Matching Land Conflict Events to Government Policies via Machine Learning Models

In partnership with

WORLD RESOURCES INSTITUTE

Laura Clark Murray, Nikhel Gupta, Joanne Burke, Rishika Rupam, Zaheeda Tshankie

Project Overview

This project aimed to provide a proof-of-concept machine-learning-based methodology to identify land conflict events in a geography and match those events to relevant government policies. The overall objective is to offer a platform where policymakers can be made aware of land conflicts as they unfold and identify existing policies that are relevant to the resolution of those conflicts.

Several Natural Language Processing (NLP) models were built to identify and categorize land conflict events in news articles and to match those land conflict events to relevant policies. A web-based tool that houses the models allows users to explore land conflict events spatially and through time, as well as explore all land conflict events by category across geography and time.

The geographic scope of the project was limited to India, which has the most environmental conflicts of all countries on Earth.

Background

Degraded land is “land that has lost some degree of its productivity due to human-caused process”, according to the World Resources Institute. Land degradation affects 3.2 billion people and costs the global economy about 10 percent of the gross product each year. While dozens of countries have committed to restore 350 million hectares of degraded land, land disputes are a major barrier to effective implementation. Without streamlined access to land use rights, landowners are not able to implement sustainable land-use practices. In India, where 21 million hectares of land have been committed to the restoration, land conflicts affect more than 3 million people each year.

AI and machine learning offer tremendous potential to not only identify land-use conflicts events but also match suitable policies for their resolution.

Data Collection

All data used in this project is in the public domain.

News Article Corpus: Contained 65,000 candidate news articles from Indian and international newspapers from the years 2008, 2017, and 2018. The articles were obtained from the Global Database of Events Language and Tone Project (GDELT), “a platform that monitors the world’s news media from nearly every corner of every country in print, broadcast, and web formats, in over 100 languages.” All the text was either originally in English or translated to English by GDELT.

Annotated Corpus: Approximately 1,600 news articles from the full News Article Corpus were manually labeled and double-checked as Negative (no conflict news) and Positive (conflict news).

Gold Standard Corpus: An additional 200 annotated positive conflict news articles, provided by WRI.

Policy Database: Collection of 19 public policy documents related to land conflicts, provided by WRI.

Approach

Text Preparation

In this phase, the articles of the News Article Corpus and policy documents of the Policy Database were prepared for the natural language processing models.

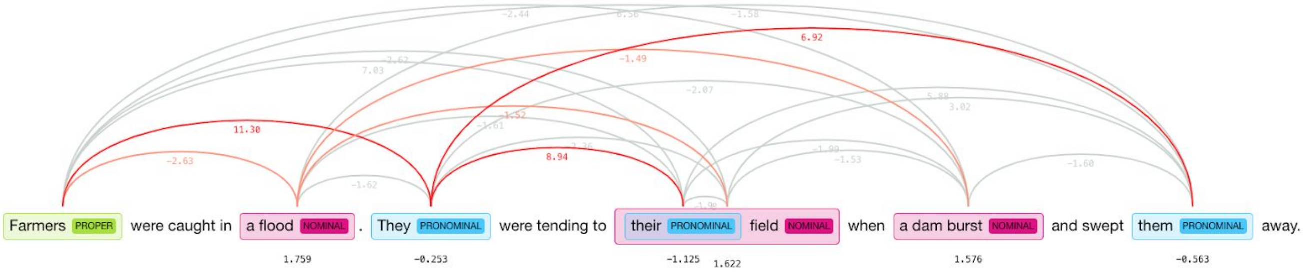
The articles and policy documents were processed using SpaCy, an open-source library for natural language processing, to achieve the following:

- Tokenization: Segmenting text into words, punctuation marks, and other elements.
- Part-of-speech (POS) tagging: Assigning word types to tokens, such as “verb” or “noun”
- Dependency parsing: Assigning syntactic dependency labels to describe the relations between individual tokens, such as “subject” or “object”
- Lemmatization: Assigning the base forms of words, regardless of tense or plurality
- Sentence Boundary Detection (SBD): Finding and segmenting individual sentences.
- Named Entity Recognition (NER): Labelling named “real-world” objects, like persons, companies, or locations.

Coreference resolution was applied to the processed text data using Neuralcoref, which is based on an underlying neural net scoring model. With coreference resolution, all common expressions that refer to the same entity were located within the text. All pronominal words in the text, such as her, she, he, his, them, their, and us, were replaced with the nouns to which they referred.

For example, consider this sample text:
“Farmers were caught in a flood. They were tending to their field when a dam burst and swept them away.”

Neuralcoref recognizes “Farmers”, “they”, “their” and “them” as referring to the same entity. The processed sentence becomes:
“**Farmers** were caught in a flood. **Farmers** were tending to their field when a dam burst and swept **farmers** away.”



Coreference resolution of sample sentences

Document Classification

The objective of this phase was to build a model to categorize the articles in the News Article Corpus as either "Negative", meaning they were not about conflict events, or "Positive", meaning they were about conflict events.

