# NewsPanda: Media Monitoring for Timely Conservation Action

**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

## Abstract

Non-governmental organizations for environmental conservation have a significant interest in monitoring conservation-related media and getting timely updates about infrastructure construction projects as they may cause massive impact to key conservation areas. Such monitoring, however, is difficult and time-consuming. We introduce **NEWSPANDA**, a toolkit which automatically detects and analyzes online articles related to environmental conservation and infrastructure construction. We fine-tune a BERT-based model using active learning methods and noise correction algorithms to identify articles that are relevant to conservation and infrastructure construction. For the identified articles, we perform further analysis, extracting keywords and finding potentially related sources. **NEWSPANDA** has been successfully deployed by the World Wide Fund for Nature teams in the UK, India, and Nepal since February 2022. It currently monitors over 80,000 websites and 1,074 conservation sites across India and Nepal, saving more than 30 hours of human efforts weekly. We have now scaled it up to cover 60,000 conservation sites globally.

## 1 Introduction

Massive floods, poaching of wildlife, waste pollution, climate change – every week, new threats impacting our environment come to light. Each of these events can cause a long chain of negative impacts if not addressed. As such, monitoring these conservation-related events is of great importance for non-governmental organizations (NGOs) focused on environmental conservation such as the World Wide Fund for Nature (WWF) to take timely action and participate in relevant conversations.

In addition to the conservation topic as a whole, many NGOs are particularly interested in monitoring news on certain subtopics. One such area is the ongoing or upcoming infrastructure projects such as roads, railways, and pipelines. Environmental impact mitigation measures proposed for these projects are often inadequate. Conservation NGOs such as WWF play a key role in advocating for more sustainable infrastructure development. Early detection and engagement of these projects could shift infrastructure planning towards more environmentally sustainable outcomes while benefiting the people that the projects intend to serve.

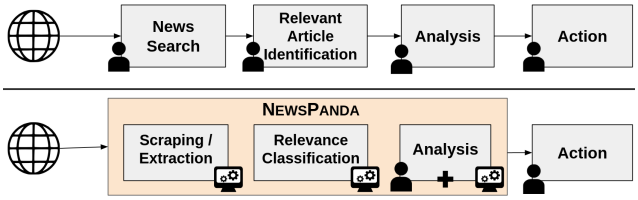



Figure 1: **Top:** Current costly and time-consuming information gathering pipeline at NGOs. **Bottom: NEWSPANDA** automates multiple steps in the pipeline, enabling humans to perform the more critical tasks (analysis and action).

However, information about conservation-related events and infrastructure plans threatening critical habitats and species is scattered across numerous sources and comes in different forms. NGOs typically learn of such information through word-of-mouth or a handful of news outlets that they check manually. This data collection process is both time-consuming and ineffective, and it can potentially fail to capture critical information in a timely manner, leaving these NGOs out of key conversations during early or ongoing stages of these developments.

To fill this gap, we develop **NEWSPANDA**, a natural language processing (NLP) based toolkit to automatically detect and analyze public news and government articles describing emerging and current threats to conservation areas. **NEWSPANDA** has five main components, which we detail in Section 3. At the core of **NEWSPANDA** is a classification module built using a BERT-based language model, which we specifically fine-tune to classify whether articles are relevant to conservation and to infrastructure.

Developing such a tool in the conservation nonprofit setting poses several unique challenges. First, labeling data is expensive. We propose an active learning-based method to selectively acquire labels on the most critical data points. Second, the data labels could be noisy since labeling an ar-

[*]These authors contributed equally.

[†]Nirmal Bhagabati was not at USAID when the research for this paper was conducted. The views and opinions expressed in this paper are those of the authors and not necessarily those of USAID.

ticle as conservation-related or not is ultimately a subjective judgement, even if we fix a labeling rubric. We adopt a state-of-the-art noise reduction algorithm (Cheng et al. 2021) to improve our model's performance.

**NEWSPANDA** was developed as a collaboration between WWF and Carnegie Mellon University (CMU). It has been successfully deployed since February 2022 and has been used by the WWF teams in the UK, India, and Nepal to monitor developments in conservation sites. The entire pipeline runs on a weekly basis, scraping and classifying relevant news articles regarding conservation and infrastructure construction related events that occurred in the past week. These articles are then visualized in WWF's GIS systems for the field teams to investigate. We also share some results through social media for the benefit of the broader civil society. Through the deployment of **NEWSPANDA**, the WWF teams have been able to save over 30 hours weekly on collecting news, which allows us at WWF to instead focus on analyzing the news and taking actions (Figure 1).

