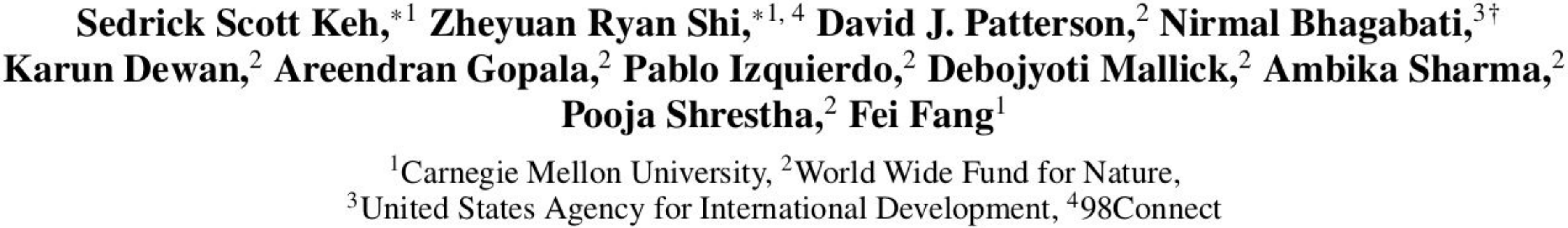
NewsPanda: Media Monitoring for Timely Conservation Action

World Wide Fund for Nature, Carnegie Mellon University

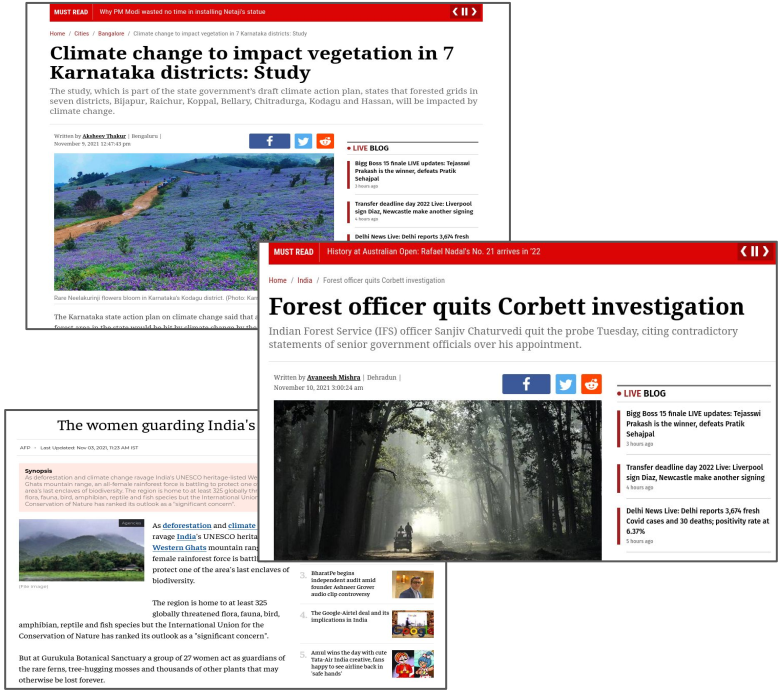


Sedrick Scott Keh,[*1] Zheyuan Ryan Shi,[*1, 4] David J. Patterson,[2] Nirmal Bhagabati,[3†]
Karun Dewan,[2] Areendran Gopala,[2] Pablo Izquierdo,[2] Debojyoti Mallick,[2] Ambika Sharma,[2]
Pooja Shrestha,[2] Fei Fang[1]

[1]Carnegie Mellon University, [2]World Wide Fund for Nature,
[3]United States Agency for International Development, [4]98Connect

Motivation

- WWF country offices spend a lot of time and resources looking through various news articles to identify **trends, events, or threats** related to **conservation and infrastructure**.
    - ○ Identifying 7-10 articles can take 2-3 days to complete
- Having a tool to **automate this process** will save a lot of time for WWF and allow them to more effectively allocate their resources.
- **How do we identify and analyze media articles for timely conservation and infrastructure actions?**



MUST READ | Why PM Modi wasted no time in installing Netaji's statue

Home / Cities / Bangalore / Climate change to impact vegetation in 7 Karnataka districts: Study

Climate change to impact vegetation in 7 Karnataka districts: Study

The study, which is part of the state government's draft climate action plan, states that forested grids in seven districts, Bijapur, Raichur, Koppal, Bellary, Chitradurga, Kodagu and Hassan, will be impacted by climate change.

Written by Aksheev Thakur | Bengaluru |
November 9, 2021 12:47:43 pm

Rare Neelakurinji flowers bloom in Karnataka's Kodagu district. (Photo: Karı

The Karnataka state action plan on climate change said that a

● LIVE BLOG

Bigg Boss 15 finale LIVE updates: Tejasswi Prakash is the winner, defeats Pratik Sehajpal
3 hours ago

Transfer deadline day 2022 Live: Liverpool sign Diaz, Newcastle make another signing
4 hours ago

Delhi News Live: Delhi reports 3,674 fresh

MUST READ | History at Australian Open: Rafael Nadal's No. 21 arrives in '22

Home / India / Forest officer quits Corbett investigation

Forest officer quits Corbett investigation

Indian Forest Service (IFS) officer Sanjiv Chaturvedi quit the probe Tuesday, citing contradictory statements of senior government officials over his appointment.

Written by Avaneesh Mishra | Dehradun |
November 10, 2021 3:00:24 am

● LIVE BLOG

Bigg Boss 15 finale LIVE updates: Tejasswi Prakash is the winner, defeats Pratik Sehajpal
3 hours ago

Transfer deadline day 2022 Live: Liverpool sign Diaz, Newcastle make another signing
4 hours ago

Delhi News Live: Delhi reports 3,674 fresh Covid cases and 30 deaths; positivity rate at 6.37%
5 hours ago

The women guarding India's

AFP • Last Updated: Nov 03, 2021, 11:23 AM IST

Synopsis
As deforestation and climate change ravage India's UNESCO heritage-listed We
Ghats mountain range, an all-female rainforest force is battling to protect one o
area's last enclaves of biodiversity. The region is home to at least 325 globally thr
flora, fauna, bird, amphibian, reptile and fish species but the International Union
Conservation of Nature has ranked its outlook as a "significant concern".

Agencies

(File image)

As deforestation and climate
ravage India's UNESCO herita
Western Ghats mountain ran
female rainforest force is battl
protect one of the area's last enclaves of biodiversity.

The region is home to at least 325 globally threatened flora, fauna, bird, amphibian, reptile and fish species but the International Union for the Conservation of Nature has ranked its outlook as a "significant concern".

But at Gurukula Botanical Sanctuary a group of 27 women act as guardians of the rare ferns, tree-hugging mosses and thousands of other plants that may otherwise be lost forever.

