Case study

Can machine learning help us measure the trustworthiness of news?

Quality, fact-based news—and trust between citizens and journalists—helps people make informed decisions about important issues. We tested whether machine learning can help us catch news articles that contain journalists' own opinions and biases.

IREX

Case study

Can machine learning help us measure the trustworthiness of news?

Samhir Vasdev
Technical Advisor for Digital Development
September 2018

IREX
Center for Applied Learning and Impact
1275 K Street, NW, Suite 600
Washington DC, 20036
www.irex.org

All photographs in this report highlight activities and participants from the Mozambique Media Strengthening Program (MSP), where this case study was implemented.

Acknowledgments

The author owes a debt of gratitude to the team of the Media Strengthening Program (MSP, funded by USAID) in Mozambique—particularly its Monitoring and Evaluation (M&E) Manager Alexandre Gavaza, whose tireless and enthusiastic commitment to this experiment made it possible. Thanks also to MSP M&E Senior Assistants Hassane Ibrahimo and Ercilia Justino, who worked closely with the machine learning software to train and test it throughout the experiment, and to MSP Chief of Party Arild Drivdal, who provided the space and support for this experiment to thrive.

This effort was funded by IREX through its Center for Applied Learning and Impact. Technical advisors Manisha Aryal provided invaluable editorial and conceptual clarity, Tara Susman-Peña media context and report feedback, and Charles Guedenet media evaluation expertise. Misha Mirny, IREX's Director for Information & Media, contributed direction, enthusiasm, and resources to put this idea into practice. Alex Cole and Josh Tong, in IREX's strategic communications team, provided guidance on framing and publishing this final product. Nicholas Benequista, Research Manager and Editor at the Center for International Media Assistance (CIMA), also shared essential feedback that informed the content and visuals in this report.

The team from Lore.Ai, the machine learning technology partner, was key to this experiment's success. Hedeer El-Showk spent countless hours testing software, training team members, analyzing and cleaning data, and explaining complex concepts to various stakeholders. He also contributed meaningfully to various sections of this report. His colleague Neeran Saraf provided guidance and context throughout to keep the project on track, as well as feedback, copy editing, and proofreading that shaped this report. Together, they exemplified the ideal technology partner, bringing humility, patience, dedication, and deep expertise to bear in this unique collaboration.

This effort was led by Samhir Vasdev, IREX's technical advisor for digital development, who designed the experiment and developed this case study.

Contents

Boxes

Key Findings

- IREX partnered with Lore.Ai to test whether machine learning software can automatically detect news articles that contains journalists' own opinions. This matters because impartial, fact-based news is a powerful indicator for the quality of media and the vibrancy of an information ecosystem.
- A team of professional media evaluators trained machine learning software to find examples of news articles that contain opinions from a body of **over 1,200 online Mozambican news articles**.
- The software identified articles that contained opinions with **95% accuracy**. This accuracy was achieved after only 16 rounds of training the software, and anecdotes from the team suggest that the software's accuracy noticeably **improved after only about 20 minutes** of "training".
- The results have promising **implications to improve efficiency, scale, and consistency** of traditionally manual and time-consuming media monitoring efforts, such as helping projects target resources more effectively to support journalists whose articles are flagged by the software.
- The process also surfaced **valuable lessons about limitations** of applying machine learning to monitoring, evaluation, and learning (MEL) in global development contexts, such as reinforcing human bias or the need to invest in indigenous machine learning talent to apply these tools sustainably.

The experiment was implemented in Mozambique, where IREX's Media Strengthening Program (MSP, funded by the United States Agency for International Development) supports Mozambican professional and community journalists and their media platforms to provide high quality information to citizens.

OVERVIEW

Can machine learning help us measure the trustworthiness of news?

Quality, fact-based news—and trust between citizens and journalists—is essential to helping people make informed decisions about important issues. Traditional methods to evaluate media content are resource-intensive and time-consuming, so we tested whether machine learning can help us catch news articles that contain journalists' own opinions and biases.

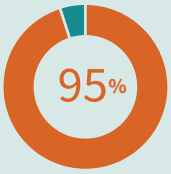
THE EXPERIMENT

Load 1,200 online news articles into machine learning software.

Show it examples of articles that contain opinionated content.

Verify and correct the software's suggestions.

KEY FINDINGS



The software found opinionated articles with a **95% accuracy** rate.

The software began finding opinions after seeing only **20 examples**.

Accuracy increased over time, despite **human error and bias**.

LESSONS & LIMITATIONS

This experiment tested only one of IREX's 18 indicators of media quality (which is whether the author inserts their own opinion into articles). Others, like citing a variety of reliable sources, are not as easy to automate.

Like many machine learning applications, human bias is codified into the software. Measuring media quality can be a subjective exercise. This software doesn't eliminate bias, but it does apply it more consistently.

More research and experimentation is necessary. Machine learning can help us spend resources more efficiently, but more exposure to the technology is needed to realize its potential appropriately and responsibly.

This partnership between **IREX** and **Lore.AI** was supported by IREX's Center for Applied Learning and Impact and tested in the Mozambique Media Strengthening Program (MSP), funded by the United States Agency of International Development.

This infographic accompanies a report containing more details, located at **www.irex.org/measuringnews**.
Visit www.irex.org or contact Samhir Vasdev (svasdev@irex.org) for more information.

Introduction

USAID's 2018 report Reflecting the Past, Shaping the Future: Making AI Work for International Development provides an important foundation for development practitioners who are considering applying machine learning in their work. Among its conclusions is the need to "actively investigate the appropriate use" of new tools like machine learning, " understanding their powers and limitations across contexts and geographies if we hope to effectively leverage them in our work".[1]

Responding to that need, this report describes the process and outcomes of an experiment conducted by IREX, a global development and education organization, and Lore.Ai, a machine learning firm. The experiment tests the feasibility of using machine learning to automatically evaluate the quality of media content in Mozambique.

Its audience includes media support practitioners who are interested in leveraging innovative digital tools to amplify their work, as well as non-governmental oganizations (NGOs) and global development organizations more broadly who are grappling with practical ways to engage with this technology and apply it meaningfully and responsibly to their work.