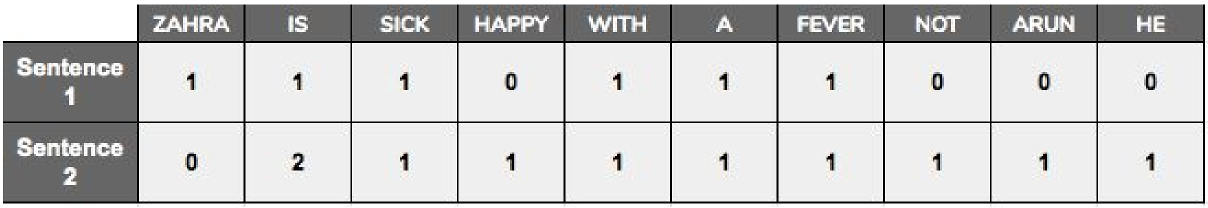
After preparation of the articles in the News Article Corpus, as described in the previous section, the texts were then prepared for classification.

First, an Annotated Corpus was formed to train the classification model. A 1,600 article subset of the News Article Corpus was manually labeled as "Negative" or "Positive".

To prepare the articles in both the News Article Corpus and Annotated Corpus for classification, the previously pre-processed text data of the articles was represented as vectors using the Bag of Words approach. With this approach, text is represented as a collection, or "bag", of the words it contains along with the frequency with which each word appears. The order of words is ignored.

For example, consider a text article comprised of these two sentences:
Sentence 1: "Zahra is sick with a fever."
Sentence 2: "Arun is happy he is not sick with a fever."

This text contains a total of ten words: "Zahra", "is", "sick", "happy", "with", "a", "fever", "not", "Arun", "he". Each sentence in the text is represented as a vector, where each index in the vector indicates the frequency that one particular word appears in that sentence, as illustrated below.



	ZAHRA	IS	SICK	HAPPY	WITH	A	FEVER	NOT	ARUN	HE
Sentence 1	1	1	1	0	1	1	1	0	0	0
Sentence 2	0	2	1	1	1	1	1	1	1	1

With this technique, each sentence is represented by a vector, as follows:

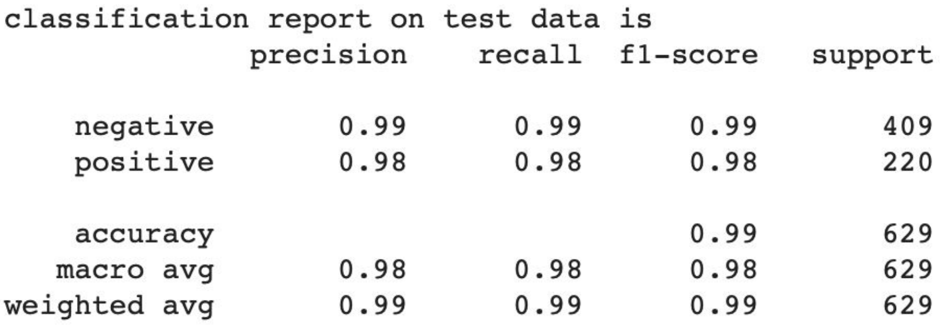
"Zahra is sick with a fever." ⟶ [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
"Arun is happy he is not sick with a fever." ⟶ [0, 2, 1, 1, 1, 1, 1, 1, 1, 1]

With the Annotated Corpus vectorized with this technique, the data was used to train a logistic regression classifier model. The trained model was then used with the vectorized data of the News Article Corpus, to classify each article into Positive and Negative conflict categories.

The accuracy of the classification model was measured by looking at the percentage of the following:
True Positive: Articles correctly classified as relating to land conflict
False Positive: Articles incorrectly classified as relating to land conflict
True Negative: Articles correctly classified as not being related to land conflict
False Negative: Articles incorrectly classified as not being related to land conflict

The “precision” of the model indicates how many of those articles classified to be about land conflict were actually about land conflict. The “recall” of the model indicates how many of the articles that were actually about land conflict were categorized correctly. An f1-score was calculated from the precision and recall scores.

The trained logistic regression model successfully classified the news articles with a precision, recall and f1-score of 98% or greater. This indicates that produced a low number of false positives and false negatives.