## 2 Related Work

**News Monitoring Systems** Although there is a rich literature on news information extraction in general domains (Ojokoh 2012; Reis et al. 2004; Dandeniya 2018) as well as some specific applications (Murray et al. 2019; Joshi, N, and Rao 2016), there has been hardly any media monitoring tool for environmental conservation and infrastructure construction. Directly using generic media monitoring tools often lead to unsatisfactory results that are not localized enough to be actionable for a specific conservation site or not relevant enough to be reliable. As a result, conservation NGOs still mostly use a manual process to collect articles. The only work on conservation news monitoring that we are aware of is a preliminary attempt by Hosseini and Coll Ardanuy (2020) that apply BERT to classify news articles. Compared to that, with **NEWSPANDA** we provide a classification module with algorithmic contributions to address challenges in using the tool in the nonprofit context, a full end-to-end information extraction and processing pipeline, and most importantly, results and lessons learned from a large scale actual deployment of the tool. This is the first comprehensive and actionable media monitoring tool for conservation and infrastructure.

**NLP for Conservation & Infrastructure** Outside of news monitoring, NLP tools have been used for various applications in conservation and infrastructure. Some analyze the relevant news articles for general insights on conservation reporting (Santos and Crowder 2021) or study their spread and impact (Wu et al. 2018). These studies are descriptive in nature and orthogonal to our work. The few studies that take the civil society stakeholder's perspective are focused on different links in the process from us. Luccioni, Baylor, and Duchene (2020) use BERT-based models to analyze corporate environment sustainability reports. Boutilier and Bahr (2020) explore mining-related texts to analyze the social license of a particular project. They target different problems from us. They assume a relevant text is readily available and try to extract meaningful insights from it. On the other hand, we work on identifying that relevant text from thousands of irrelevant texts in the first place and leave the insight extraction to professional organizations like WWF that have been doing that for years.

## 3 NEWSPANDA Overview

**NEWSPANDA** toolkit consists of five modules as illustrated below and in Figure 2a. During pilot study and deployment (Section 8), this entire pipeline is run on a weekly basis.

1. **Information Retrieval Module**: We use the `NewsAPI` scraper (Lisivick 2018) with the names of conservation sites taken from a curated list of conservation areas.
2. **Relevance Classification Module**: We classify articles along two dimensions, namely *Conservation Relevance* and *Infrastructure Relevance*, through a large pretrained language model fine-tuned with our collected dataset. Details of this model are explained in Section 5.
3. **Article Postprocessing Module**: The article postprocessing module has 3 parts: a keyword extractor which extracts keywords, an event extractor which extracts event trends, and a geolocator which provides location coordinates. We discuss these features in Section 6.
4. **Visualization Module**: After the relevant articles are identified, we visualize them in our GIS system at WWF, which we can further analyze and act upon (Section 8).
5. **Social Media Module**: In parallel to the visualization module, another downstream application for **NEWSPANDA** is WILDLIFENEWSINDIA, [1] a Twitter bot we built from **NEWSPANDA** that shares weekly relevant conservation-related articles on social media (Section 8).

## 4 Dataset

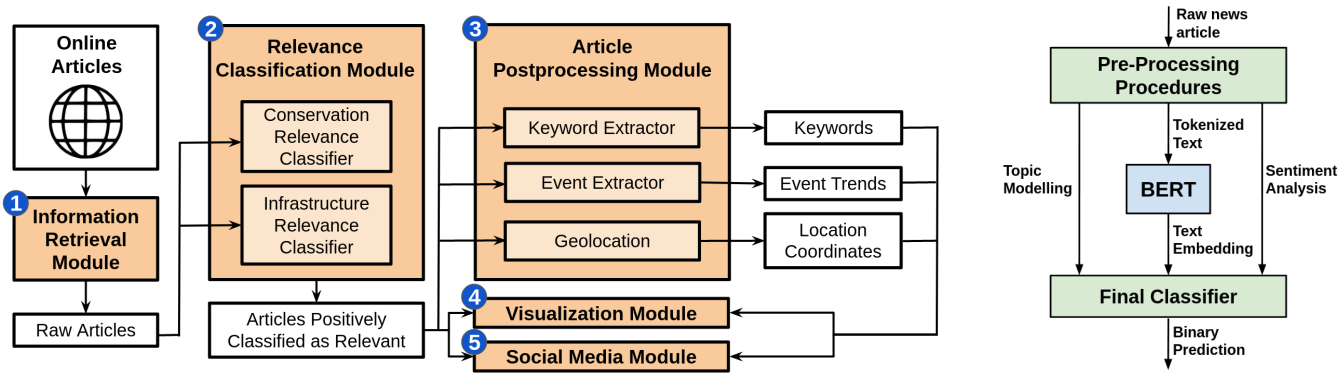
We use two main datasets for developing **NEWSPANDA**. First, we use an existing corpus (WHS-CORP) by Hosseini and Coll Ardanuy (2020) consisting of articles scraped using World Heritage Sites as keywords and labelled by domain experts. Second, we scrape and label our own corpus (INFRACORP), which is a more focused, timely, and fine-grained upgrade over WHS-CORP. Both datasets contain English articles relevant to conservation, but they differ in terms of the locations of the conservation sites used, as well as the time frame of the articles. These are detailed below.

### 4.1 WHS-CORP Dataset

WHS-CORP contains over 44,000 articles from 2,974 different sources covering 224 World Heritage Sites from around the world. Scraping was done using `NewsAPI`'s Python library from a list of curated conservation sites of interest. Aside from the title and the content, it also contains other metadata such as the publication site, the author, and the date of publication. Articles in WHS-CORP span the time frame from January 2018 to October 2019.