[1] https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf p.75

Context

Quality, fact-based, impartial journalism drives vibrant information systems.[2] It builds trust with citizens, holds leaders accountable for their actions, and lays the foundation for meaningful, informed debate that strengthens democracy and propels societies forward. It also provides a bedrock amidst the waves of misinformation that cloud judgment, reinforce prejudice, and mislead citizens and leaders. Quality journalism is in high demand; despite what some may expect, audiences are hungry for quality information when it matters.[3]

But as the importance of quality, impartial media persists, and the sheer volume of content increases exponentially, the need to measure that quality—in other words, to evaluate whether journalists cite reliable sources, avoid bias reporting, embrace impartiality, and other characteristics common in journalistic codes of ethics—is more critical than ever.[4] Taking stock of how media meets or falls short of these criteria offers various advantages. It helps to diagnose and prioritize media support efforts and to track their impact, and it can give consumers a sense of which media outlets are more reliable and trustworthy. And since more people today discover news through computer algorithms than through human editors, knowing which news is most impartial can help those algorithms deliver more fact-based information.[5]

Despite its importance, measuring the quality of media content is challenging for a number of reasons. Often, trained human evaluators need to manually find, index, and read scores of news stories sourced from multiple different outlets, evaluating them according to different indicators of media quality. This is the process currently used for IREX's **Media Content Analysis Tool** (MCAT), a framework for

Box 1: What is the MCAT?

The Media Content Analysis Tool applies content analysis, a well-established evaluation methodology, to systematically score selected text against 18 well-defined indicators of media quality. Scores are calculated for each story and averaged across a representative sample of stories over a period of time.

The results are useful not only to monitor change in the quality and content of media produced, but also to diagnose capacity weaknesses and to ultimately address them.

[2] Learn more at https://www.irex.org/sites/default/files/node/resource/vibrant-information-paper.pdf
[3] https://www.cima.ned.org/blog/audiences-worldwide-hungry-quality-news-actively-search-matters
[4] See for example https://www.spj.org/pdf/spj-code-of-ethics.pdf
[5] https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf p.15

evaluating media content quality across dozens of indicators. The framework has been tested and used for decades in multiple countries.[6] When evaluators are sufficiently trained, the MACT yields reliable insight into the trustworthiness of news, but it can be a slow, inefficient, and expensive process (see the "Problem" section).

In Mozambique, through its Media Strengthening Program (MSP), IREX has been supporting media development efforts for several years. Among many other activities, MSP includes a team of evaluators who have manually clipped, read, and evaluated **thousands of news articles** over the years as part of the MCAT process. This team of trained evaluators reads each article and scores it on 18 indicators that together provide a picture of the overall quality of the article. These indicators sit within four categories: whether and how the article cites sources; the article's relevance and newsworthiness; its structure; and its impartiality.[7]

Since more peple today discover news through computer algorithms than through human editors, knowing which news is most impartial can help those algorithms deliver more fact-based information.

Programs like the MidiaLab, incubated by the Media Strengthening Program in Mozambique, provide support to journalists like these on tools and techniques to become leaders in their field.

The actual evaluation process is tracked in an Excel spreadsheet. Armed with MCAT analyses of thousands of news articles, the MSP team can evaluate the effectiveness of its efforts to train journalists. The MCAT also empowers other stakeholders like USAID, who funds the Mozambique program, to keep a pulse on how well the media in the country covers various sectors. The data can also be useful to any media support practitioner seeking to identify gaps for further improvement. For instance, if the MCAT reveals that articles about the economy at a particular news outlet are of better quality than articles about health and nutrition, resources could be reallocated for deeper journalist training at its health and nutrition desk.

Box 2: The urgency of fact-based journalism

Propaganda and disinformation are as old as the news media itself, but in recent years they have shaped geopolitical and social currents in powerful ways, and concepts like "fake news" have entered the global public consciousness. This trend accompanies growing attacks on press freedom—even in stable democracies—and waning trust in news media.

It is critical that citizens, media practitioners, and other concerned stakeholders are equipped with the tools and skills to discern fact from fiction in the news they produce and consume. Increasing people's understanding of the information they live with and the information that they actually need can help increase demand for factual information.

Better understanding of our information systems and how to navigate them play a critical role in developing vibrant information systems and more just and prosperous societies.

[6] Defining "quality" of journalism is no easy task, in part because it varies depending on context (for instance, the MCAT's indicators vary according to different countries). However, our definition of "quality" is explained via the 18 MCAT indicators in Mozambique. See Annex 1 for details.

[7] These indicators vary based on the country where MCAT is conducted and are founded in the basic journalistic principles that can be perceived when consuming content. See Annex 1 for a list of all 18 MCAT indicators in Mozambique, where this experiment was conducted.

Problem

Although the MCAT is an effective methodology that has been replicated in other country contexts, it still has its limitations. Manually monitoring and evaluating news articles across a growing spectrum of online and offline sources is a **time-consuming and expensive process**. In the case of the Mozambique MCAT, a team of two evaluators alone is responsible for tracking and evaluating about 25 news articles every day.

This means that, spending between 10 and 25 minutes on each article, each evaluator spends an average of **nearly four hours every day** on this task. In any under-resourced context, losing just one evaluator (as was the case during this pilot, when one evaluator left the team) can lead to a significant backlog, especially considering the resources that must be re-invested to train up a new evaluator.

These conditions result in slower evaluations, which ultimately limits the depth of the MCAT analysis. It also limits the team's ability to be nimble and responsive to changing contexts (for instance, in the days after a contentious election, the team might want to evaluate the quality of election coverage more quickly than their resources permit).

Another consideration is the issue of **inconsistency**. Although evaluators are trained to the same high standards, human factors can limit the consistency of how they evaluate media content. For instance, one MCAT indicator requires evaluators to confirm whether statements in a news article are "supported by evidence." An audit randomly selected seven evaluated news articles (the team had already conducted an MCAT on these) and asked the evaluators to re-code those articles. The Mozambique MSP team found that, in this second pass for that specific indicator, 4 of those 7 articles were coded differently than their original scores. This points to a limitation in the MCAT methodology that could

be addressed to improve the quality of media content analysis by making that analysis more consistent.