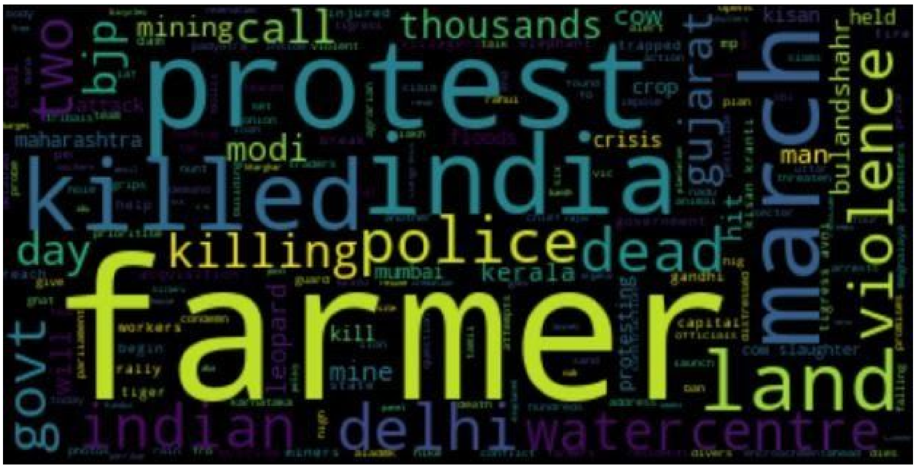
```
classification report on test data is
              precision    recall  f1-score   support

    negative       0.99      0.99      0.99       409
    positive       0.98      0.98      0.98       220

    accuracy                           0.99       629
   macro avg       0.98      0.98      0.98       629
weighted avg       0.99      0.99      0.99       629
```

Classification report using a test dataset and logistic regression model

Categorize by Conflict Events

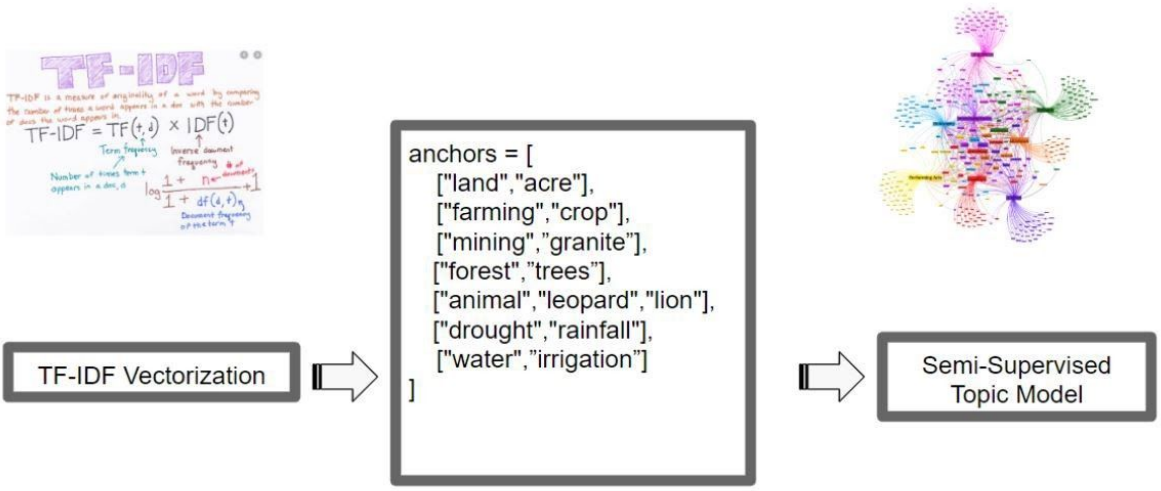
The objective of this phase was to build a model to identify the set of conflict events referred to in the collection of positive conflict articles and then to classify each positive conflict article accordingly.

A word cloud of the articles in the Gold Standard Corpus gives a sense of the content covered in the articles.



A topic model was built to discover the set of conflict topics that occur in the Positive conflict articles. We chose a semi-supervised approach to topic modeling to maximize the accuracy of the classification process. We chose to use CorEx (Correlation Explanation), a semi-supervised topic model that allows domain knowledge, as specified by relevant keywords acting as “anchors”, to guide the topic analysis.

To align with the Land Conflict Policies provided by WRI, seven relevant core land conflict topics were specified. For each topic, correlated keywords were specified as “anchors” for the topic.



TF-IDF Vectorization

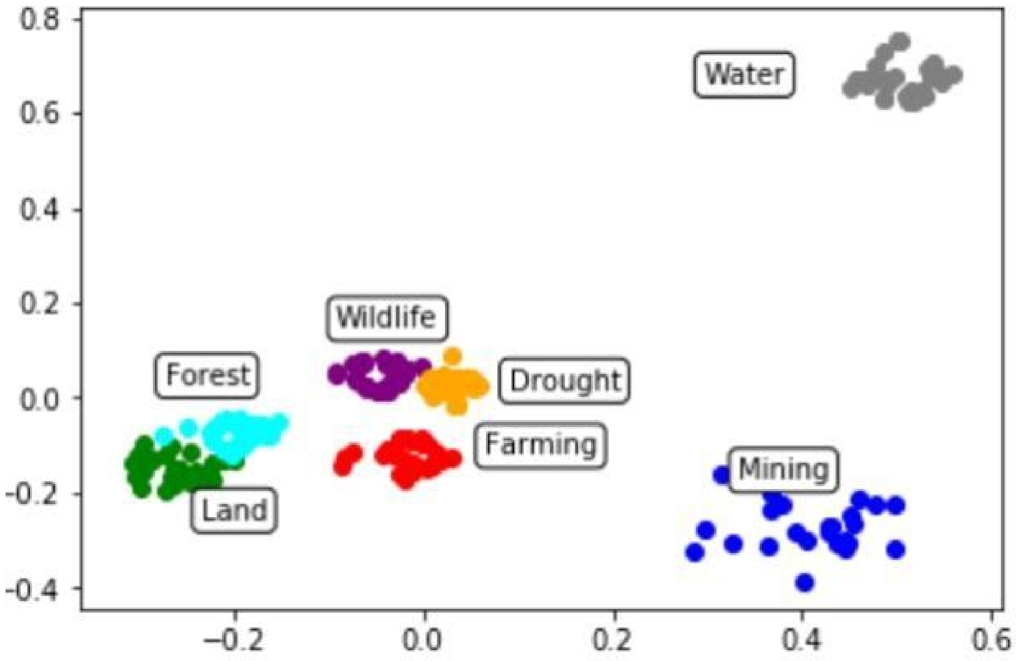
```
anchors = [
    ["land","acre"],
    ["farming","crop"],
    ["mining","granite"],
   ["forest","trees"],
   ["animal","leopard","lion"],
   ["drought","rainfall"],
    ["water","irrigation"]
]
```