After these articles were gathered, a subset of 928 articles were sampled and manually annotated for *Conservation Relevance* by domain experts familiar with conservation. *Conservation Relevance* denotes whether an article discusses

[1] `https://twitter.com/WildlifeNewsIND`




(a) Diagram of overall NEWSPANDA pipeline, with the five key modules in orange boxes. Generated outputs of NEWSPANDA are in the white boxes.

(b) Conservation and infrastructure classification models.

Figure 2: NEWSPANDA pipeline (2a) and model diagram for conservation and infrastructure relevance classifiers (2b).

threats or impacts to wildlife and environment conservation in general, e.g. poaching, forest development, natural disasters. We use this labelled dataset for training our model.

### 4.2 INFRACORP Dataset

As opposed to WHS-CORP which focuses on global conservation sites, INFRACORP specifically focuses on conservation sites in India and Nepal. The INFRACORP corpus contains 4,137 articles (150 for Nepal and 3,987 for India) from 1,074 conservation sites across the two countries. All articles were taken in the two-year span from November 2019 to November 2021. We use `NewsAPI` to search for the official names of the conservation sites, or alternative and/or local names for the sites as recorded at WWF.

Given the data availability as well as the annotator capacity of the local domain experts from India and Nepal, we labeled all the 150 articles from Nepal and only 1,000 articles from India. We detail the selection criteria for selecting these 1,000 articles in Section 5.2. Annotation for INFRACORP was done along two dimensions: *Conservation Relevance* and *Infrastructure Relevance*. Here, *Conservation Relevance* is similar to the one described for WHS-CORP in Section 4.1. Among the articles which were labelled as positive for *Conservation Relevance*, we further categorize whether it is relevant to infrastructure. This covers issues such as new roads in forested areas and construction projects near national parks. By definition *Infrastructure Relevance* is a subset of *Conservation Relevance*. Each article was annotated by two domain experts, one from WWF UK, and another from either WWF India or WWF Nepal. We provided the annotators with a descriptive rubric for labeling in each dimension, as well as concrete examples of edge cases. The following was one such example in our instructions:

> Articles describing tourism or wildlife or natural beauty of a national park, but without talking about environmental impacts or threats to wildlife and conservation, do not count as positive for *Conservation Relevance*.

Where the two sets of labels disagree, the authors closely inspect the articles and decide on the final labels.

## 5 Relevance Classification Module

We highlight the structure of our NEWSPANDA classification module and other key techniques used during training.

### 5.1 Classification Model

The backbone of the NEWSPANDA classification model is a BERT model (Devlin et al. 2019) with a linear classification head. BERT is a Transformer-based language model trained using masked language modelling and next sentence prediction objectives on large-scale corpora of books and articles. This large-scale pretraining, as well as its ability to effectively encode context, leads to superior performance on a wide variety of tasks. We adapt BERT to the domain of conservation and infrastructure, and we fine-tune it to perform news article classification. In Section 7, we explore different variants of the BERT model (such as RoBERTa).

One key change we make to the BERT model is that in the final linear head after the main BERT layers, instead of only considering the BERT vector outputs, we also incorporate other features, namely sentiment analysis and topic modelling, as shown in Figure 2b. We hypothesize that including these additional features will provide the model with more useful information that will help classify whether or not a particular article is relevant to infrastructure or conservation. For instance, if an article has topic vectors that align with other articles covering forest habitats, but it has an overwhelmingly positive sentiment, then we may suspect that it could be a tourism-related feature article instead of a conservation-related news article (which are often more neutral or negative in terms of sentiment).

For sentiment analysis, we extract the sentence polarity scores of the article title, its description, and its content, giving us three sentiment scores per article. This is done on a scale of $-1.0$ to $+1.0$, with $-1.0$ representing the most negative score and $+1.0$ representing the most positive score. Sentiment analysis was done using the `textblob` package (Loria 2018). Meanwhile, for topic extraction, we consider the entire training corpora of WHS-CORP and INFRACORP, and train a Latent Dirichlet Allocation (LDA) model to identify topic clusters. We use 50 topics for the LDA model and

implemented it using `scikit-learn` (Pedregosa et al. 2011). Lastly, for the main BERT model, we concatenate the title, description, and content of each article, and we use this concatenated text as input to our classifier. For cases where the article is missing certain features (e.g. no description), we simply supply an empty string for that feature. The vectors from the three steps (i.e. BERT model, sentiment analysis, topic modelling) are then concatenated, and this final vector is used as the input to the final classification head to generate a binary prediction. Specific implementation settings and other hyperparameters can be found in Section 7.1.

