New tool monitors wildlife conservation in low-resource languages

July 17 2024, by Marylee Williams

Credit: Unsplash/CC0 Public Domain

Activists on the front lines of wildlife conservation routinely monitor news articles for information about infrastructure projects that could threaten at-risk animals. But that monitoring required more staff time

than organizations on the ground could spare.

Researchers at Carnegie Mellon University helped ease this burden by working with the World Wildlife Fund (WWF) for Nature to develop a tool that monitors and identifies media articles related to environmental conservation.

Once a week, WWF India needed two full-time workers to monitor news and identify issues related to wildlife conservation, said Fei Fang, an associate professor in the Software and Societal Systems Department (S3D) at Carnegie Mellon University's School of Computer Science.

CMU researchers worked with the WWF to develop media-monitoring tools that allow staff to spend less time analyzing news about infrastructure and environmental conservation and more time advocating for and protecting wildlife.

The tools have been expanded to include media monitoring in low-resourced languages like Hindi and Nepali to gather news from communities where wildlife is especially at risk.

"We are trying to identify the news articles relevant to environmental conservation in a timely fashion for multiple languages and especially for those low-resource languages where we don't have a lot of label data," Fang said.

Fang deployed her first model, NewsPanda, in the United Kingdom, India and Nepal in 2022. On a weekly basis, the toolkit automatically detected and analyzed news and government articles written in English describing threats to conservation areas.

A pretrained large language model (LLM) classified the articles as relevant to conservation and infrastructure. The NewsPanda team

created their dataset with WWF Nepal and India, labeling more than 1,000 articles. Along with scraping and analyzing the articles, NewsPanda also placed them on a map and created a bot to share articles via social media.

Workers at WWF who used NewsPanda asked Fang if her team could update this tool for articles written in local languages, like Hindi and Nepali. But staff at these organizations did not want to label 1,000 articles again to create the training data needed for these languages.

Fang said her research team needed to find a more efficient way of assisting with local media monitoring. She reached out to Lei Li, an assistant professor in CMU's Language Technologies Institute (LTI) who works on multilingual natural language processing.

"Where the text classification and information extraction technology is right now, natural language processing tools work well for high-resource languages—like English, Spanish, German, French and Chinese—because you need labeled data to do supervised training," Li said.

"Once you want to add a new language where you don't have the annotated data, it doesn't work well. This is the exact problem we are trying to solve. We are trying to understand the text of these articles and extract the most important information in another language without much human-labeled data."

WWF Nepal agreed to help the research team develop this tool. Initially, the CMU research team tried commercially available machine translation tools, but it wasn't producing quality translations from English to Nepali. So researchers created NewsSerow, a news monitoring system that uses an LLM to summarize and classify articles written in Nepali. The tool is named after a serow, an animal found in Nepal.

The technology used to create NewsSerow isn't novel, but how the tools are put together is powerful, Fang said. NewsSerow has three modules: summarization, classification and reflection. Summarization uses GPT-3.5 turbo, an LLM similar to OpenAI's ChatGPT, to summarize the information in the article in three sentences in a particular language, like Nepali.

Then, using the article's title and summary, the text is classified as relevant or not relevant to conservation with an explanation about this classification. Researchers used in-context learning in the LLM to develop the classification module.

They provided 10 examples, which included the title, summary, classification label and an explanation of the articles provided by an expert in the area. The process meant staff at WWF Nepal didn't have to label more than 1,000 articles, they just had to label 10.

Finally, NewsSerow performs a reflection, which double-checks if the tool's relevancy classification is accurate. The reflection module is optional, and researchers added it to decrease the number of false positives.

Researchers found NewsSerow performed comparably to other news summarization and classification models that required much more training data.

"That's exactly what we want to achieve. We want this workflow we built for NewsSerow to be used for other low-resource languages," Fang said "It's difficult when you want to establish a tool for a new language, but a domain expert is asked to label 300, 500 or 1,000 articles for us. It's not that hard to ask them to label 10. That's doable."

Researchers are working with WWF India to expand this tool to work on

media monitoring in Hindi and other languages, and to expand to other sources such as social media.

A paper detailing the system is available on the *arXiv* preprint server.

More information: Sameer Jain et al, Where It Really Matters: Few-Shot Environmental Conservation Media Monitoring for Low-Resource Languages, *arXiv* (2024). DOI: 10.48550/arxiv.2402.11818

Provided by Carnegie Mellon University