3. BharatPe begins independent audit amid founder Ashneer Grover audio clip controversy

4. The Google-Airtel deal and its implications in India

5. Amul wins the day with cute Tata-Air India creative, fans happy to see airline back in 'safe hands'

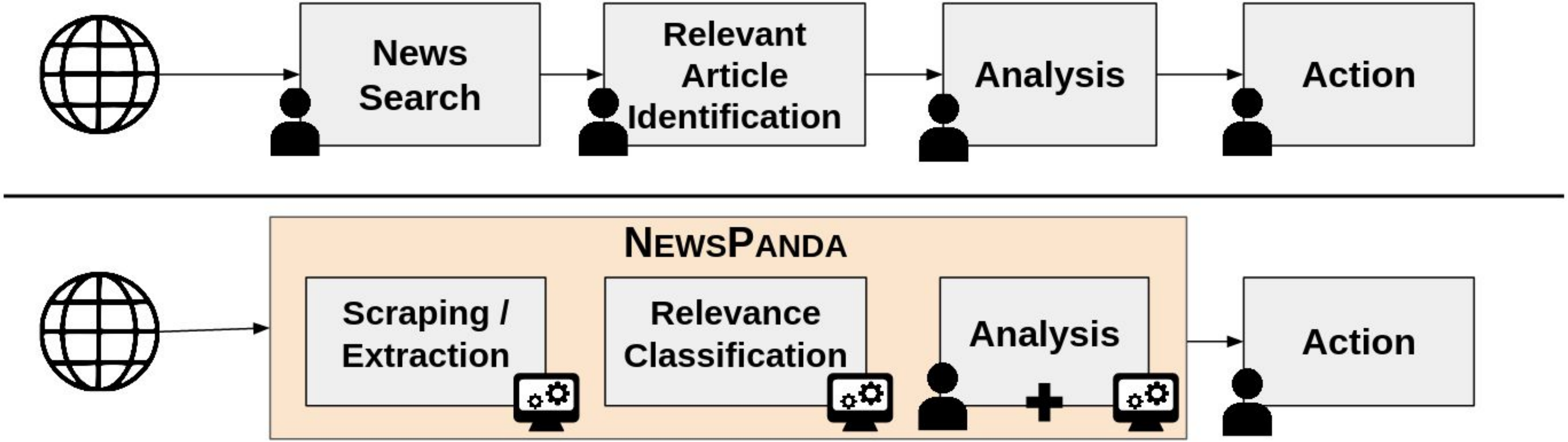
Motivation

- WWF country offices spend a lot of time and resources looking through various news articles to identify **trends, events, or threats** related to **conservation** and **infrastructure**.
  - Identifying 7-10 articles can take 2-3 days to complete
- Having a tool to **automate this process** will save a lot of time for WWF and allow them to more effectively allocate their resources.
- **How do we identify and analyze media articles for timely conservation and infrastructure actions?**

Infrastructure?

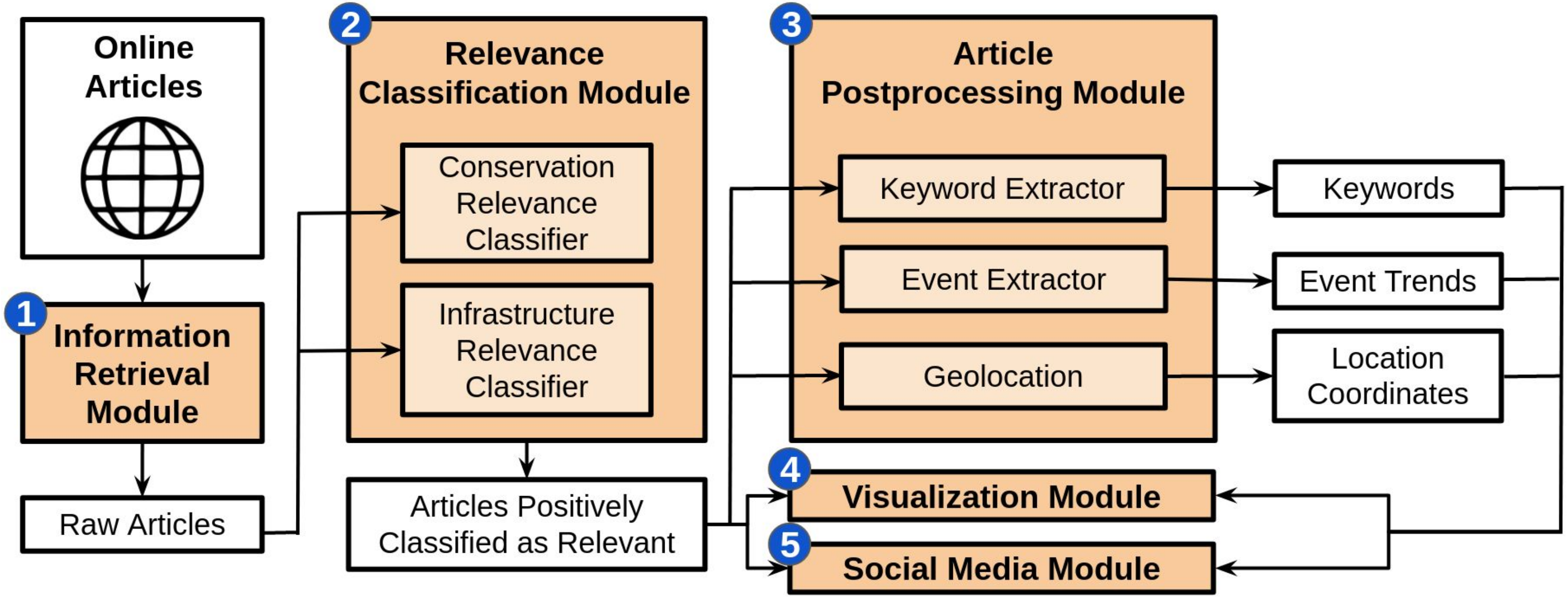
- roads, railways, pipelines, etc.
- Usually high-impact and long-term
- These articles usually cover upcoming developments, which is where WWF can truly perform the necessary interventions

NewsPanda motivation



NewsPanda automates multiple steps in the pipeline, enabling humans to perform the more critical tasks (analysis and action).