### 5.2 Active Learning

Annotating a dataset is costly. In curating our INFRACORP dataset, we need to be mindful of which specific articles to label in order for our model to learn most efficiently. For this selection process, we first fine-tune a pretrained RoBERTa-base model on the existing WHS-CORP dataset, based on the *Classification Relevance*. To make this preliminary model as close to our final model as possible, we also incorporate the topic modelling and sentiment analysis features, as shown in Figure 2b. Because this is only a preliminary model, we forego doing extensive hyperparameter tuning and decided to just select a setting that worked decently well: with a learning rate of 1e-5, batch size of 16, and training for 10 epochs, we were able to get an F-score of 0.61 on WHS-CORP. Using this trained model, we then generate *Classification Relevance* predictions for all articles in the INFRACORP corpus, together with the corresponding softmax scores. We treat these softmax scores as a measure for the classification confidence of the model: if the softmax is close to 0 or close to 1, then it means that the model is very certain with its prediction, while if the softmax is close to 0.5, then it means the model is unsure with its prediction.

We then select 300 articles which our model is least confident about. We hypothesize that selecting these "difficult" rows will have the greatest impact on model performance. We call this active learning-based dataset INFRACORP-A. To verify the effectiveness of active learning, we also randomly sample 300 articles to label, which we call INFRACORP-R. We will later evaluate how this compares with the actively selected dataset on a randomly selected test set of 400 samples in our ablation study (Section 7.3).

### 5.3 Noisy Label Correction

Our dataset is labelled by two sets of domain expert annotators from WWF. Although we provided detailed criteria for labelling each article, there is always room for some subjectivity in the process. This resulted in the two sets of labels not agreeing with each other on over $10\%$ of the data points. Although, as mentioned in Section 4.2, we did manage to obtain the "ground truth" label for a small subset of INFRACORP for model evaluation purposes, doing that for every single article is prohibitively expensive – much more expensive than the (not cheap) process of having either annotator providing a (noisy) label. Therefore, in order for **NEWSPANDA** to work well once deployed, we need to be able to learn well from the potentially noisy labels only.

More formally, let $x_n$ be the embedding of an article along with its sentiment and topic modeling vectors as described in Section 5.1. Let $y_n$ be the true label of this article. The task is to make an accurate prediction on the dataset $\{(x_n, y_n) : n = 1 \dots N\}$ when we only have access to the noisy data $\{(x_n, \tilde{y}_n) : n = 1 \dots N\}$ where $\tilde{y}_n$ is the label that we get from either of the two annotators, and the true labels $y_n$ are the final labels that we decide on after resolving conflicts.

To address this challenge, we adapt the CORES$^2$ loss (Cheng et al. 2021) noise correction algorithm, which is an extension of the earlier peer loss (Liu and Guo 2020). Peer loss frames the task of learning from noisy labels as a peer prediction problem. In practice, the loss for each $(x_n, y_n)$ data point can be calculated using the standard cross entropy loss with $(x_n, y_n)$, modified with a loss calculated using a randomly sampled input $x_{n_1}$ and an *independently* randomly sampled label $y_{n_2}$. That is, we have

$$\ell_{\text{PEER}}(f(x_n), \tilde{y}_n) := \ell(f(x_n), \tilde{y}_n) - \alpha \cdot \ell(f(x_{n_1}), \tilde{y}_{n_2})$$

where $\alpha > 0$ is a tunable parameter. Meanwhile, CORES$^2$ replaces the random sampling from peer loss with a confidence regularizer defined as follows:

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

where $\tilde{D}$ is the dataset, $\tilde{Y}$ is a noisy label, and $\beta > 0$ is a tunable parameter. Following Cheng et al. (2021), we calculate this confidence regularizer term using an estimate of the noise prior probability. We test both peer loss and CORES$^2$ loss, and report results in our ablation study (Section 7.3).

## 6 Article Postprocessing Module

Once the relevant articles are identified using the model, we then perform a few post-processing steps to extract key information and make them easier to analyze and visualize.

### 6.1 Keyword Extractor

Keywords are important, as they allow the easy summarization, categorization, and grouping of news articles. Furthermore, we also use these keywords as hashtags in our social media module (Section 8). To extract keywords, we use an extensive list of conservation-related keywords maintained at WWF. and search the article for exact matches. In addition, we also use Named Entity Recognition systems to extract the salient words in each article. To perform this, we use a BERT-based model trained on the CoNLL 2003 Named Entity Recognition dataset (Tjong Kim Sang and De Meulder 2003). The keywords extracted using these two methods are then concatenated to form the final set of keywords.

### 6.2 Event Extractor

To track the progress of infrastructure projects, it is often not enough to just view a single article in isolation. Rather, news regarding these projects often builds up over a period of weeks or months. To help provide this context, we create an automated event extractor, which leverages our INFRACORP dataset, including both the labelled articles as well as the unlabelled articles. Given a new article $a$, our goal is to find past articles $P_a$ which are closely related to $a$. We first

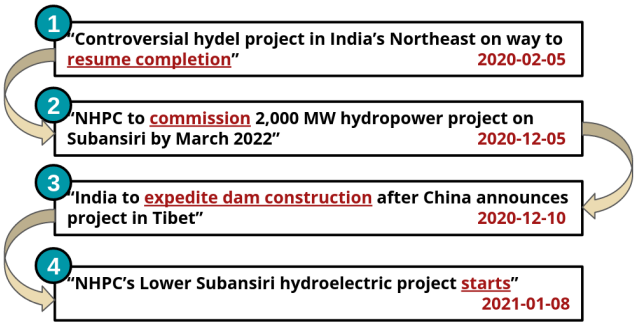



Figure 3: Example of events selected by the Event Extractor (Section 6.2) by date. The progression of the project is highlighted by the phrases in red underline.