NO CONTINENTE AFRICANO

Uma em cada três mulheres utiliza contraceptivos modernos

HISTÓRIAS DE VIDA

Conduzir autocarro é sonho de infância

SOCIEDADE

Aleitamento materno exclusivo até aos seis meses de vida cobre apenas metade das crianças

As taxas de aleitamento materno exclusivo continuam baixas no país. Só metade das crianças é que são alimentadas apenas de leite materno durante os primeiros seis meses de vida

ALEITAMENTO REDUZ MORTALIDADE INFANTIL

6 MESES

2 ANOS

EM MASSINGA

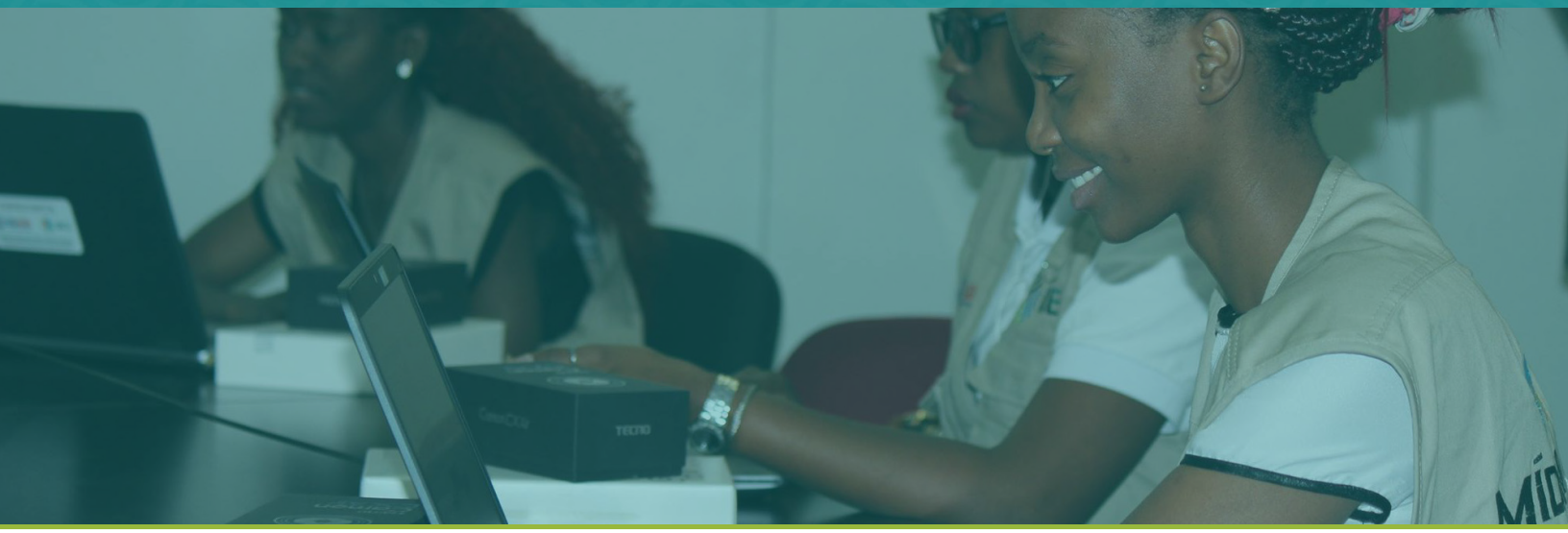
Níveis de seroprevalência preocupam sector da Saúde

Tuberculose associada ao HIV/SIDA

Quatro funcionários da Saúde punidos por maus tratos em Malema

The MCAT evaluators clipped, scanned, and reviews thousands of articles like these from Mozambican media outlets to monitor the quality of news journalism in the country. Crucially, this massive body of evaluated media content can be used to check a machine learning software's accuracy against real evaluations. Rigorous and large "training datasets", as these are called, are one of the **main limiting factors** for machine learning efforts. Media support initiatives like MSP can partner with technologists to provide these important datasets.

Both of these challenges—**inefficiencies** and **inconsistencies**—ultimately impede MCAT's utility for evaluating, learning from, and improving journalists' work. This has direct consequences for the quality of media content at a time when impartial, fact-based reporting is needed now more than ever.



Opportunity

The challenges associated with manually evaluating the quality of media content in Mozambique inspired a simple question: **How could machine learning technologies make evaluating media content more time-efficient, consistent, and scalable?**[8]

The opportunities presented by machine learning are clear. As USAID recently reported, machine learning shows "tremendous potential for helping to achieve sustainable development objectives globally," including by improving efficiencies or providing new insights that can amplify the impact of global development programs.[9]

For instance, in the media support sector, an NGO could have a real-time window of the quality of the hundreds of news articles that its trained journalists were publishing every week. This could help the NGO measure anything from the effectiveness of its training efforts in a particular program to the health and vibrancy of an information system. Other potential use cases about leveraging machine learning to amplify media support efforts are in the "What's Next?" section.

The opportunities provided by machine learning include purposes beyond supporting and strengthening media. The underlying concept of the idea behind this experiment—training a computer to automatically detect characteristics about media content—applies in many other contexts, including life-saving ones. For example, the 2014 Ebola crisis in west Africa—the most widespread outbreak of the deadly virus in history—was exacerbated by reluctance among affected populations to seek treatment. According to USAID, the outbreak "was driven as much by misinformation,"—such as rumors that bleach sprayed by health workers

Box 3: What is machine learning, and why use it?

Machine learning allows computers to find patterns in data and use those patterns to make predictions.

Enabled by advances in everyday computing power that allows rapid analysis of a lot of information, it can recognize patterns across large swaths of data.

This report discusses technical terms related to machine learning using fuzzy and informal definitions. For more context and details about key terms, consider reviewing USAID's 2018 report about machine learning in global development: https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf.

[8] Machine learning is an umbrella term for a discipline that combines conventional statistical methods, like regressions, with advanced computing power to learn about, model, and predict behavior in the real world.

[9] https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf, p.4

Team member Alexandre Gavaza, the Monitoring and Evaluation Specialist for IREX's Media Strengthening Program in Mozambique, and two valuators train machine learning software to automatically find opinions in news articles.

was the actual cause of the disease—as it was by other factors like weak health systems."[10] Being able to track rumors such as these in near real-time as they spread across widely used social media and news platforms could give responders a more accurate picture of the problem and help them develop holistic emergency responses that include fact-based counter-narratives.