NewsPanda pipeline

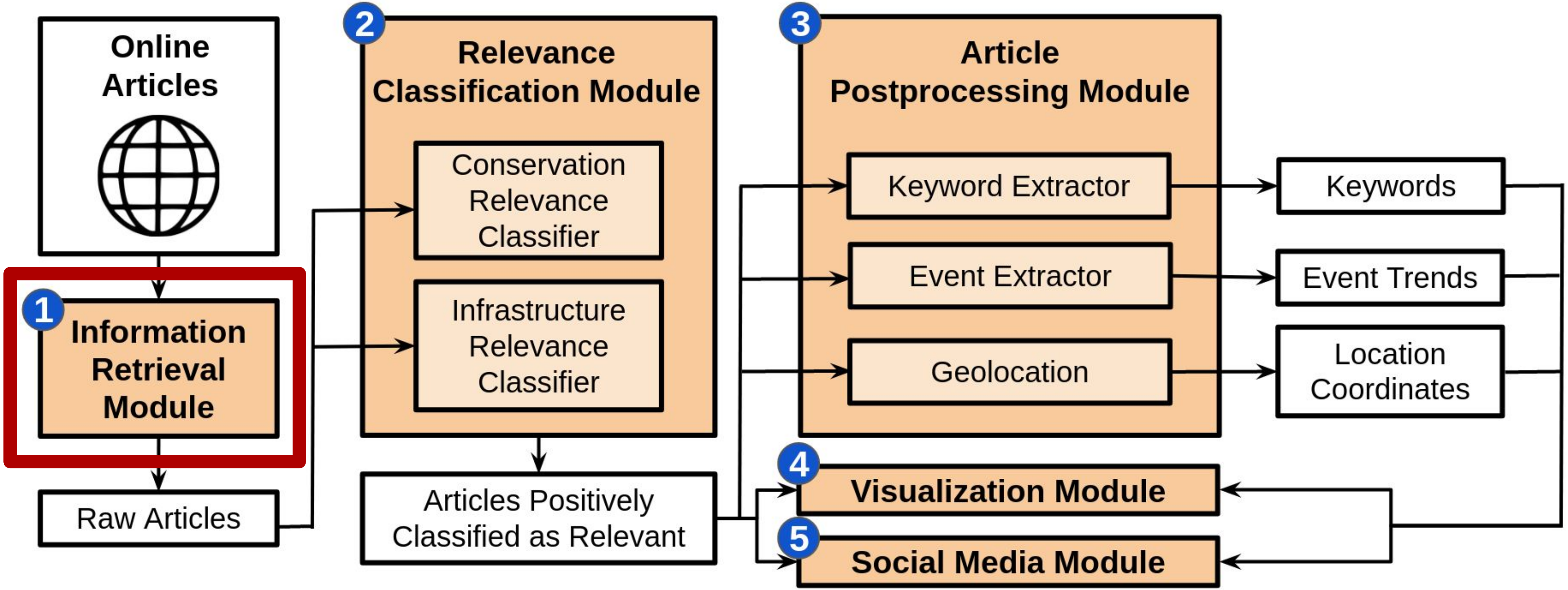

Online Articles
1
Information Retrieval Module
Raw Articles
2
Relevance Classification Module
Conservation Relevance Classifier
Infrastructure Relevance Classifier
Articles Positively Classified as Relevant
3
Article Postprocessing Module
Keyword Extractor
Event Extractor
Geolocation
Keywords
Event Trends
Location Coordinates
4
Visualization Module
5
Social Media Module

This entire pipeline is ran on a weekly basis.

Dataset

- **Initial dataset** – We start off with two datasets, with labels along two dimensions: conservation relevance and infrastructure relevance
  a. **WHS-Corp dataset (44,000 articles; 928 with labels)**
     - from Hosseini and Coll Ardanuy (2020)
     - global news articles covering World Heritage Sites around the world
     - only contains labels for conservation relevance
  b. **InfraCorp dataset (4,137 articles; 1,000 with labels)**
     - our own dataset which we scrape + annotate
     - focus specifically on India and Nepal
     - scraping done using NewsAPI
     - each of the 1,000 articles is annotated by two domain experts at WWF

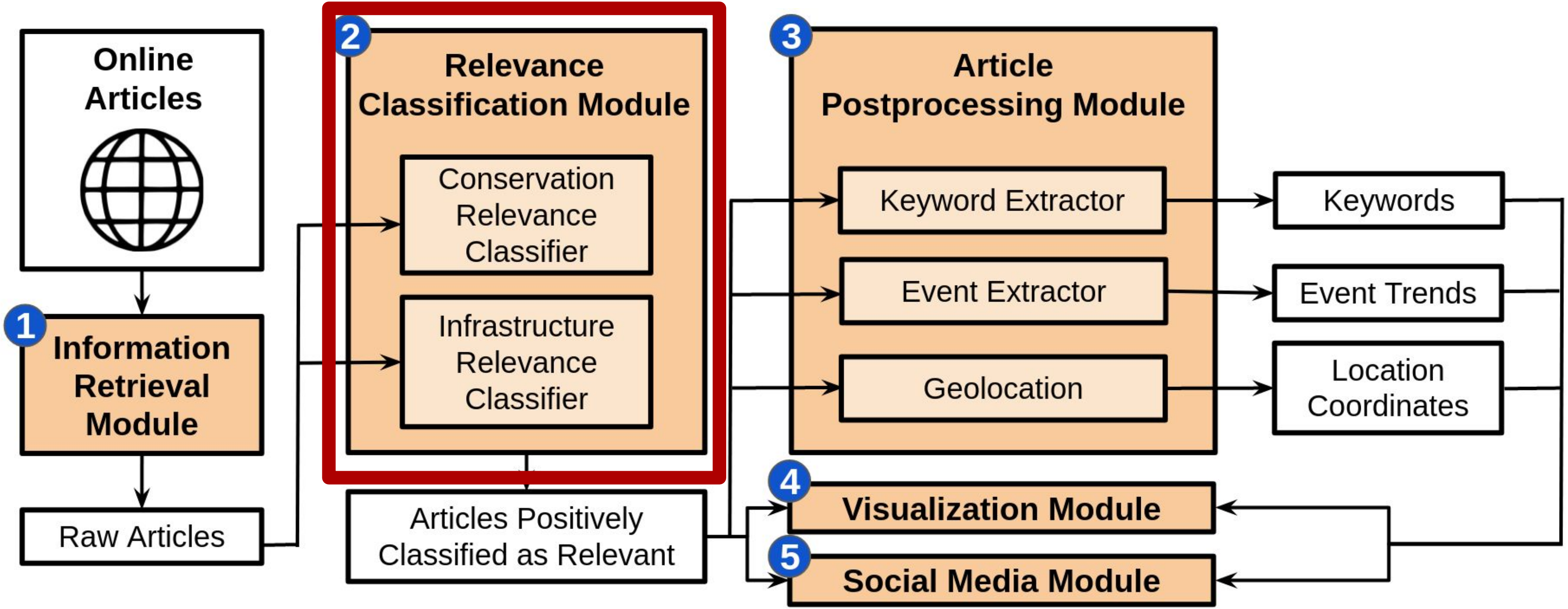
NewsPanda pipeline


Online Articles
1
Information Retrieval Module
Raw Articles
2
Relevance Classification Module
Conservation Relevance Classifier
Infrastructure Relevance Classifier
Articles Positively Classified as Relevant
3
Article Postprocessing Module
Keyword Extractor
Event Extractor
Geolocation
Keywords
Event Trends
Location Coordinates
4
Visualization Module
5
Social Media Module

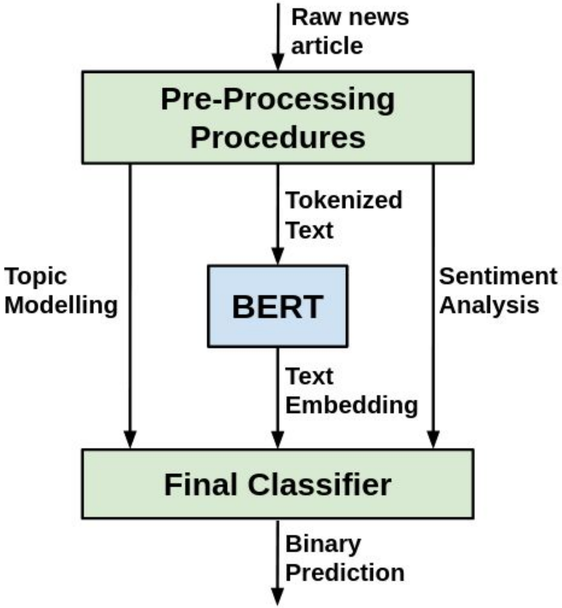
1. Information Retrieval Module

- Use the NewsAPI scraper with search terms taken from a list of curated conservation sites by WWF
- Focus on India and Nepal
- This is ran on a weekly basis

NewsPanda pipeline


Online Articles
1
Information Retrieval Module
Raw Articles
2
Relevance Classification Module
Conservation Relevance Classifier
Infrastructure Relevance Classifier
Articles Positively Classified as Relevant
3
Article Postprocessing Module
Keyword Extractor
Event Extractor
Geolocation
Keywords
Event Trends
Location Coordinates
4
Visualization Module
5
Social Media Module