gather all previous articles which are from the same conservation site. Next, we create a graph $G_a$, where each article is a node, and two nodes share an edge if the corresponding articles share $\geq k$ common keywords (from Section 6.1). Here, $k$ is an adjustable parameter depending on how loosely connected we want $G_a$ to be. For our data, we use $k = 3$. Once the graph $G_a$ is constructed, we then define an "event" to be the maximal clique containing $a$, and we report all such events. A sample chain of events is shown in Figure 3.

### 6.3 Geolocation

To aid with visualization (Section 8), we perform geolocation on the classified news articles, based on the search terms used to retrieve them. To extract the latitude and longitude coordinates, we leverage an extensive directory of conservation sites from WWF, and we use the directory to map conservation sites to their corresponding coordinates. For cases where there is no match to the directory, we use the geolocation feature of the `geopy` package.

## 7 Experiments and Results

Here, we discuss results of our in-lab experiments (Section 7.1, 7.2) and ablation studies to verify our hypotheses (Sections 7.3). Results from our real-world deployment with WWF are discussed in the succeeding section (Section 8).

### 7.1 Experiment Settings

**Baselines** We compare the performance of our **NEWSPANDA** model with the following baselines:

1. **Keyword model**: We consider a naive model that checks for the count of certain keywords. We curate two sets of "conservation-related keywords" and "infrastructure-related keywords". If an article contains more than $k$ "conservation-related keywords", then it is considered to be relevant to conservation (likewise for infrastructure).
2. **RNN-based models**: We tokenize each article, then pass the embedding to RNN models, where the hidden state of the last layer is used as input to the final classification layer. We use two types of RNN models, namely GRUs (Bahdanau, Cho, and Bengio 2014) and LSTMs (Hochreiter and Schmidhuber 1997).

| 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*

Table 1: Average scores for *Conservation Relevance* (Table 1a) and *Infrastructure Relevance* (Table 1b), taken over 10 random seeds. Standard deviations are shown in parentheses, except for the keyword model, which is deterministic.

3. **BERT-based models**: We fine-tune a pretrained BERT-base (Devlin et al. 2019) and RoBERTa-base model (Liu et al. 2019), where we add a classification head after the final layer to perform relevance classification.

**Evaluation Metrics** Since our task is binary classification, we measure the accuracy, precision, recall, and F1-score. For precision, recall, and F1, we consider only the scores of the positive class. All metrics are calculated separately for *Conservation Relevance* and *Infrastructure Relevance*.

**Data** For *Conservation Relevance*, we train on the INFRACORP dataset (consisting of both INFRACORP-A and INFRACORP-R), as well as the WHS-CORP dataset. For *Infrastructure Relevance*, since WHS-CORP does not contain infrastructure labels, we only train using INFRACORP. We split the training data into an 80-20 training-validation split. For evaluation, we use the test split of INFRACORP for both *Conservation Relevance* and *Infrastructure Relevance*.

**Implementation Settings** For the GRU and LSTM models, we use a batch size of 128, a hidden size of 128, and a dropout of 0.2. We train for 10 epochs with a learning rate of 1e-4. Meanwhile, for BERT, RoBERTa, and **NEWSPANDA**, we train for 10 epochs with a batch size of 4 and a learning rate of 1e-5. We use RoBERTa for the backbone model of **NEWSPANDA**. For all models, we use the Adam optimizer and a final linear head size of 768. Model selection is done by considering the best validation F1-score.

### 7.2 Results and Analysis

Experimental results are shown in Tables 1a and 1b. We observe that indeed, adding the sentiment analysis and topic modelling features, as well as the CORES$^2$ loss for noisy label correction, aids in predictions for both *Conservation Relevance* and *Infrastructure Relevance*, providing an improvement over both BERT-base and RoBERTa-base.

| 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) |

Table 2: Evaluation scores for *Conservation Relevance* for INFRACORP-A compared with INFRACORP-R, taken over 10 random seeds with standard deviations in parentheses.

| 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) |

Table 3: Evaluation scores for *Conservation Relevance* for two noise correction methods, averaged over 10 random seeds with standard deviations in parentheses.

Our data is quite imbalanced: >80% of the articles are not relevant. This manifests itself in the discrepancies between accuracy and F1-score. We observe, for example, that the naive keyword model has very high accuracy scores but very low F1-scores, which indicates that it predicts a lot of zeros (hence the high accuracy), but is not able to predict the relevant articles well. The RNN-based models (LSTM and GRU) seem to perform relatively poorly, achieving an F1-score of around 0.5. This could also be attributed to the data imbalance, since these RNN-based models are generally not as robust to imbalanced datasets. In contrast, the BERT and RoBERTa models perform quite well, with F1-scores >0.7 for conservation and >0.75 for infrastructure, and precision/recall scores also around that range. This indicates that these transformer-based models are able to generalize quite well and successfully capture the notions of *Conservation Relevance* and *Infrastructure Relevance*. Lastly, **NEWSPANDA** offers significant improvement over the RoBERTa-base model (F1 t-test $p$-value $= 0.018$ for conservation and $0.033$ for infrastructure), showing the positive effects of incorporating information such as the emotion and topics over simply considering the article text in isolation.