The opportunity to use machine learning specifically to support and strengthen media remains relatively underdeveloped. For example, despite emerging methods to apply artificial intelligence in the fight against fake news (see Box 4), a lack of training data—that is, examples of fake news evaluated and tagged by professionals for the software to learn from—means that even some of the best AI models are only 65% accurate.[11] IREX's MCAT is therefore an invaluable asset through which to incubate this test, as it offers not only a rigorous methodology for evaluating media quality but also a training dataset of thousands of articles.

If we are to take full advantage of the promises of this technology, we as practitioners need more practical experience and exposure to machine learning—including understanding its limitations. For instance, we know that measuring the quality of media content necessarily involves some level of human bias, but how

Box 4: AI in the fight against misinformation

Frontline journalists, activists, and civic technologists have developed various tools that test how machine learning can help tackle fake news.

For example, Chequeabot (http://chequeado.com/automatizacion/) automatically identifies claims in the media and matches them with existing fact checks, to tell readers how reliable a news article is. Other tools, like FakerFact (https://www.fakerfact.org/) and Deep News (https://www.deepnews.ai/) use machine learning for similar outcomes.

Tools like Chequeabot help human fact checkers prioritize which articles, out of thousands, their team should focus on during the week. This affirms our finding that machine learning tools can help us to prioritize the time and energy of our teams more efficiently, leading to cost savings and impact.

A lack of training data means that some of the best AI models to catch fake news are only 65% accurate. IREX's MCAT offers a training set of thousands of articles evaluated over the years by trained professionals.

[10] See https://blog.usaid.gov/2016/09/tracking-rumors-to-contain-disease-the-case-of-deysay-in-liberias-ebola-outbreak/

[11] See https://medium.com/mit-technology-review/even-the-best-ai-for-spotting-fake-news-is-still-terrible-5afe0f026d94

will automating that process inadvertently further entrench those biases? (This question and others are explored more in the Limitations and Lessons section). As USAID frames it, we as development practitioners have the "responsibility" to understand and influence how these technologies are applied and how they influence the communities we work in.[12]

Journalists participating in the Media Strengthening Program in Mozambique build skills on producing fact-based, impartial, quality news. Machine learning offers the chance to measure the quality of these products at a scale and speed that exceeds what human evaluators could do.

This report responds to both of these opportunities: testing whether machine learning can improve IREX's efforts to support quality media, while also providing a concrete case study for other stakeholders of how machine learning can be applied to our work. Our hope is that this experiment contributes to a growing global commons of shared experiences and good practices among development actors about promises and pitfalls of working with machine learning tools in practice.

Reading and coding articles is indispensable for the MSP program but demands so many hours. With a machine learning tool, coders would be able to focus on other activities, ranging from administrative to field work.

ALEXANDRE GAVAZA, M&E MANAGER

[12] https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf, p.4

Experiment

This illustration depicts a simplified version of the experiment. The process was not as linear as this image might suggest. For instance, defining the problem statement required some back-and-forth between different stakeholders. Step 5 (evaluating results) happened many times, often on phone calls between the technology partner and the MCAT evaluators who trained the software in Mozambique.

Background

Our experiment was a **minimum viable prototype** (MVP), which is a technique used to help teams collect the maximum amount of learning with the least effort, often in the form of testing or validating a concept. The conclusions are often just enough to validate or invalidate further investment. Framing this experiment as an MVP served a simple purpose: resources at any NGO are limited, so it's neither realistic nor smart to invest heavily in a machine learning experiment like this without deeper knowledge of its potential and limitations. Rather, in keeping with our commitment to agile development, IREX's approach was to develop an MVP to validate a hypothesis, and then continue investing if the initial results are positive and promising. This report is about this MVP.

The actual structure of the MVP was co-designed in a collaborative way between IREX and Lore. This meant that the specific problem statement and scope of the experiment was defined together, which helped both organizations develop a deeper technical understanding of the issues while building shared ownership over the experiment and its desired outcomes. Crucially, designing the scope of the MVP together also helped IREX to understand the limitations of the technology, demystify misconceptions about its potential, and manage expectations.

Framing this experiment as an MVP (minimum viable prototype) served a simple purpose: resources at any NGO are limited, so it's neither realistic nor smart to invest heavily in a machine learning experiment like this without first testing its potential and limitations.

IREX's MCAT measures media quality in Mozambique along 18 indicators. For this experiment, the team narrowed down on one of them to test: whether reporters insert their own opinions into news articles.

Narrowing down to a specific problem to test

Over a period of six months, the MCAT evaluators trained Lore's machine learning software, called Salient, to detect opinions in the text of news articles. This test—whether the reporter inserts his or her own opinion into a news article—is one of the 18 indicators of media quality that the MCAT measures (see Annex 1 for more details about the MCAT). Identifying an opinion in a news article involves reading the article and looking for subjective language or first-person perspective; often these phrases are associated with opinions that don't belong in fact-based news reporting.[13]

Choosing to teset only one of the 18 MCAT indicators in this MVP was not an easy decision, since it would narrow the scope of our results (we would only be testing whether one indicator of media quality could be automatically measured, instead media quality as a whole). Nevertheless, narrowing down to one indicator was key to making this MVP manageable and feasible, while still honoring the spirit and vision of this exercise. By proving this small step was possible, we would lay the foundation to build upon this pilot later, potentially expanding to other MCAT categories.

Developing the dataset

Once we understood which part of the MCAT we would be applying machine learning to, it was time to use the software. To do this, the team automatically

Box 5: What does an opinion look like?

Although buzzwords like "I believe..." or "I think..." help coders identify opinions in news articles, the distinction, often, is less clear. Here are some examples of sentences containing opinions that the team used to train Salient (translated by the Mozambique team from their original Portuguese texts):

- "It already surpassed the simple level of "tightening the belt" and now literally goes to tightening citizen's neck…"
- "Contrary to what citizens see on their dinner plates, the Frelimo government says that the [economic] balance is positive and growing."

These examples provide a glimpse into the varied language and contextual hints that Salient needed to "learn" in order for this experiment to be a success.

[13] Of course, opinions have an important place in journalism, such as in the form of editorials. This experiment only searched articles that are meant to be fact-based, like news reports.

scanned the websites of nine leading print media outlets and imported over 1,200 online articles about topics relevant to the objectives of IREX's Media Strengthening Program in the country, ranging from health and nutrition to transparency and governance.[14] The articles were in Portuguese, which is the native language of the Mozambican MSP team (this is important, since not all languages and scripts are supported by all machine learning software).[15]

In this screenshot of the Salient interface, coders have highlighted sentences (in red) that reflect an opinion in a news article. The presence of journalists' opinions in objective news articles is one indicator of bias and poor impartiality.