Semi-Supervised Topic Model

The trained topic model provided 3 words for each of the seven topics:

Topic #1: land, resettlement, degradation
Topic #2: crops, farm, agriculture
Topic #3: mining, coal, sand
Topic #4: forest, trees, deforestation
Topic #5: animal, attacked, tiger
Topic #6: drought, climate change, rain
Topic #7: water, drinking, dams

The resulting topic model is 93% accurate. This scatter plot uses word representations to provide a visualization of the model’s classification of the Gold Standard Corpus and hand-labeled positive conflict articles.



Visualization of the topic classification of the Gold Standard Corpus and Positive Conflict Articles

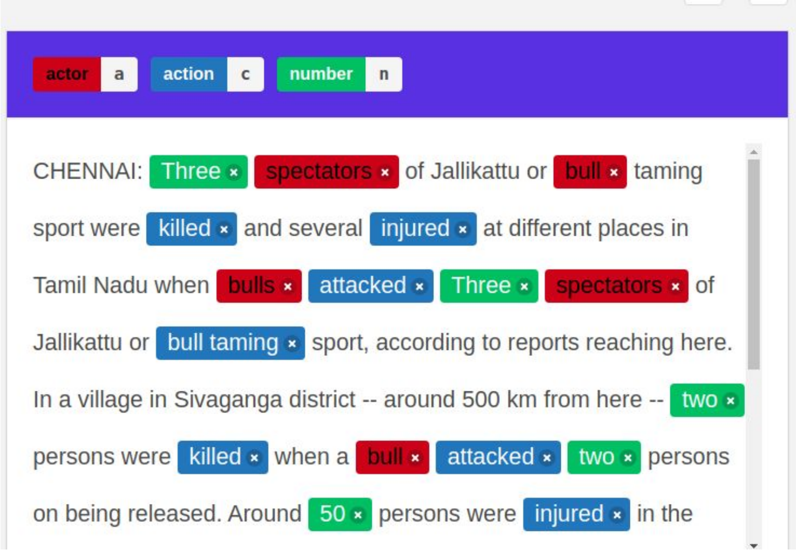
Identify the Actors, Actions, Scale, Locations, and Dates

The objective of this phase was to build a model to identify the actors, actions, scale, locations, and dates in each positive conflict article.

Typically, names, places, and famous landmarks are identified through Named Entity Recognition (NER). Recognition of such standard entities is built-in with SpaCy's NER package, by which our model detected the locations and dates in the positive conflict articles. The specialized content of the news articles required further training with "custom entities" — those particular to this context of land conlficts.

All the positive conflict articles in the Annotated Corpus were manually labeled for "custom entities":
Actors: Such as "Government", "Farmer", "Police", "Rains", "Lion"
Actions: Such as "protest", "attack", "killed"
Numbers: Number of people affected by a conflict

This example shows how this labeling looks for some text in one article:



