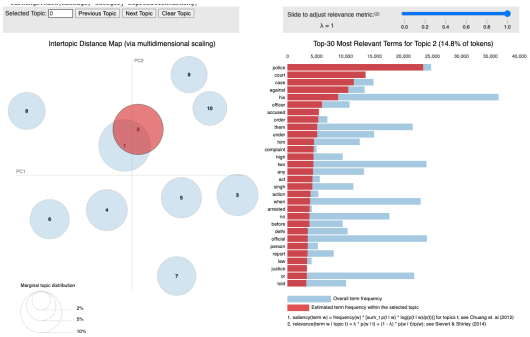
Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 1 (15.6% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0 Previous Topic Next Topic Clear Topic

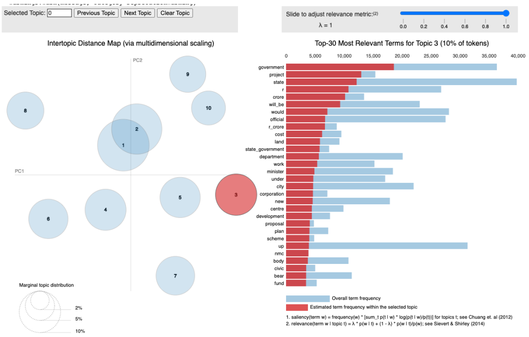
Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 2 (14.8% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0 Previous Topic Next Topic Clear Topic

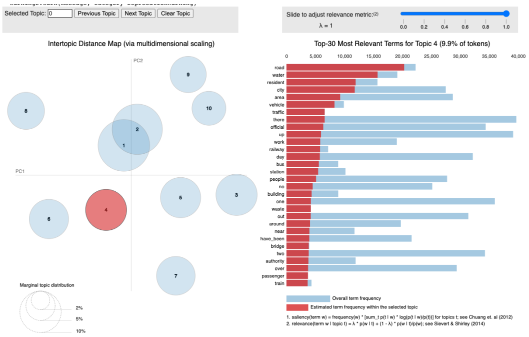
Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 3 (10% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0 Previous Topic Next Topic Clear Topic

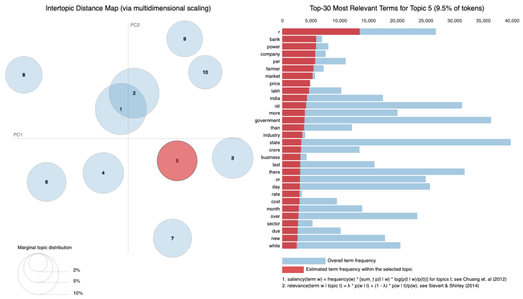
Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 4 (9.9% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

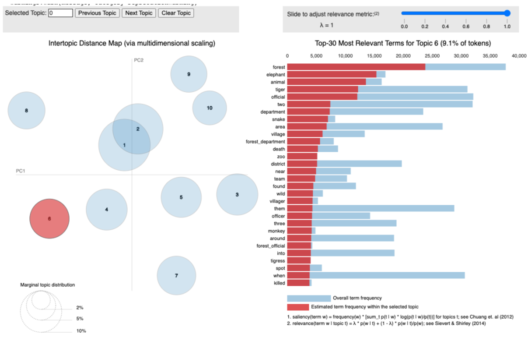
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 5 (9.5% of tokens)

Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 0 Previous Topic Next Topic Clear Topic

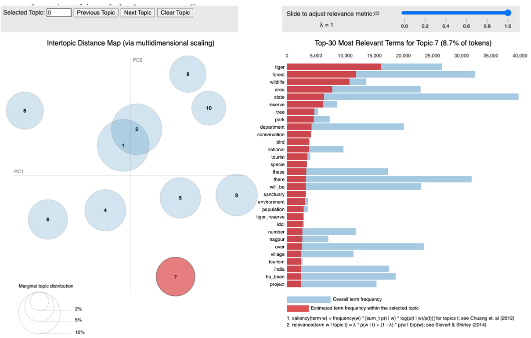
Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (9.1% of tokens)

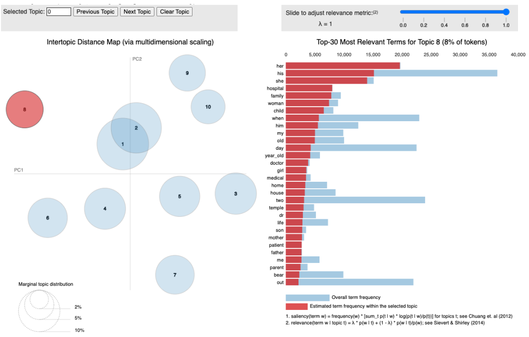
Overall term frequency
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



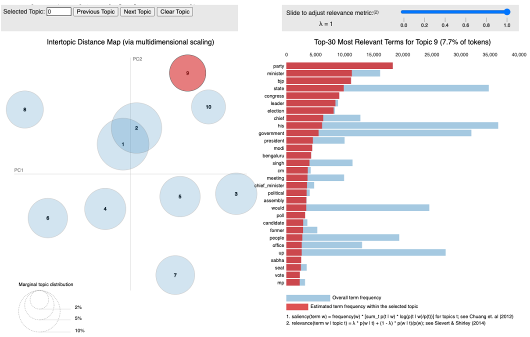
Top-30 Most Relevant Terms for Topic 7 (8.7% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



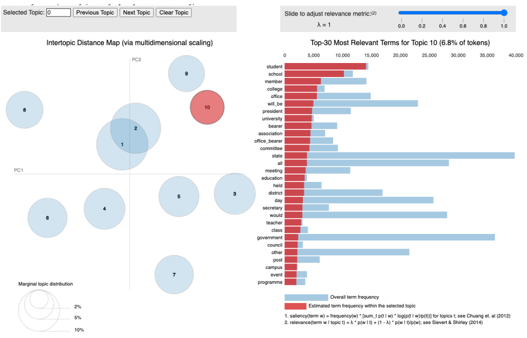
Top-30 Most Relevant Terms for Topic 8 (8% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



Top-30 Most Relevant Terms for Topic 9 (7.7% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



Top-30 Most Relevant Terms for Topic 10 (6.8% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)