2. Relevance Classification Module

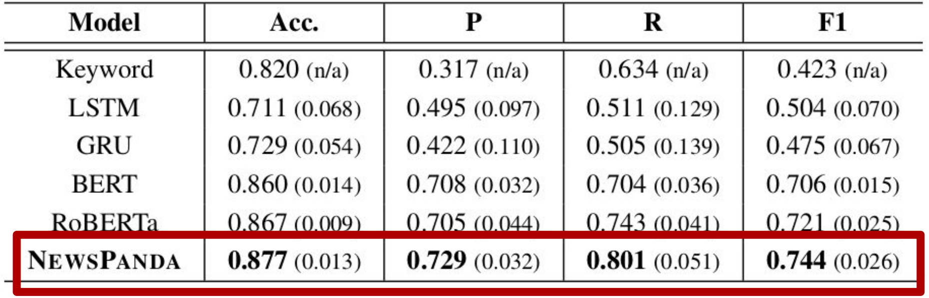


We include the following features for each article:

- BERT embedding
- Sentiment analysis score
- Topic modelling vector

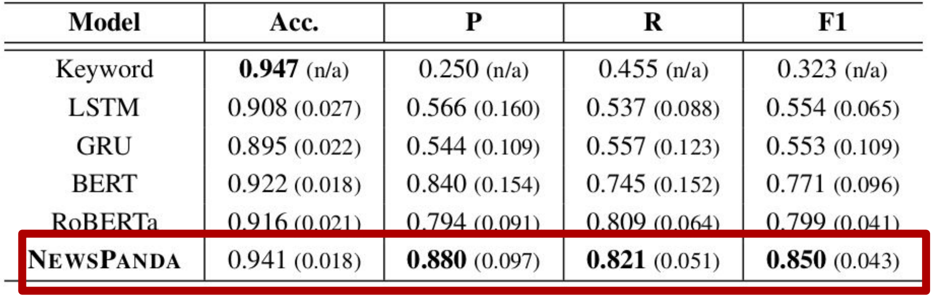
Prediction is done along two dimensions: **conservation relevance** and **infrastructure relevance**

2. Relevance Classification Module



Model	Acc.	P	R	F1
Keyword	0.820 (n/a)	0.317 (n/a)	0.634 (n/a)	0.423 (n/a)
LSTM	0.711 (0.068)	0.495 (0.097)	0.511 (0.129)	0.504 (0.070)
GRU	0.729 (0.054)	0.422 (0.110)	0.505 (0.139)	0.475 (0.067)
BERT	0.860 (0.014)	0.708 (0.032)	0.704 (0.036)	0.706 (0.015)
RoBERTa	0.867 (0.009)	0.705 (0.044)	0.743 (0.041)	0.721 (0.025)
NEWSPANDA	**0.877** (0.013)	**0.729** (0.032)	**0.801** (0.051)	**0.744** (0.026)

(a) Scores for *Conservation Relevance*



Model	Acc.	P	R	F1
Keyword	**0.947** (n/a)	0.250 (n/a)	0.455 (n/a)	0.323 (n/a)
LSTM	0.908 (0.027)	0.566 (0.160)	0.537 (0.088)	0.554 (0.065)
GRU	0.895 (0.022)	0.544 (0.109)	0.557 (0.123)	0.553 (0.109)
BERT	0.922 (0.018)	0.840 (0.154)	0.745 (0.152)	0.771 (0.096)
RoBERTa	0.916 (0.021)	0.794 (0.091)	0.809 (0.064)	0.799 (0.041)
NEWSPANDA	0.941 (0.018)	**0.880** (0.097)	**0.821** (0.051)	**0.850** (0.043)

(b) Scores for *Infrastructure Relevance*

NewsPanda performs the best across all the models and baselines.

Recall the InfraCorp Dataset

This leads us to two key questions:

1. How do we best select which 1,000 articles out of the 4,137 to label?
2. How to we best handle label noise in our annotated dataset?

b. **InfraCorp dataset (4,137 articles; 1,000 with labels)**

- our own dataset which we scrape + annotate
- focus specifically on India and Nepal
- scraping done using NewsAPI
- each of the 1,000 articles is annotated by two domain experts at WWF

Recall the InfraCorp Dataset

This leads us to two key questions:

1. **How do we best select which 1,000 articles out of the 4,137 to label?**
2. How to we best handle label noise in our annotated dataset?

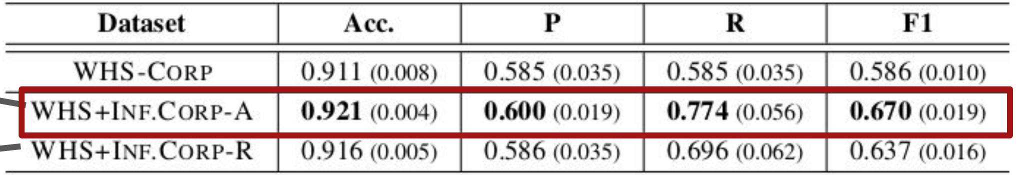
Confidence-based active learning:

- Train an initial model using the available WHS-Corp dataset
- Select the 1,000 most “difficult” articles, i.e. the articles which the initial model is “least confident” about.

2. Relevance Classification Module

Ablation study: Confidence-based active learning

- Select two sets of 300 articles – set A is actively selected, and set R is randomly selected.



actively selected

randomly selected

Dataset	Acc.	P	R	F1
WHS-CORP	0.911 (0.008)	0.585 (0.035)	0.585 (0.035)	0.586 (0.010)
WHS+INF.CORP-A	**0.921** (0.004)	**0.600** (0.019)	**0.774** (0.056)	**0.670** (0.019)
WHS+INF.CORP-R	0.916 (0.005)	0.586 (0.035)	0.696 (0.062)	0.637 (0.016)

Using the actively selected set gives a larger performance gain as compared to using a randomly selected set.

Recall the InfraCorp Dataset

This leads us to two key questions:

1. How do we best select which 1,000 articles out of the 4,137 to label?
2. **How to we best handle label noise in our annotated dataset?**

Noisy label correction methods:

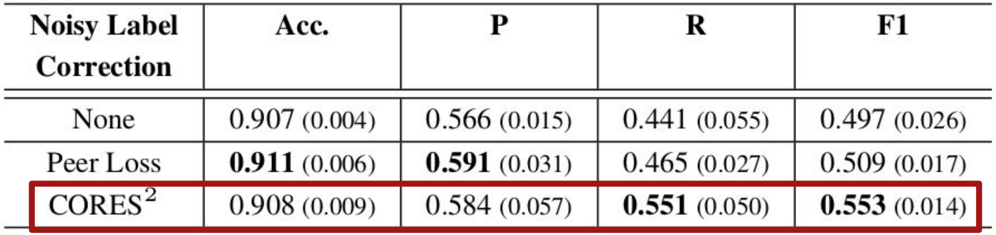
- Adapt the CORES^2 loss (Cheng et al. 2021) noise correction algorithm
- Extension of earlier peer loss algorithm – frames the task of learning from noisy labels as a peer prediction problem

$$\ell_{\text{CORES}}(f(x_n),\tilde{y}_n) := \ell(f(x_n),\tilde{y}_n) - \beta\cdot\mathbb{E}_{\mathcal{D}_{\tilde{Y}|\tilde{D}}}[\ell(f(x_n),\tilde{Y})]$$