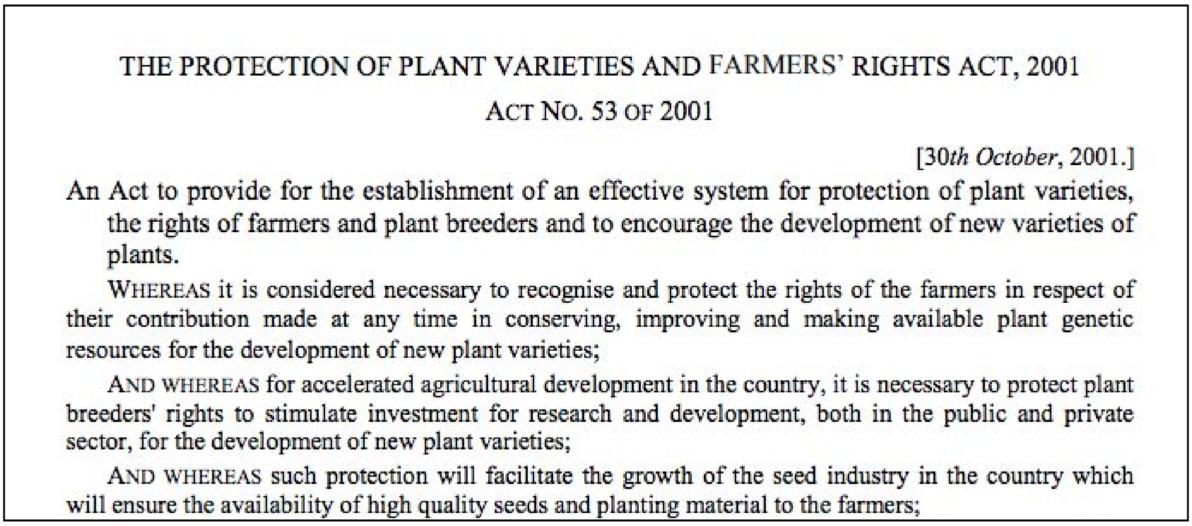
CHENNAI: Three spectators of Jallikattu or bull taming sport were killed and several injured at different places in Tamil Nadu when bulls attacked Three spectators of Jallikattu or bull taming sport, according to reports reaching here. In a village in Sivaganga district -- around 500 km from here -- two persons were killed when a bull attacked two persons on being released. Around 50 persons were injured in the

These labeled positive conflict articles were used to train our custom entity recognizer model. That model was then used to find and label the custom entities in the news articles in the News Article Corpus.

Match Conflicts to Relevant Policies

The objective of this phase was to build a model to match each processed positive conflict article to any relevant policies.

The Policy Database was composed of 19 policy documents relevant to land conflicts in India, including policies such as the “Land Acquisition Act of 2013”, the “Indian Forest Act of 1927”, and the “Protection of Plant Varieties and Farmers’ Rights Act of 2001”.



THE PROTECTION OF PLANT VARIETIES AND FARMERS’ RIGHTS ACT, 2001

ACT NO. 53 OF 2001

[30*th October*, 2001.]

An Act to provide for the establishment of an effective system for protection of plant varieties, the rights of farmers and plant breeders and to encourage the development of new varieties of plants.

WHEREAS it is considered necessary to recognise and protect the rights of the farmers in respect of their contribution made at any time in conserving, improving and making available plant genetic resources for the development of new plant varieties;

AND WHEREAS for accelerated agricultural development in the country, it is necessary to protect plant breeders' rights to stimulate investment for research and development, both in the public and private sector, for the development of new plant varieties;

AND WHEREAS such protection will facilitate the growth of the seed industry in the country which will ensure the availability of high quality seeds and planting material to the farmers;

Excerpt of a 2001 policy document related to agriculture

A text similarity model was built to compare two text documents and determine how close they are in terms of context or meaning. The model made use of the “Cosine similarity” metric to measure the similarity of two documents irrespective of their size.

Cosine similarity calculates similarity by measuring the cosine of an angle between two vectors. Using the vectorized text of the articles and the policy documents that had been generated in the previous phases as described above, the model generated a collection of matches between articles and policies.

Visualization of Conflict Event and Policy Matching

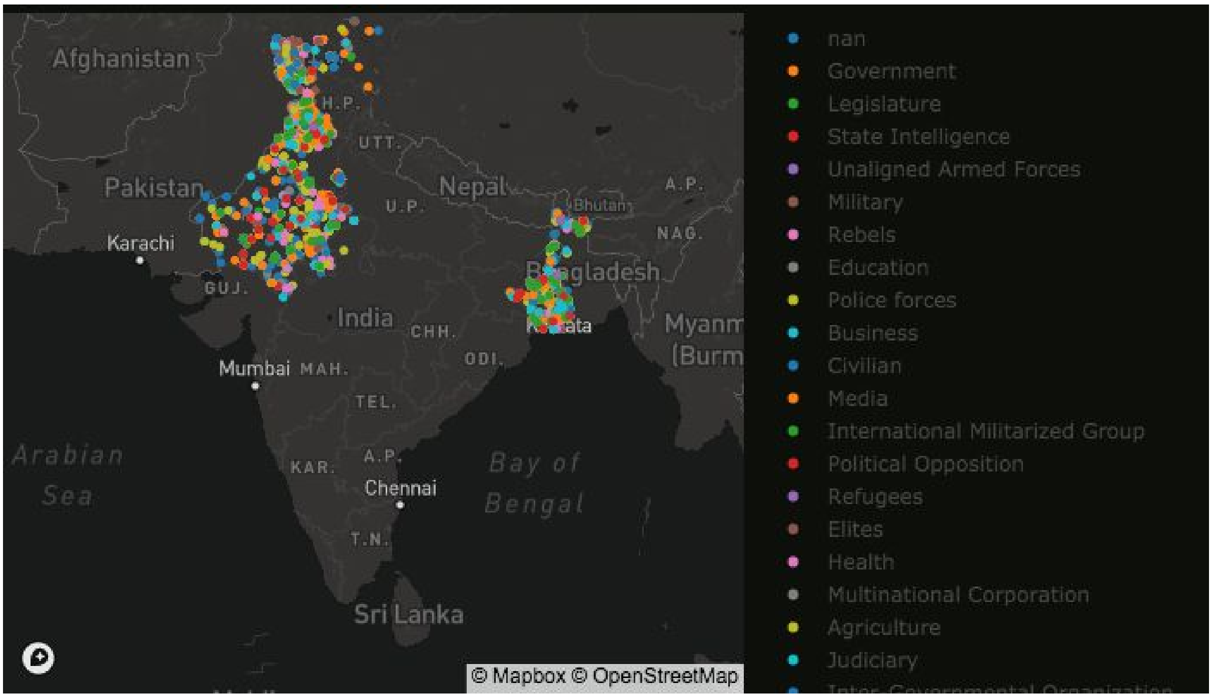
The objective of this phase was to build a web-based tool for the visualization of the conflict event and policy matches.