### 7.3 Ablation Study

In this section, we investigate the effect of the different components of the **NEWSPANDA** model. More specifically, we evaluate the active learning strategy (Section 5.2) and the noisy label correction strategy (Section 5.3).

**Active Learning** We compare the effect with training on actively-sampled data (INFRACORP-A) and randomly-sampled data (INFRACORP-R). Each of these datasets contain 300 India articles, as detailed in Section 5.2 and 4.2. We append these articles to the existing WHS-CORP to create the final data for training. We use the RoBERTa model for these experiments. Results are shown in Table 2.

For both INFRACORP-A and INFRACORP-R, we see an improvement over just using WHS-CORP. Indeed, training with more data will result in better performance, regardless of how the data is sampled. We also observe that adding actively sampled data results in a larger improvement than adding randomly sampled data across all metrics (F1 t-test $p$-value $= 0.004$). This verifies the effectiveness of our hypothesized confidence-based data selection for annotation.

**Noisy Label Correction** We examine the effect of the noise correction methods outlined in Section 5.3, by comparing the effect of using peer loss, CORES$^2$ loss, and standard cross entropy loss. Based on INFRACORP, we use the labels supplied by one of the two annotators for the training set, and the final calibrated labels for the test set. Hyperparameter search was done for both peer loss and CORES$^2$ loss to find the optimal values of $\alpha = 0.05$ and $\beta = 0.05$. We trained for 20 epochs with a learning rate of 2e-5.

From Table 3, we observe that for accuracy and precision, all three losses perform very similarly, with peer loss performing the highest by a small margin. For recall and F1, peer loss and the standard loss perform at comparable levels, while CORES$^2$ loss performs better than both (F1 t-test $p$-value $= 0.001$). This is likely because the confidence regularizer used in CORES$^2$ works better than the random sampling used by peer loss. Both peer and CORES$^2$ loss might work even better if we had more training data than the current 600 in INFRACORP. In the end, given the positive results of CORES$^2$, we used it in our **NEWSPANDA** model.

## 8 Deployment and Impact

**NEWSPANDA** has been used at WWF teams since February 2022. In this section, we describe the deployment process, results, and lessons learned.

### 8.1 Pilot Study

The first stage of **NEWSPANDA** deployment, which is the pilot study, started in February 2022 and ran for around one month. Every week, the CMU team scraped the news articles and ran the entire **NEWSPANDA** pipeline, forwarding the outputs to the WWF teams to examine and provide feedback. During this pilot phase, the WWF and CMU teams identified a range of operational and technical issues in the initial version of **NEWSPANDA**.

First, in order for **NEWSPANDA** to fit into the established workflow of WWF, it needs to be integrated into its GIS system. During the pilot, we realized that it is crucial to add the geolocation of each article (Section 6.3) and format the model output according to the specifications of the GIS platform used at WWF. Figure 4 shows how **NEWSPANDA**'s results get integrated into the GIS system, with the red areas being the locations where we identify a relevant article.

We also discovered that while `NewsAPI` has a good collection of global news sources, it fails to include some relevant sources in the local context. With the suggestions from the WWF team, we incorporated additional sources that often yield relevant local articles. One such site is Parivesh, which contains proposals of infrastructure projects in India.

Finally, we discovered that some conservation sites' names often lead to irrelevant search results and hence inefficiencies in the pipeline. We revised the search terms, or in some cases, dropped the sites from the list.

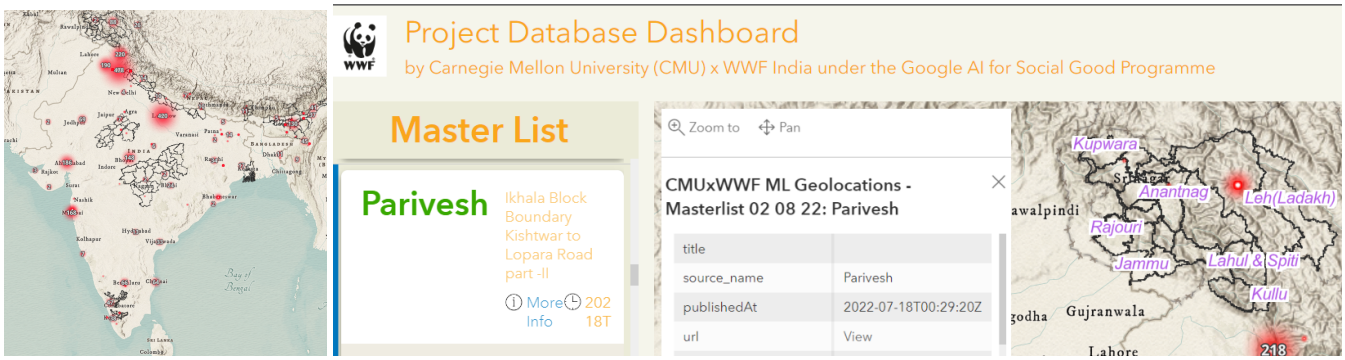



Figure 4: Left: The highlighted red areas indicate clusters of articles found by our model. Right: The WWF GIS system, where each relevant article is shown on the map with its corresponding key details.