Training the software

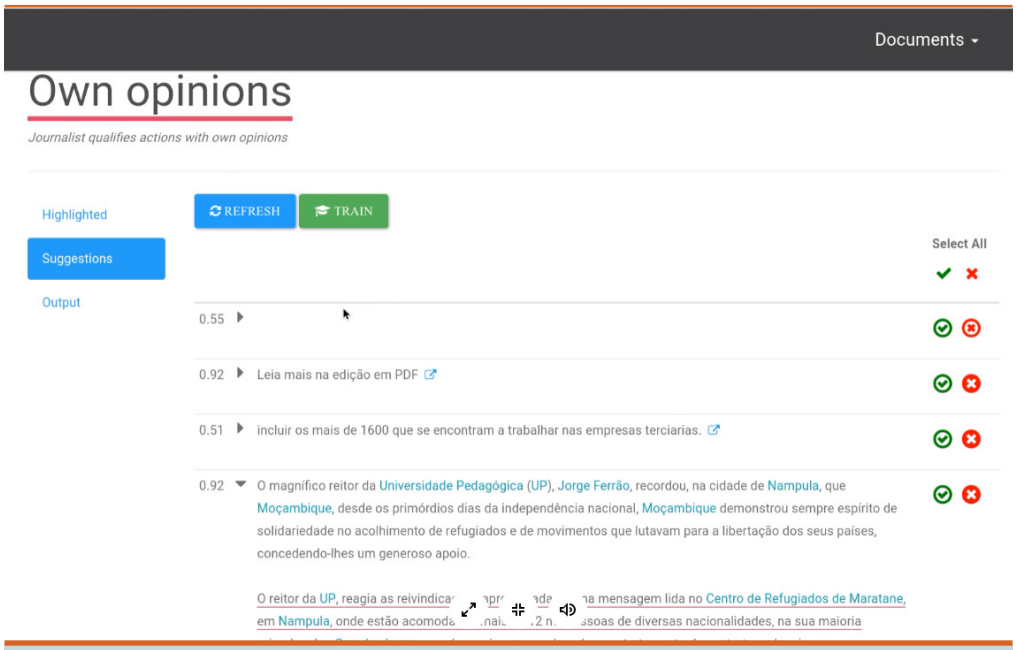
Once the articles were loaded into Salient for analysis, the evaluators began training the tool to identify opinions in the text. To find examples of opinionated text in the articles, evaluators would search for keywords and phrases (like "I believe" or "I think") using Salient's search engine. Then, they would review the articles in the search results to find examples of opinions in the text. Using Salient's "highlighter" tool, they clicked on sentences in the article to show Salient examples of where opinions are present. In the background, Salient would learn

[14] In total, 1,229 articles were ingested into Salient's platform. 21 of these were manually added by the MSP team as scanned articles, while the rest were scraped directly from the online websites of the media sources. More information is found in Annex 2.

[15] Not all of the 1,229 articles were used to train Salient. Rather, the software learns from the highlighted examples submitted by the user (as explained later in this section). The 1,200 articles were simply a starting set used for analysis in this project.

to find similar examples of opinions in the text, based on what evaluators were tagging with the highlighter feature. This process took about one minute per article.

In parallel, evaluators could interact with the tool using a training loop feature. In this mode, Salient would present evaluators with a sentence that it thought was similar to those highlights—that is, sentences that the software thought also contained opinions. The evaluators would review these sentences and confirm that the suggestion was either correct (that there was indeed an opinion, making this a "true positive" result) or incorrect (that Salient thought it contained an opinion but actually it did not, making this a "false positive" result). Overall, the team conducted this feedback loop 51 times.[16]



After a few rounds of training, the software begins to suggest sentences that it believes contain opinions. Coders click the green or red icons to give feedback to the system.

With each round of feedback, the team captured data about Salient's performance, tracking how it changed over time. This performance data included several statistical measures including accuracy, precision, and recall. The next section explores these results.

Box 6: How did the software pick what to get evaluators' feedback on?

Salient typically shows suggestions that are somewhere near the "decision boundary"—that is, examples it isn't quite sure are positive or negative. If it concludes that something is either positive or negative, it usually won't show it.

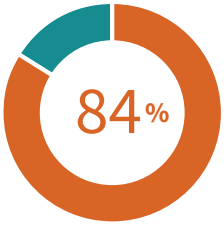
The reason for this is, except in rare cases, asking users to manually correct the software's suggestion is asking them to do something outside their normal workflow. So, to minimize the burden, we want to extract as much information as possible each time they give feedback.

Examples that the software is least sure of are the ones it needs the most help with, so and having the users tell it one way or the other gives it the most information possible.

[16] This same process was also repeated with "negative" samples—that is, articles with sentences that the software thought did not include opinions. See Box 6 to learn more about how Salient chose which sentences to present to evaluators for their feedback.

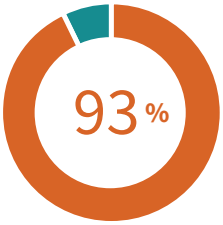
Results

Precision



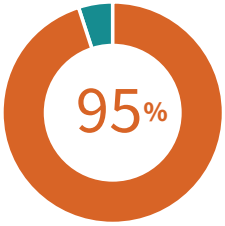
Percent of articles identified as opinionated that actually are.

Recall



Percent of all opinionated articles that the software identified as opinionated.

Accuracy



A type of average between precision and recall.

The software achieved these results after only **16 rounds of training**.

The software began finding opinions after seeing only **20 examples**.

Accuracy increased over time, despite **human error and bias**.

Important note: The accuracy rate (95%) is from the *best-performing* model of many models that the software used in the final round of training. The precision and recall rates are from the *average* of all models that the software used in the final round. This helps explain why the accuracy (95%)—which is a type of average of precision and recall—is greater than precision (84%) and recall (93%): they are from different models.