An application was created using the Plotly Python Open Source Graphing Library. The web-based tool houses the models and allows users to explore land conflict events spatially and through time, as well as explore all land conflict events by category across geography and time.

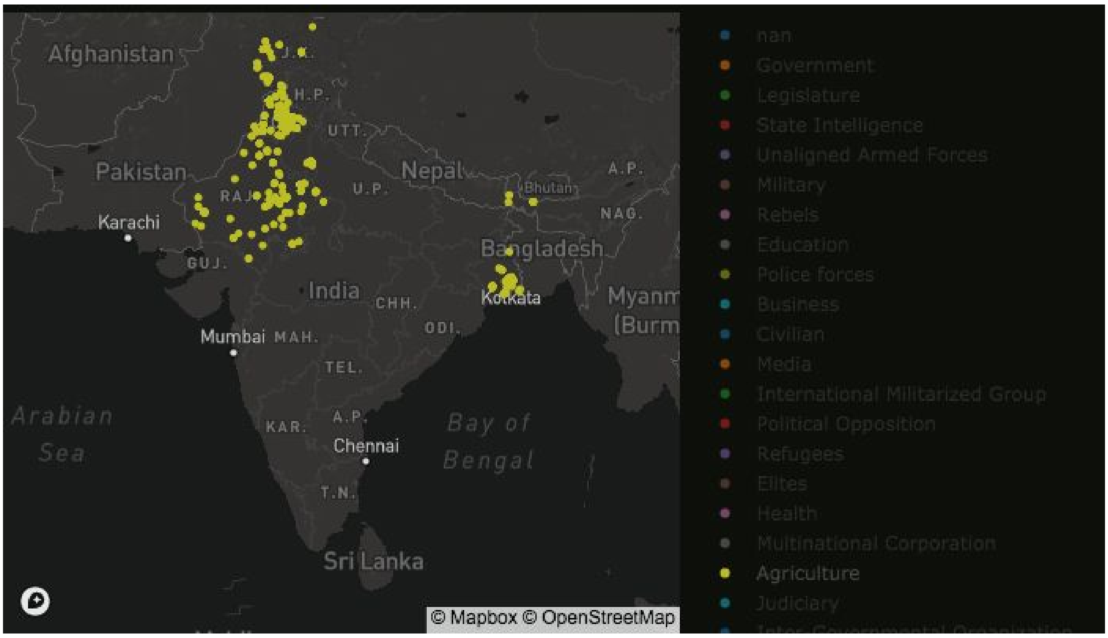
The map displays land conflict events detected in the News Article Corpus for the selected years and regions of India.

Conflict events are displayed as color-coded dots on a map. The colors correspond to specific conflict categories, such as “Agriculture” and “ Environmental”, and actors, such as “Government”, “Rebels”, and “Civilian”.