2. Relevance Classification Module

Ablation study: Noisy label correction algorithms

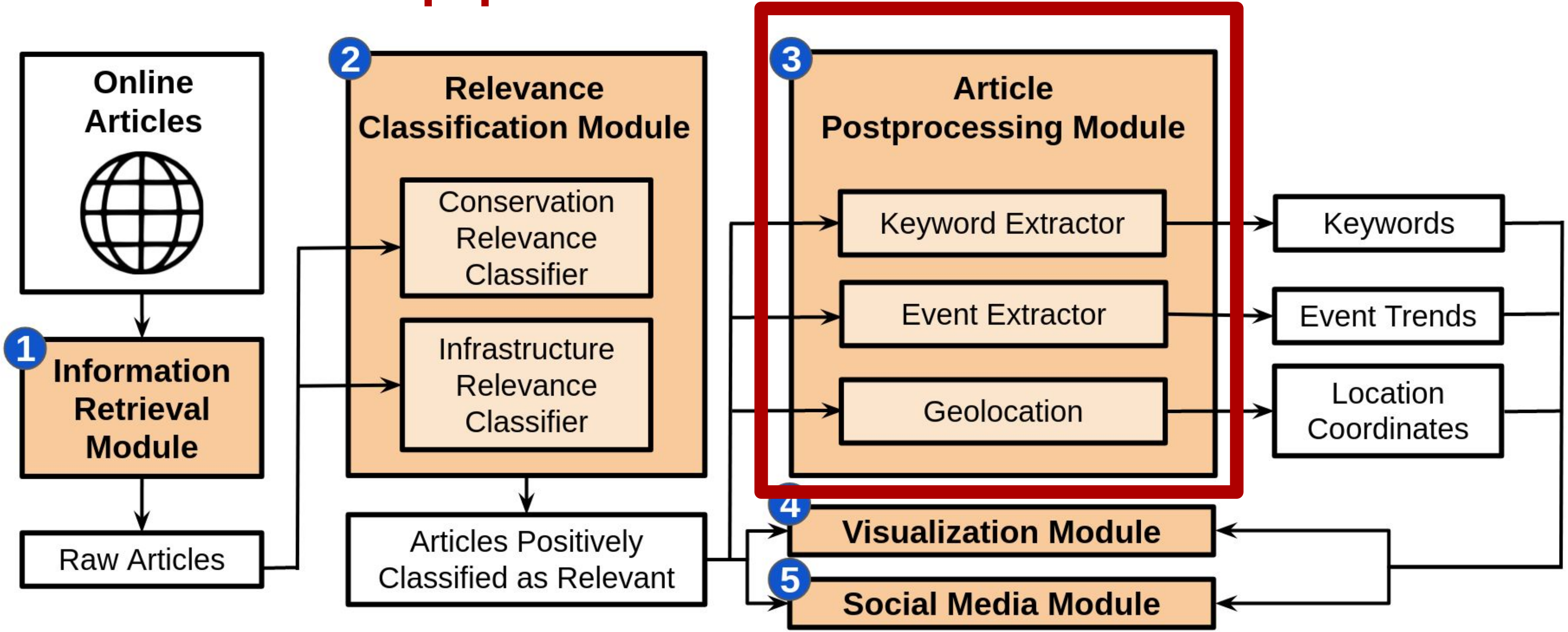
- Investigate the effects of using peer loss and CORES^2 loss



Noisy Label Correction	Acc.	P	R	F1
None	0.907 (0.004)	0.566 (0.015)	0.441 (0.055)	0.497 (0.026)
Peer Loss	**0.911** (0.006)	**0.591** (0.031)	0.465 (0.027)	0.509 (0.017)
CORES^2	0.908 (0.009)	0.584 (0.057)	**0.551** (0.050)	**0.553** (0.014)

Using CORES^2 loss yields the best performance compared to using Peer Loss and using no noisy label correction.

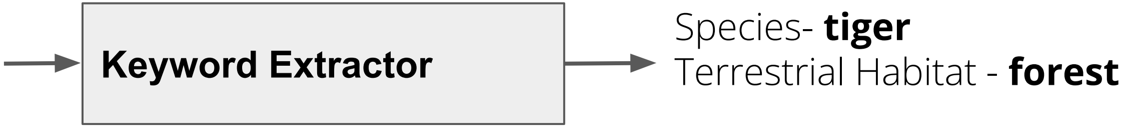
NewsPanda pipeline


Online Articles
1
Information Retrieval Module
Raw Articles
2
Relevance Classification Module
Conservation Relevance Classifier
Infrastructure Relevance Classifier
Articles Positively Classified as Relevant
3
Article Postprocessing Module
Keyword Extractor
Event Extractor
Geolocation
Keywords
Event Trends
Location Coordinates
4
Visualization Module
5
Social Media Module

Keyword extraction

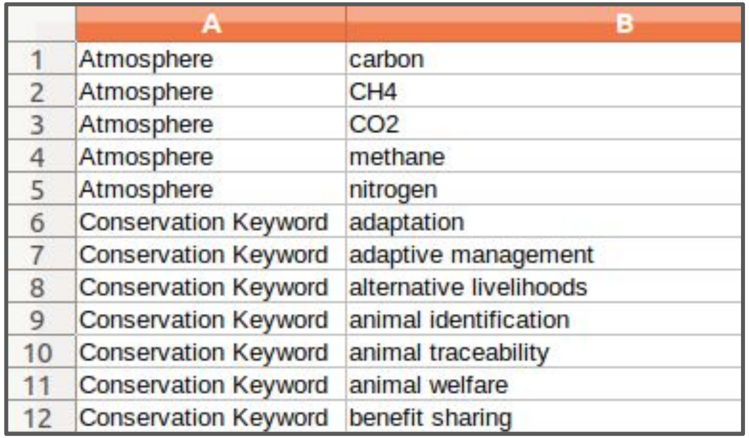
(example paragraph)

A 45-year-old man from Chamrajanagar in Karnataka was arrested by the Forest Department for attempting to hunt wild animals and also setting fire in **forest** areas in the Talavadi Forest Range in the Sathyamangalam **Tiger** Reserve here.



Keyword Extractor

Species- **tiger**
Terrestrial Habitat - **forest**



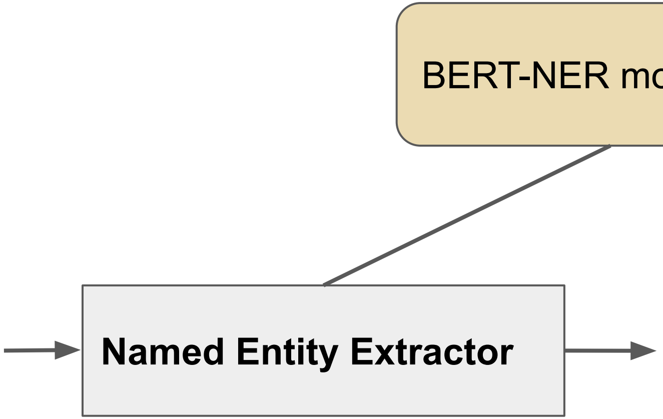
	A	B
1	Atmosphere	carbon
2	Atmosphere	CH4
3	Atmosphere	CO2
4	Atmosphere	methane
5	Atmosphere	nitrogen
6	Conservation Keyword	adaptation
7	Conservation Keyword	adaptive management
8	Conservation Keyword	alternative livelihoods
9	Conservation Keyword	animal identification
10	Conservation Keyword	animal traceability
11	Conservation Keyword	animal welfare
12	Conservation Keyword	benefit sharing

We use a list of around 1000 keywords (level 1 and level 2), then check for matches in the text.