This experiment tested whether software can automatically evaluate the impartiality of online news articles, specifically by identifying opinions in news articles. **The results prove that it is feasible for machine learning to help us find opinions in news articles**. More importantly, the software identified these articles reliably enough to apply it to monitoring and evaluating media content, at a **scale and speed** that far outpace conventional human evaluators (a matter of seconds, instead of minutes or hours).

The results also expose some **important limitations**. For example, the software can find opinions in articles, but it needs more training before being able to determine whether those opinions belong to the author or whether they serve a newsworthy purpose (such as quoting a source or political commentator). More about these considerations are in the Limitations and lessons section.

The following paragraphs dive deeper into these results from a statistical point of view, followed by the implications of these results in the context of a media support program. Although the language in this section is technical, it is also written in accessible language in order to demystify some concepts related to machine learning to non-technical specialists.

Key highlights of our results

- **The software recognized sentences containing opinions within the dataset of 1,200 articles with 95% accuracy**.[17] This means that, 95 out of 100 times it tries, the software can find an article containing an opinion.
- **The precision rate reached 84%.** This means that, if the software thinks that 100 articles contain opinions, 84 of them actually do.
- **The recall rate reached 93%**. This means that out of every 100 articles that actually did contain opinions, the software found 93 of them but missed 7.
- **Accuracy and precision increased the more that the model was trained**. There is a clear relationship between the number of times the evaluators trained the software and the accuracy and precision of the results. The recall results did not improve over time as consistently.
- **The software's ability to "learn" was almost immediately evident**. Anecdotally, the evaluation team noticed a marked improvement in the accuracy of the software's suggestions after showing it only twenty sentences that had opinions. The accuracy, precision, and recall results highlighted above were achieved after only sixteen rounds of training the software.

Next, let's look in more details at the numbers.

Box 7: What results data have we left out, and why?

In most machine learning exercises, the results (like accuracy and precision) of the models against a training set should be close to 1—after all, the models are trained against the very same data that they're being applied to!

In our experiment, we can see several data points where the models perform poorly even against the training set. We believe this is due to human error explained in the "Lessons and Limitations" section.

To make these charts more accessible and readable while maintaining their integrity, we have made some adjustment and "hidden" this human error. Annex 3 contains a table that highlights exactly which results were omitted.

A look at the numbers: accuracy

Chart 1 on the next page explains details about the accuracy results of this experiment. It shows the results of sixteen rounds of training over a two-week

[17] USAID defines "accuracy" this as "the fraction of correct predictions made by a model," (see https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf p.39). We define accuracy using the F1 score of the models, which itself is the harmonic mean of precision and recall (see Box 8).

period.[18] Each dot depicts a statistical measurement called an F1 score (marked on the y-axis), which is essentially a measure of a test's accuracy. The x-axis is the number of "positive samples" the team highlighted in Salient. A positive sample is a single example (a sentence) of opinionated text in any article.

In Chart 1, **blue dots** represent the software's accuracy when it applied its models to the training dataset—that is, the very same news articles that it was trained on by the MCAT evaluators. We would expect this to be 1, or close to 1, since the model is based on those very news articles, so the model should accurately identify them.[19]

Orange dots represent the software's accuracy after applying its models to the test dataset, which is a set of news articles that the software had never before encountered or analyzed (but that human evaluators had seen, in order to compare the model's results against reality). These orange dots demonstrate how well the software performs when trying to find opinions in news articles that it is seeing for the first time.

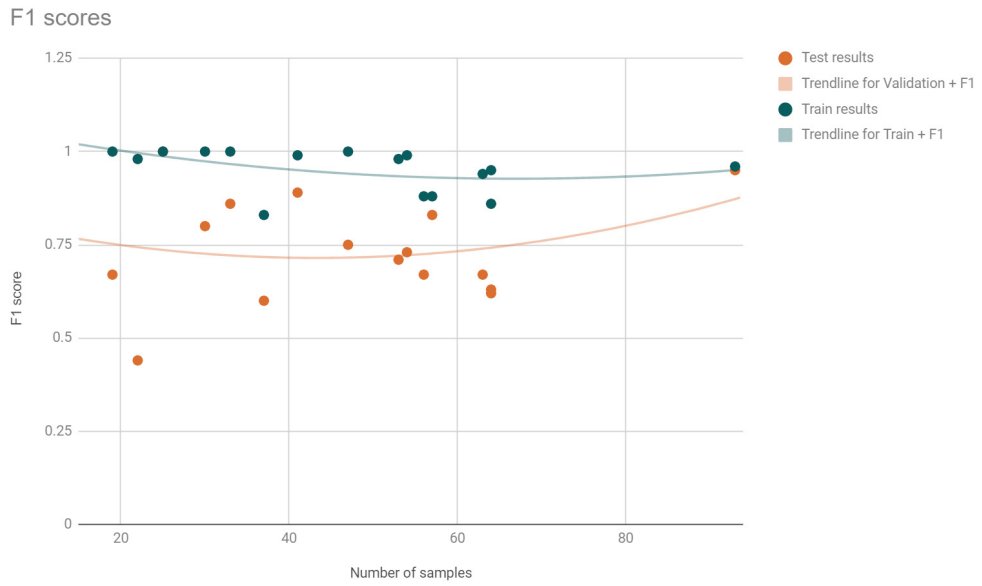


Chart 1: Accuracy. F1 scores (in orange), which indicate how accurately the software can find sentences that contain opinions in news articles it's never seen before, increase as more samples are fed into Salient's system. The F1 accuracy reaches 95% after 93 samples.

The software achieved 95% accuracy, which means that, 95 out of 100 times it tries, the software can correctly find an article containing an opinion. The higher the accuracy, the more confidence we can have with its predictive capabilities.

[18] The team actually trained the software 51 times, but the first 27 times used an earlier version of Salient so their results are omitted. The other 24 times used an updated version of Salient, but 8 of those times are excluded from this analysis due to factors that can be attributed to human error.

[19] The dots that have lower accuracy on the training set are usually a reflection of human error and mistakes (explained in more detail in the Limitations and Lessons section).

A look at the numbers: precision and recall

Accuracy only paints one part of the picture about the software's performance. The tool could be finding examples of opinions that aren't actually opinions, or it might be missing some opinions altogether. That's why we have metrics like **precision** and **recall**. These tell us how many of the articles identified by the software are false positives (articles that the software thinks have opinions but really do not) or false negatives (articles that the software did not think have opinions but really do). Chart 2 visualizes how these are calculated.