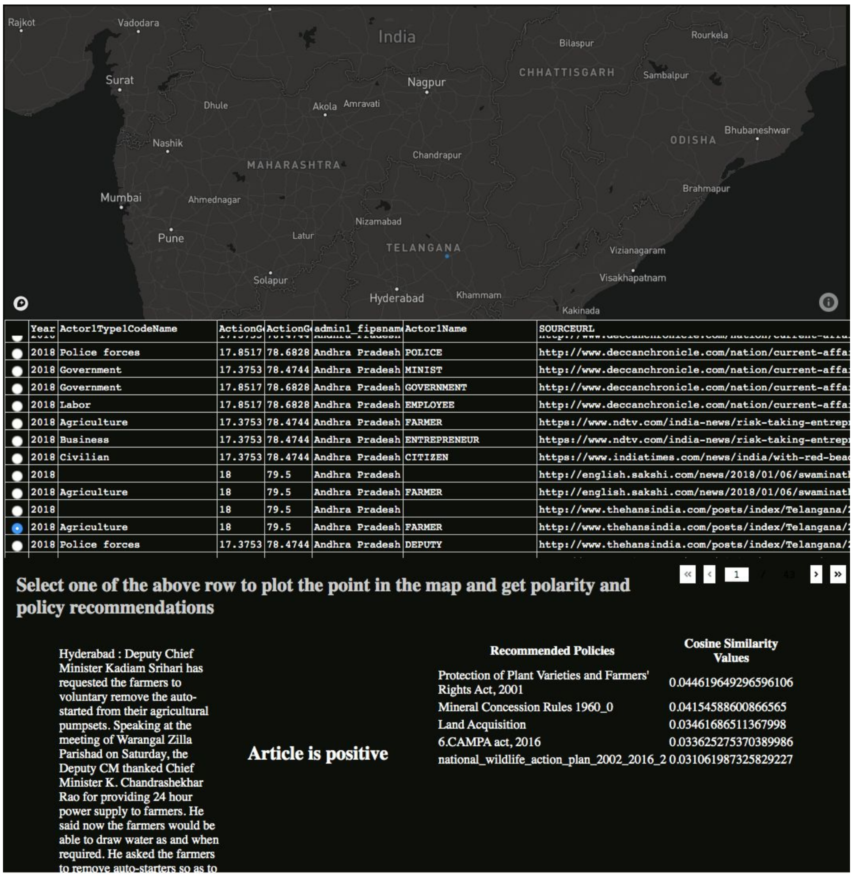
In this example, the tool displays geo-located land conflict events across five regions of India in 2017 and 2018.



By selecting a particular category from the right column, only those conflicts related to that category are displayed on the map. Here only the Agriculture-related subset of the events shown in the previous example are displayed.



News articles from the select years and regions are displayed below the map. When a particular article is selected, the location of the event is shown on the map. The text of the article is displayed along with policies matched to the event by the underlying models, as seen in the example below of a 2018 agriculture-related conflict in the Andhra Pradesh region.



	Year	Actor1Type1CodeName	ActionG	ActionG	admin1_fipsnam	Actor1Name	SOURCEURL
○	2018	Police forces	17.8517	78.6828	Andhra Pradesh	POLICE	http://www.deccanchronicle.com/nation/current-affa
○	2018	Government	17.3753	78.4744	Andhra Pradesh	MINIST	http://www.deccanchronicle.com/nation/current-affa
○	2018	Government	17.8517	78.6828	Andhra Pradesh	GOVERNMENT	http://www.deccanchronicle.com/nation/current-affa
○	2018	Labor	17.8517	78.6828	Andhra Pradesh	EMPLOYEE	http://www.deccanchronicle.com/nation/current-affa
○	2018	Agriculture	17.3753	78.4744	Andhra Pradesh	FARMER	https://www.ndtv.com/india-news/risk-taking-entrepr
○	2018	Business	17.3753	78.4744	Andhra Pradesh	ENTREPRENEUR	https://www.ndtv.com/india-news/risk-taking-entrepr
○	2018	Civilian	17.3753	78.4744	Andhra Pradesh	CITIZEN	https://www.indiatimes.com/news/india/with-red-bea
○	2018		18	79.5	Andhra Pradesh		http://english.sakshi.com/news/2018/01/06/swaminat
○	2018	Agriculture	18	79.5	Andhra Pradesh	FARMER	http://english.sakshi.com/news/2018/01/06/swaminat
○	2018		18	79.5	Andhra Pradesh		http://www.thehansindia.com/posts/index/Telangana/
●	2018	Agriculture	18	79.5	Andhra Pradesh	FARMER	http://www.thehansindia.com/posts/index/Telangana/
○	2018	Police forces	17.3753	78.4744	Andhra Pradesh	DEPUTY	http://www.thehansindia.com/posts/index/Telangana/

« ‹ 1 › »

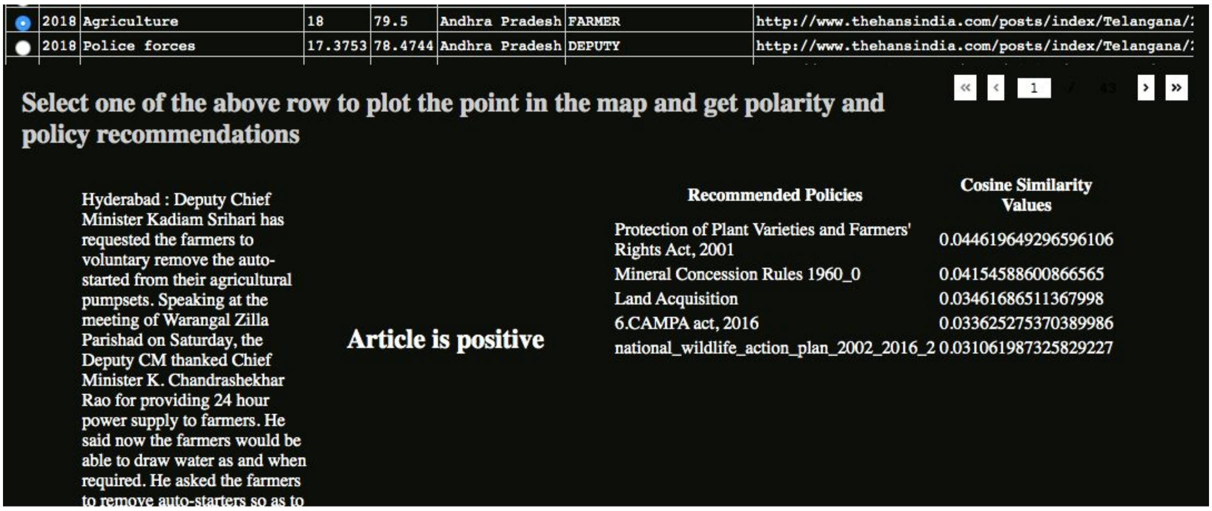
Select one of the above row to plot the point in the map and get polarity and policy recommendations