Named entity recognition

(example paragraph)

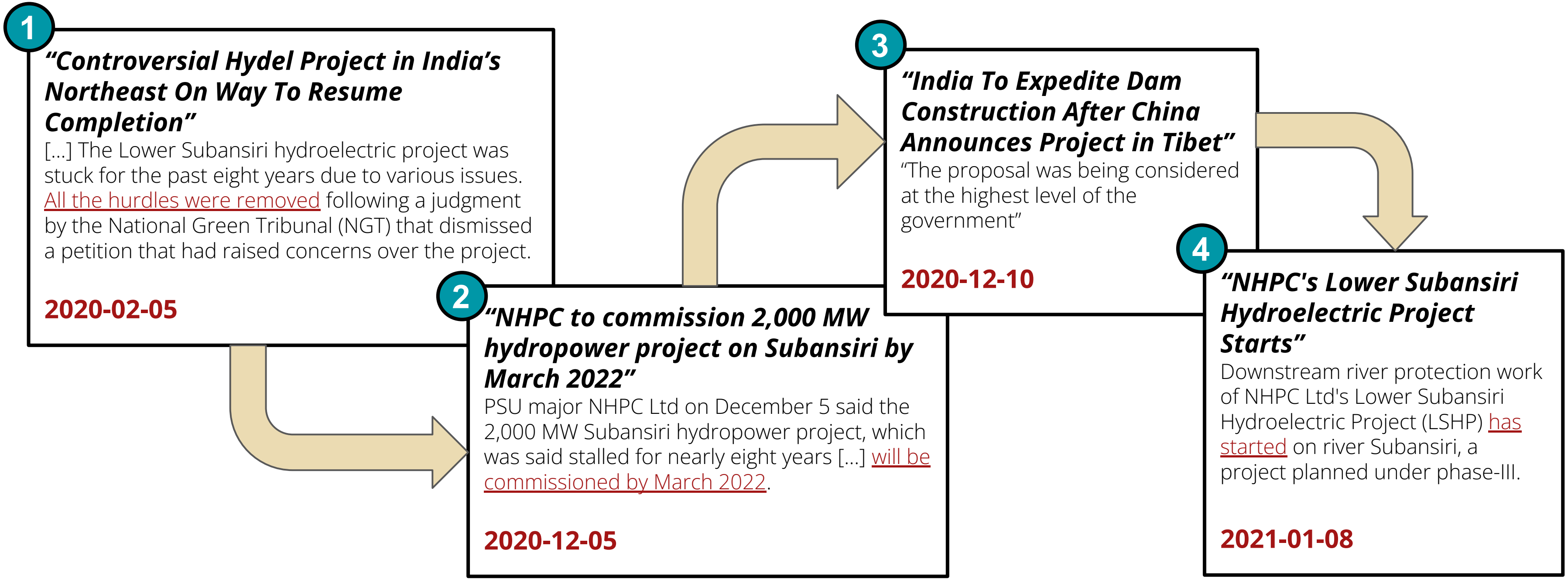
A 45-year-old man from **Chamrajanagar** in **Karnataka** was arrested by the **Forest Department** for attempting to hunt wild animals and also setting fire in forest areas in the **Talavadi Forest Range** in the **Sathyamangalam Tiger Reserve** here.



LOC - Chamrajanagar
LOC - Karnataka
ORG - Forest Department
LOC - Talavadi Forest Range
LOC - Sathyamangalam Tiger Reserve

Event timeline

We search for entity **“Subansiri”**, then filter with the keyword “**hydroelectric**”.



1

“Controversial Hydel Project in India’s Northeast On Way To Resume Completion”

[...] The Lower Subansiri hydroelectric project was stuck for the past eight years due to various issues. All the hurdles were removed following a judgment by the National Green Tribunal (NGT) that dismissed a petition that had raised concerns over the project.

2020-02-05

2

“NHPC to commission 2,000 MW hydropower project on Subansiri by March 2022”

PSU major NHPC Ltd on December 5 said the 2,000 MW Subansiri hydropower project, which was said stalled for nearly eight years [...] will be commissioned by March 2022.

2020-12-05

3

“India To Expedite Dam Construction After China Announces Project in Tibet”

“The proposal was being considered at the highest level of the government”

2020-12-10

4

“NHPC's Lower Subansiri Hydroelectric Project Starts”

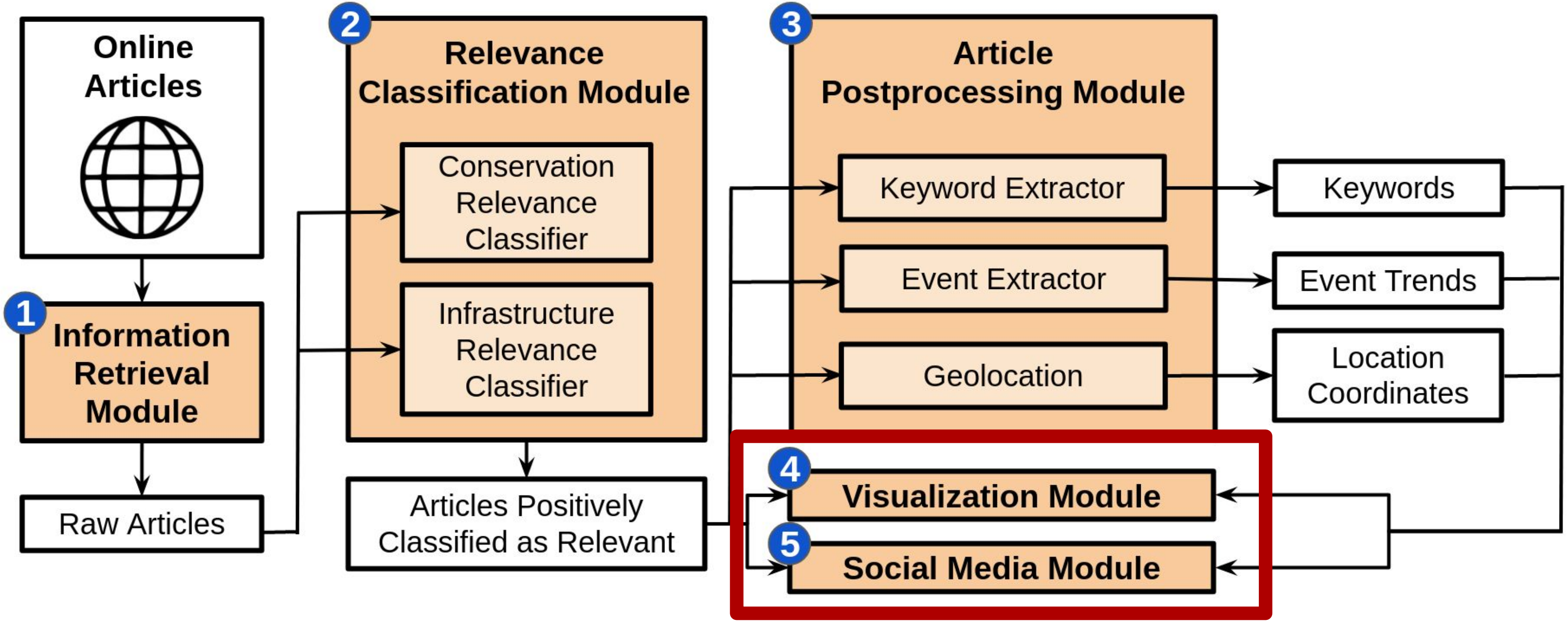
Downstream river protection work of NHPC Ltd's Lower Subansiri Hydroelectric Project (LSHP) has started on river Subansiri, a project planned under phase-III.