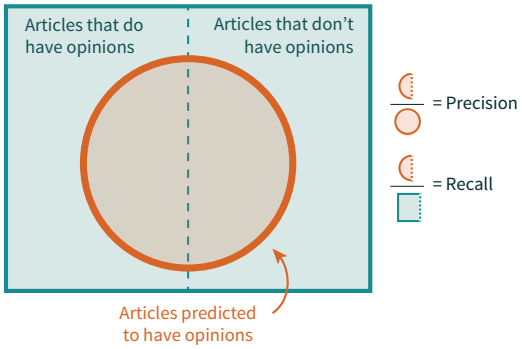


Chart 2: Explaining precision and recall. This chart explains how precision and recall are calculated.

The precision of this model was 84%, which means that 84% of the sentences that Salient believes to contain opinions actually do contain opinions (while 16% do not). Chart 3 below shows the precision results over time.

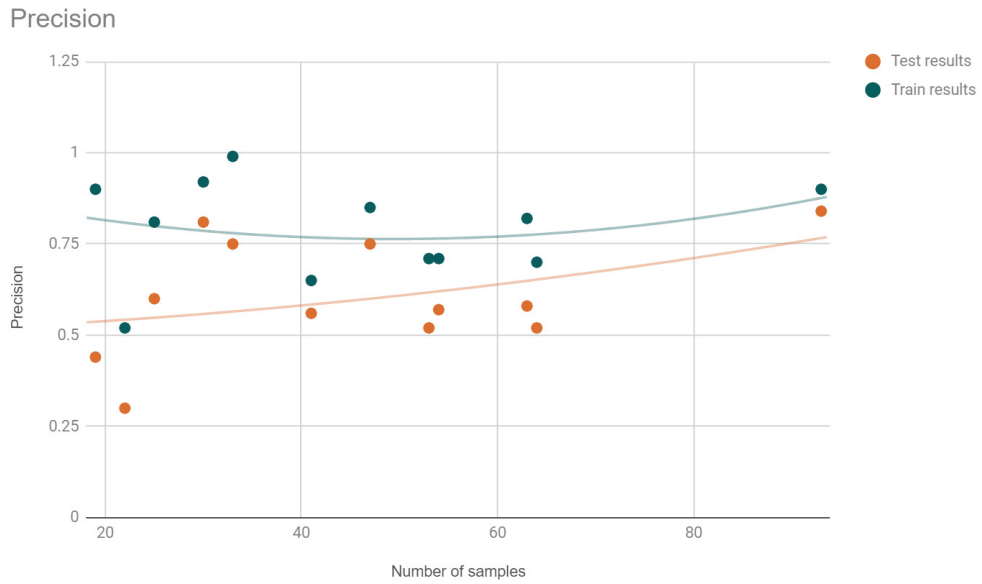


Chart 3: Precision. The software's precision (in orange) indicates how many sentences that the software thinks are opinionated actually do have opinions.

The software achieved 93% recall, which means that out of every 100 articles that actually did contain opinions, Salient found 93 of them but missed 7. However, unlike accuracy and precision metrics, the software didn't exhibit a general upwards trend. Rather, the trendline actually regresses in the middle of the pilot, at one point dropping to 56%. We believe this is due mainly to human error (described in more detail in the Limitations and Lessons section). See Chart 4 for more details.

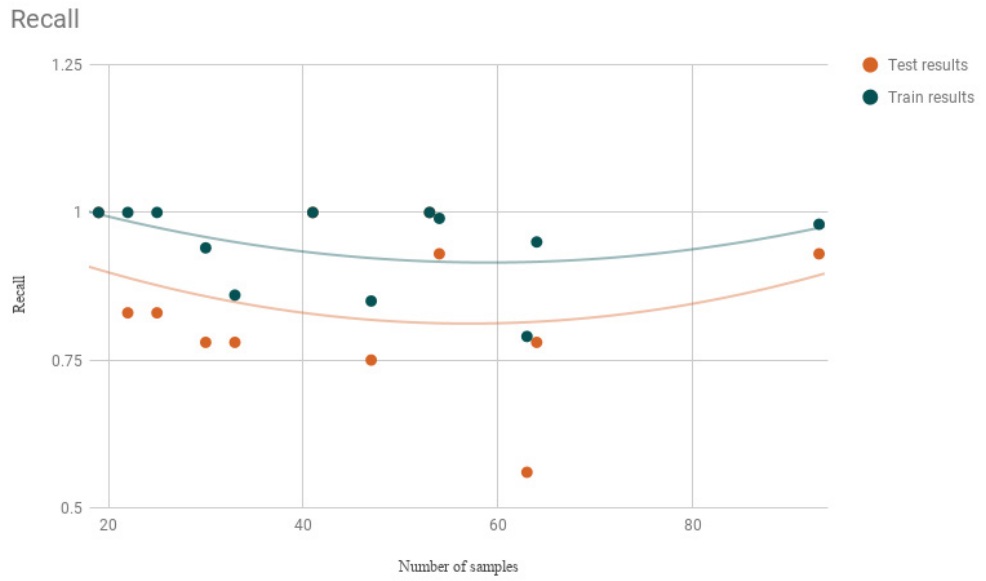


Chart 4: Recall. The software's recall (in orange) compares how many sentences the software found to contain opinions against how many actually did contain opinions. Although recall did not increase as consistently as accuracy and precision, it reached 93%.

Box 8: Should we care more about accuracy, precision, or recall?

As USAID explains, accuracy is "...the fraction of correct predictions made by a model. Accuracy doesn't distinguish between false positives and false negatives, so two models could have the same overall accuracy but make very different types of errors." To help us understand these, we have tools like precision and recall.

Whether you care about precision, recall, accuracy, or other methods of measurement depends on your context. For instance, if you're using a model to predict who might have a particular disease, you can afford to have low precision; this would just lead you to identify some people who don't actually have the disease, but at least you'll also capture those who actually do have it. But you can't afford to have low recall. You want to ensure that, if there are 20 people who have the disease, your model is identifying as many of them as possible.

To help summarize the results, Chart 5 on the next page visualizes the software's performance. In this chart, each dot represents one article. The ratio of dots in the different sections of the chart reflect actual performance of precision and recall (rounded to the nearest ten percent). This visual reveals the scale of the software's accuracy compared to false positives (precision) and false negatives (recall).