Hyderabad : Deputy Chief Minister Kadiam Srihari has requested the farmers to voluntary remove the auto-started from their agricultural pumpsets. Speaking at the meeting of Warangal Zilla Parishad on Saturday, the Deputy CM thanked Chief Minister K. Chandrashekhar Rao for providing 24 hour power supply to farmers. He said now the farmers would be able to draw water as and when required. He asked the farmers to remove auto-starters so as to

Article is positive

Recommended Policies	Cosine Similarity Values
Protection of Plant Varieties and Farmers' Rights Act, 2001	0.044619649296596106
Mineral Concession Rules 1960_0	0.04154588600866565
Land Acquisition	0.03461686511367998
6.CAMPA act, 2016	0.033625275370389986
national_wildlife_action_plan_2002_2016_2	0.031061987325829227

Here is a closer look at the article and matched policies in the example above.



●	2018	Agriculture	18	79.5	Andhra Pradesh	FARMER	http://www.thehansindia.com/posts/index/Telangana/
○	2018	Police forces	17.3753	78.4744	Andhra Pradesh	DEPUTY	http://www.thehansindia.com/posts/index/Telangana/

Select one of the above row to plot the point in the map and get polarity and policy recommendations

Hyderabad : Deputy Chief Minister Kadiam Srihari has requested the farmers to voluntary remove the auto-started from their agricultural pumpsets. Speaking at the meeting of Warangal Zilla Parishad on Saturday, the Deputy CM thanked Chief Minister K. Chandrashekhar Rao for providing 24 hour power supply to farmers. He said now the farmers would be able to draw water as and when required. He asked the farmers to remove auto-starters so as to

Article is positive

Recommended Policies	Cosine Similarity Values
Protection of Plant Varieties and Farmers' Rights Act, 2001	0.044619649296596106
Mineral Concession Rules 1960_0	0.04154588600866565
Land Acquisition	0.03461686511367998
6.CAMPA act, 2016	0.033625275370389986
national_wildlife_action_plan_2002_2016_2	0.031061987325829227

Next Steps

This overview describes the results of a pilot project to use natural language processing techniques to identify land conflict events described in news articles and match them to relevant government policies. The project demonstrated that NLP techniques can be successfully deployed to meet this objective.

Potential improvements include refinement of the models and further development of the visualization tool. Opportunities to scale the project include building the library of news articles with those published from additional years and sources, adding to the database of policies, and expanding the geographic focus beyond India.

Opportunities to improve and scale the pilot project

Improvements	Refine models
	Further development of visualization tool
Scale	Expand library of articles with content from additional years and sources
	Expand the database of policies
	Expand the geographic focus beyond India

About the Authors

Laura Clark Murray is the Chief Partnership & Strategy Officer at Omdena. Contact: laura@omdena.com

Nikhel Gupta is a physicist, a Postdoctoral Fellow at the University of Melbourne, and a machine learning engineer with Omdena.

Joanne Burke is a data scientist with MUFG and a machine learning engineer with Omdena.

Rishika Rupam is a Data and AI Researcher with Tilkal and a machine learning engineer with Omdena.

Zaheeda Tshankie is a Junior Data Scientist with Telkom and a machine learning engineer with Omdena.

Omdena Project Team

Kulsoom Abdullah, Joanne Burke, Antonia Calvi, Dennis Dondergoor, Tomasz Grzegorzek, Nikhel Gupta, Sai Tanya Kumbharageri, Michael Lerner, Irene Nanduttu, Kali Prasad, Jose Manuel Ramirez R., Rishika Rupam, Saurav Suresh, Shivam Swarnkar, Jyothsna sai Tagirisa, Elizabeth Tischenko, Carlos Arturo Pimentel Trujillo, Zaheeda Tshankie, Gabriela Urquieta

Partners

This project was done in collaboration with Kathleen Buckingham and John Brandt, our partners with the World Resources Institute (WRI).

About Omdena

Omdena is an innovation platform for building AI solutions to real-world problems through global collaboration. Omdena is a partner of the United Nations AI for Good Global Summit 2020.