## 8.2 Deployment Results

After we resolved the above issues, we proceeded with the actual deployment. The procedure was similar to the pilot phase, except that at this phase, the focus is to evaluate the performance of **NEWSPANDA**. The WWF teams closely inspected the model predictions each week and provided ground truth labels for each article. The label feedback allowed the CMU team to retrain the model regularly. This stage ran from March 2022 to July 2022. Table 4 shows the aggregated results over 5 months of evaluation results from WWF India, Nepal, and UK. WWF UK labeled the first half of the deployment for all locations and India/Nepal labeled the second half for news articles in their respective countries.

Overall, **NEWSPANDA** continued to show great performance in *Conservation Relevance* during real-world deployment. Across all evaluations, the precision scores are consistently high, indicating that almost all of the articles reported by **NEWSPANDA** are indeed relevant. We intentionally tuned the model towards this direction – when almost everything that the model flagged is relevant, it would greatly help with establishing the trust in the model at the early stage of deployment. As we continue developing the model, we aim to improve the model towards achieving higher recall, to be able to capture more relevant articles.

On the other hand, on *Infrastructure Relevance* for India, the model's performance was worse than the offline experiments. Upon further inspection, we discovered that the majority of mistakes were in fact only 2-4 original pieces of news that were paraphrased by various news sources into 20-40 articles. Since there are only a few *Infrastructure Relevance* positive articles to start with, this had a big impact on the model performance. Meanwhile, such phenomenon did not occur in our offline experiments because there we randomly sampled news from a large corpus for labeling.

Aside from overall metrics, we also highlight individual success stories. Figure 4(right) shows a concrete example where **NEWSPANDA** made a difference. In early August, 2022, **NEWSPANDA** detected a new project of Ikhala Block Boundary Kishtwar to Lopara Road and highlighted it in the WWF GIS system. Upon further investigation by WWF staff, 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. Such stories are happening frequently since the deployment of **NEWSPANDA**. Using the tool's outputs integrated into our internal GIS systems, the WWF staff are continuously coordinating with our field teams to examine the status and report on relevant projects and areas.

| | 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 |

Table 4: Aggregated scores of **NEWSPANDA** on weekly articles from March 2022 to July 2022.

## 8.3 Qualitative and Quantitative Comparison with Current Practice

Prior to **NEWSPANDA**, WWF had already been monitoring media for conservation-related articles (Figure 1). However, most of these efforts were not very structured or logged. It is thus difficult to draw head-to-head comparisons between **NEWSPANDA** and WWF's existing approach. That said, we still provide qualitative and quantitative evidence supporting the merit of **NEWSPANDA** over the current practice.

Two months into the deployment, the CMU team carried out semi-structured interviews with their WWF colleagues who have been using **NEWSPANDA** outputs in their work. The purpose was to understand how WWF teams liked the toolkit and to elicit possible suggestions for improvement. Some quotes from the interviews are as follows.

> "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."

> "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."

This improvement in news collection also helped with the downstream task – infrastructure impact assessment.

> "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 micro-level improvement in this single task has also resulted in macro-level organizational change:

> "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."

The WWF Nepal team has been putting together weekly news digests for conservation sites in Nepal. Although this dataset is small and has no negative labels, this is the only quantitative comparison between **NEWSPANDA** and current practice we can make. We find that our model is able to identify 62% of the articles in the news digest. This is a relatively good performance as we had extremely limited articles (only 150) about Nepali conservation sites to train the model.

### 8.4 Sustainable Deployment and Broader Impact

Encouraged by the success of **NEWSPANDA** at the initial stages, we are working to scale it to more sites and permanently deploy **NEWSPANDA** as part of the WWF computing infrastructure. We have been collecting news articles for more than 60,000 sites globally and applying our trained model to classify them on a weekly basis since April 2022. We will collect more labeled data to fine-tune the model for different countries or sites and evaluate the effectiveness of our system in a global scale. We are also shifting our system to a cloud server. The server is owned and maintained by the WWF team, rather than the CMU team, to ensure sustainable deployment. The CMU team will continue to provide technical support as well as tutorials to help WWF eventually grow in-house capability of sustaining the project.