2021-01-08

Geolocation

- Important to integrate into WWF’s GIS systems
- Use a directory of conservation sites from WWF to map articles to their coordinates
- If there is no match in directory, we use the geopy package

NewsPanda pipeline


Online Articles
1
Information Retrieval Module
Raw Articles
2
Relevance Classification Module
Conservation Relevance Classifier
Infrastructure Relevance Classifier
Articles Positively Classified as Relevant
3
Article Postprocessing Module
Keyword Extractor
Event Extractor
Geolocation
Keywords
Event Trends
Location Coordinates
4
Visualization Module
5
Social Media Module

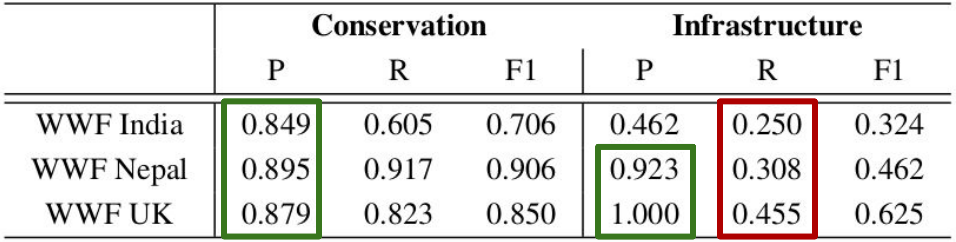
Deployment

- NewsPanda has been deployed by WWF teams in India, Nepal, and the UK since February 2022
- Three stages of deployment:
  1. Pilot study (February 2022)
     - Goal: Test out the pipeline and identify some operational and technical issues in the initial version of NewsPanda
  2. Initial deployment (March 2022 to July 2022)
     - Goal: Evaluate the performance of NewsPanda
  3. Sustainable deployment (August 2022 onwards)
     - Goal: Make pipeline more automatic and cloud-based

Deployment Results

Quantitative results:

- Each week, the WWF teams from India, Nepal, and the UK evaluated the articles classified by NewsPanda



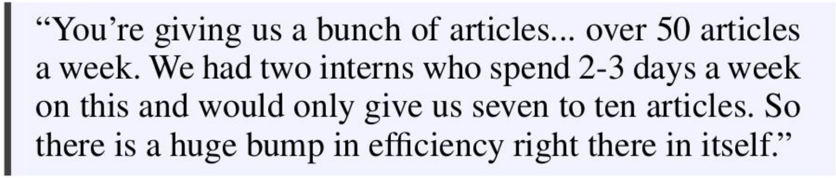
	Conservation			Infrastructure		
	P	R	F1	P	R	F1
WWF India	0.849	0.605	0.706	0.462	0.250	0.324
WWF Nepal	0.895	0.917	0.906	0.923	0.308	0.462
WWF UK	0.879	0.823	0.850	1.000	0.455	0.625

- **High precision values** = trustworthy and reliable system
- **Low recall for infrastructure** = misses out on potential articles; needs to improve more on positively identifying relevant articles

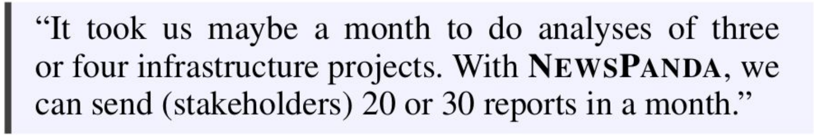
Deployment Results

Qualitative results:

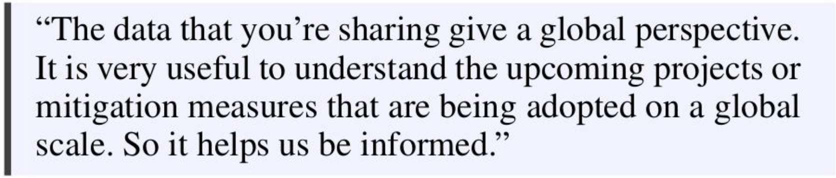
- Two months into deployment, the CMU team carried out semi-structured interviews with their WWF colleagues who have been using NewsPanda outputs in their work



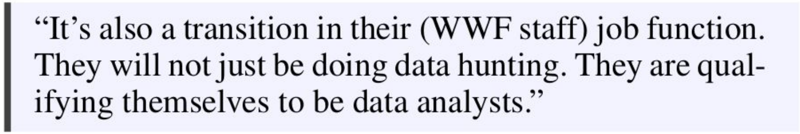
> "You're giving us a bunch of articles... over 50 articles a week. We had two interns who spend 2-3 days a week on this and would only give us seven to ten articles. So there is a huge bump in efficiency right there in itself."



> "It took us maybe a month to do analyses of three or four infrastructure projects. With **NEWSPANDA**, we can send (stakeholders) 20 or 30 reports in a month."

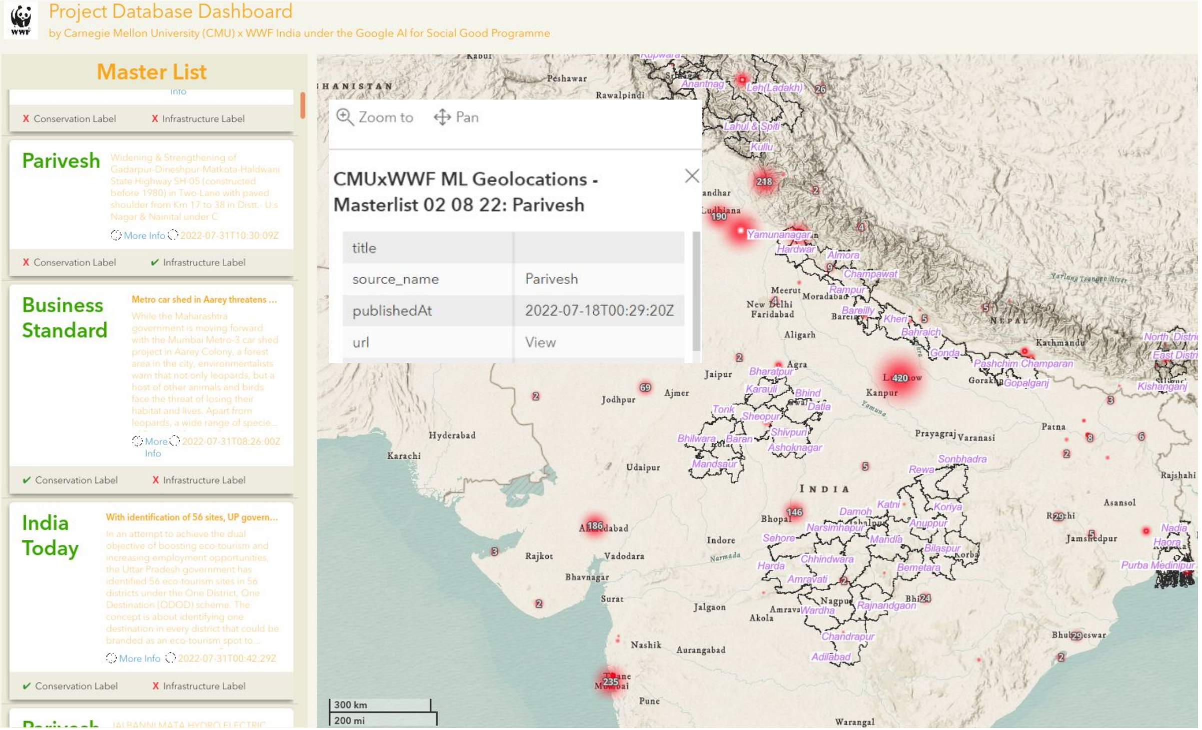


> "The data that you're sharing give a global perspective. It is very useful to understand the upcoming projects or mitigation measures that are being adopted on a global scale. So it helps us be informed."