Much as this project was a collaboration between WWF and CMU, **NEWSPANDA** could also be valuable to the broader civil society. Thus, we also developed a social media module in the form of a Twitter bot called WILDLIFENEWSINDIA. The bot periodically tweets a selected set of relevant articles. In addition to tweeting links to articles, we also use the keywords from **NEWSPANDA**'s keyword extractor (Section 6.1) to generate salient hashtags. Sample tweets are shown in Figure 5. Currently, WILDLIFENEWSINDIA is focused on conservation-related articles in India. As we continue working on this project, we hope to scale this to a global level, so that any organization or individual interested in conservation can benefit from the tool.

### 8.5 Lessons Learned

This 1.5 year long and counting collaboration has yielded many valuable lessons for both WWF and CMU. We have already mentioned some of those when discussing the data collection, noise reduction, and the pilot study. We would like to highlight two more generalizable lessons below.

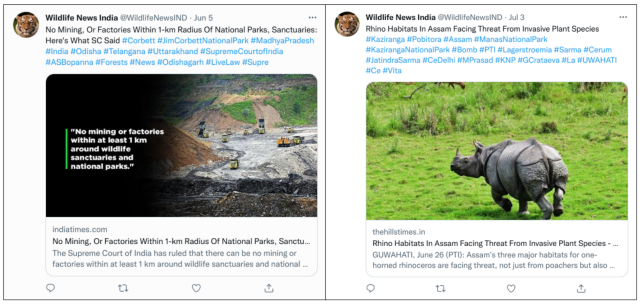



Figure 5: Sample tweets of WILDLIFENEWSINDIA

Problem identification is an iterative process and rapid prototyping helps surface unforeseen needs. The event extractor in Section 6.2 was not initially part of the agenda, because without a prototype of the classification model readily available, it was difficult for WWF to realize what could be done with it. However, after several iterations of communication and exploring the classification results, the need to track the development related to a single project/location became clear to us. This was made possible by the rapid prototyping, where the CMU team used viable algorithms that may not be optimal but are quick to implement to demonstrate the possibilities of the toolkit and the way forward.

It is the various "not-so-AI" components that realize the promise of an AI for nonprofit project on the ground. While the classification module in Section 5 is the engine of **NEWSPANDA**, the postprocessing module in Section 6 and the visualization module in Figure 4 are key in getting the information in a consumable format, and ultimately the buy-in at WWF. Each of the latter two modules requires at least as much engineering effort and careful design as the classification module. We call on future AI for nonprofit projects to pay enough attention to all the infrastructure around the AI part, in order to deliver the real impact that we hoped for.

## 9 Conclusion

In this paper, we designed and deployed **NEWSPANDA**, a toolkit for extracting, classifying, and analyzing articles related to conservation and infrastructure. We showed empirically that our proposed **NEWSPANDA** model classifies better than baseline methods for both *Conservation Relevance* and *Infrastructure Relevance*. We also presented quantitative and qualitative evaluation of our proposed system in the real world as well as its impact on the WWF teams in UK, India, and Nepal.

Currently **NEWSPANDA** mainly focuses on a few countries and we are expanding it to a global scale. But incorporating more conservation sites is just the beginning. To do it right, we also need to cover more languages and more local media sources. This is especially important for the global south. Many high-impact local developments might never reach international news outlets. The ability to capture these local sources, especially if they are not written in English, is something we are currently working on.

## References


Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Boutilier, R.; and Bahr, K. 2020. A Natural Language Processing Approach to Social License Management. *Sustainability*, 12.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *ICLR*.

Dandeniya, D. 2018. An Automatic e-news Article Content Extraction and Classification. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 196–202.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hosseini, K.; and Coll Ardanuy, M. e. a. 2020. Data Study Group Final Report: WWF.

Joshi, K.; N, B.; and Rao, J. 2016. Stock Trend Prediction Using News Sentiment Analysis. *International Journal of Computer Science and Information Technology*, 8: 67–76.

Lisivick, M. 2018. NewsAPI Python Library. https://github.com/mattlisiv/newsapi-python.

Liu, Y.; and Guo, H. 2020. Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Loria, S. 2018. textblob Documentation. *Release 0.15*, 2.

Luccioni, S.; Baylor, E.; and Duchene, N. 2020. Analyzing Sustainability Reports Using Natural Language Processing. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Murray, L. C.; Gupta, N.; Burke, J.; Rupam, R.; and Tshankie, Z. 2019. Matching Land Conflict Events to Government Policies via Machine Learning Models.

Ojokoh, B. 2012. Automated Online News Content Extraction. *International Journal of Computer Science Research and Application*, 2: 2–12.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Reis, D. C.; Golgher, P. B.; Silva, A. S.; and Laender, A. F. 2004. Automatic Web News Extraction Using Tree Edit Distance. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, 502–511. New York, NY, USA: Association for Computing Machinery. ISBN 158113844X.

Santos, B. S.; and Crowder, L. B. 2021. Online News Media Coverage of Sea Turtles and Their Conservation. *BioScience*, 71(3): 305–313.

Tjong Kim Sang, E. F.; and De Meulder, F. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147.

Wu, Y.; Xie, L.; Huang, S.-L.; Li, P.; Yuan, Z.; and Liu, W. 2018. Using social media to strengthen public awareness of wildlife conservation. *Ocean & Coastal Management*, 153: 76–83.