> "It's also a transition in their (WWF staff) job function. They will not just be doing data hunting. They are qualifying themselves to be data analysts."

4. Visualization Module



Project Database Dashboard

by Carnegie Mellon University (CMU) x WWF India under the Google AI for Social Good Programme

Master List

Info

X Conservation Label X Infrastructure Label

Parivesh Widening & Strengthening of Gadarpur-Dineshpur-Matkota-Haldwani State Highway SH-05 (constructed before 1980) in Two-Lane with paved shoulder from Km 17 to 38 in Distt.- U.s Nagar & Nainital under C

More Info 2022-07-31T10:30:09Z

X Conservation Label ✔ Infrastructure Label

Business Standard Metro car shed in Aarey threatens ...

While the Maharashtra government is moving forward with the Mumbai Metro-3 car shed project in Aarey Colony, a forest area in the city, environmentalists warn that not only leopards, but a host of other animals and birds face the threat of losing their habitat and lives. Apart from leopards, a wide range of specie...

More Info 2022-07-31T08:26:00Z

✔ Conservation Label X Infrastructure Label

India Today With identification of 56 sites, UP govern...

In an attempt to achieve the dual objective of boosting eco-tourism and increasing employment opportunities, the Uttar Pradesh government has identified 56 eco-tourism sites in 56 districts under the One District, One Destination (ODOD) scheme. The concept is about identifying one destination in every district that could be branded as an eco-tourism spot to...

More Info 2022-07-31T00:42:29Z

✔ Conservation Label X Infrastructure Label

Parivesh JALBANNI MATA HYDRO ELECTRIC

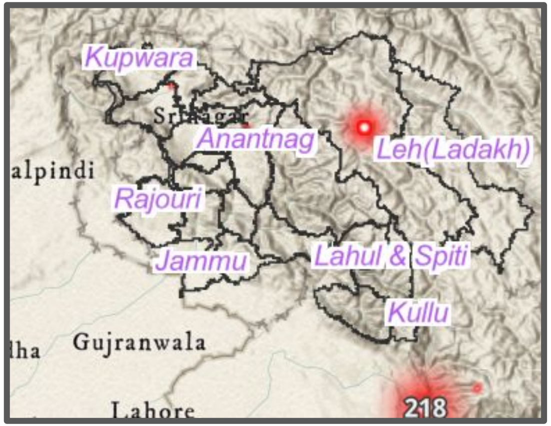
CMUxWWF ML Geolocations - Masterlist 02 08 22: Parivesh

title	
source_name	Parivesh
publishedAt	2022-07-18T00:29:20Z
url	View

The NewsPanda results are integrated into WWF’s GIS systems, which is especially useful for the field teams.

4. Visualization Module – Success Story

- August 2022: NewsPanda highlighted **Ikhala Block Boundary Kishtwar to Lopara Road** in the WWF GIS system
- Upon further investigation, it is found that the project would **divert 5.9 hectares of forest land**
- More importantly, WWF found that the project was still at its **pre-proposal stage**. This means WWF would be able to take early action and possibly participate in relevant conversations.

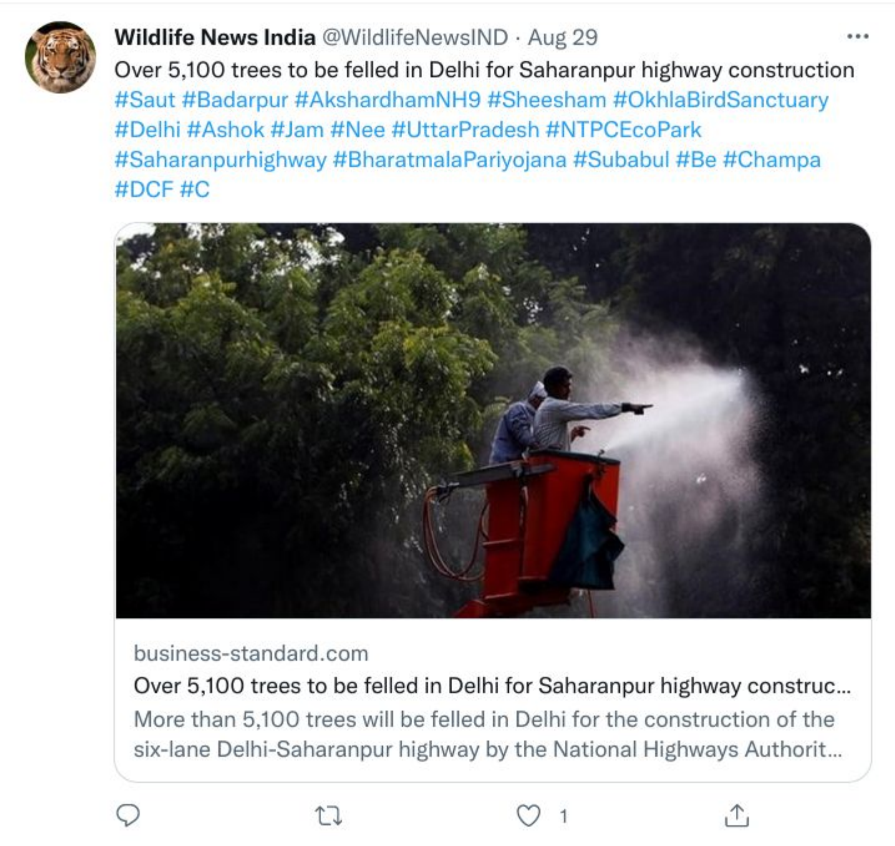


5. Social Media Module

For the general public to benefit from NewsPanda, we also developed a Twitter bot which tweets links and hashtags (keywords) to the relevant weekly articles.

@WildlifeNewsIND

Go follow and share! :)



Wildlife News India @WildlifeNewsIND · Aug 29

Over 5,100 trees to be felled in Delhi for Saharanpur highway construction #Saut #Badarpur #AkshardhamNH9 #Sheesham #OkhlaBirdSanctuary #Delhi #Ashok #Jam #Nee #UttarPradesh #NTPCEcoPark #Saharanpurhighway #BharatmalaPariyojana #Subabul #Be #Champa #DCF #C

business-standard.com

Over 5,100 trees to be felled in Delhi for Saharanpur highway construc...

More than 5,100 trees will be felled in Delhi for the construction of the six-lane Delhi-Saharanpur highway by the National Highways Authorit...

Current Steps + Future Work

- Expand to **more languages** and **more local media sources**
- Starting with **Nepali** language articles
- Multilingual language models (e.g. multilingual BERT, XLM-R)
    - Initial results show **good generalizability** to other languages

- **Ongoing challenge: Annotating training data for multiple languages is costly. How can we best use NewsPanda in a few-shot or zero-shot manner?**