Despite the promising results of this experiment, the data also show significant fluctuations in the accuracy, precision, and recall of predictions for the test dataset. It's likely that these fluctuations can be pegged to endogenous and manageable factors like human error, but **more feedback loops and training sessions are needed** to reinforce the conclusions and achieve more consistent accuracy, precision, and recall scores.

Nevertheless, we are confident that this minimum viable prototype has at least validated the opportunity for machine learning to accurately find opinionated content from within a large dataset of online articles. The "What's next?" section highlights some specific ways to build on the initial progress of this MVP.

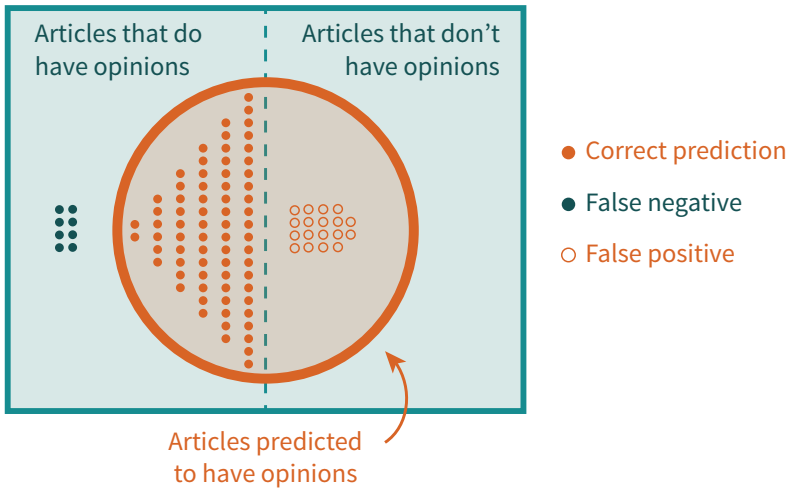


Chart 5: Visualization of results. This chart visualizes the results as false negatives and false positives.

Implications for media support programs

The results of this experiment suggest that software can be trained, even by people with limited technology skills and in a relatively short amount of time, to identify opinions in news articles. But identifying opinions in news articles is only one of the 18 indicators of the MCAT assessment, so it's not yet possible to assert that machine learning can fully transform how media support practitioners monitor and evaluate the impact of their work. That being said, this MVP offers a glimpse into the types of efficiencies that machine learning can bring to media support work.

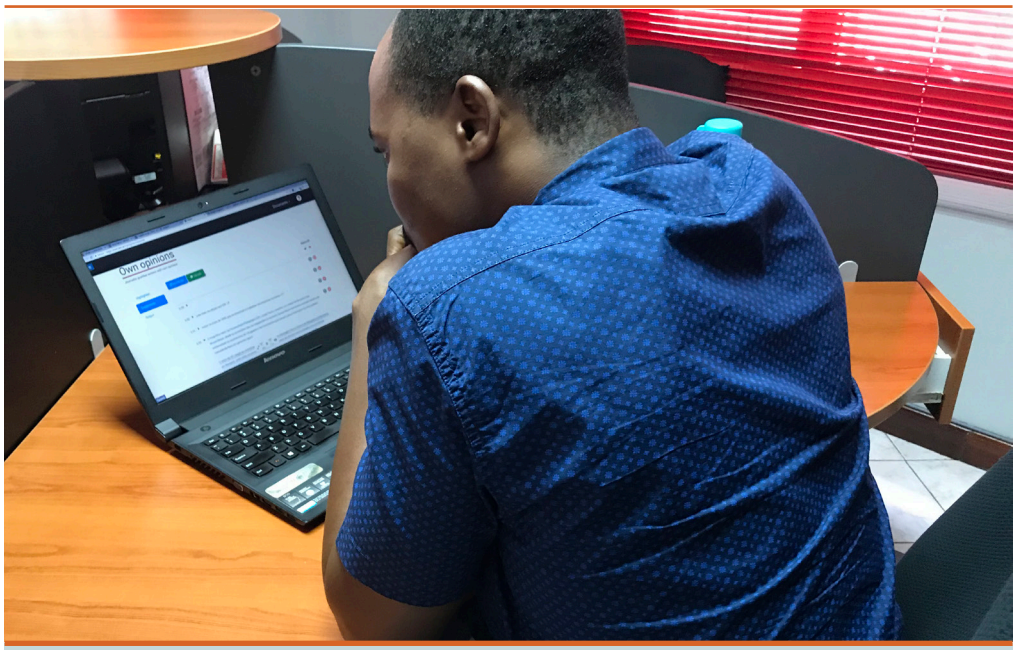
First, machine learning can help practitioners prioritize which information to focus on from a vast amount of data. The conventional MCAT process requires human evaluators to manually read articles to determine their quality. This limits them to a few dozen articles a day, and they might review several news articles before encountering one that requires more attention or analysis.

About 70% of the MSP monitoring and evaluation team's time is spent on activities related to evaluating articles. Automating any pieces of this process will help them spend their time more efficiently.

In contrast, tools like Salient can help them rapidly distill thousands of articles in a matter of minutes, helping them to funnel and prioritize only those articles which need more investigation. This hints at the possibility of increasing the scale of analysis such that it becomes possible to efficiently describe and understand media systems in general. For instance, machine learning could help us explore whether there is a relatively higher ratio of opinionated content in news articles for certain topics in a country or region. Answers to questions like this can transform how media practitioners understand and support robust media ecosystems.

These benefits of scale that come with sifting through vast amounts of data lead to time-saving efficiencies. Analyzing 1,200 articles for their impartiality would be a weeks-long project for a team of evaluators. In fact, **about 70% of the MSP monitoring and evaluation team's time is spent on activities related**

to evaluating articles. After being trained, Salient's models could scan these articles in seconds to help the evaluators focus their resources on more high-value activities, like reviewing only news articles with opinions to identify which journalists need more support and training. This could lead to significant cost savings for a program, allowing teams to spend less time finding issues and more time supporting the journalists who need the most support.



Team member Alexandre Gavaza reviews sentences that the machine learning software thinks contains opinions, giving feedback on each one to improve its algorithm.

A third benefit is around consistency. As explained in the "Problem" section, even well-trained specialists can vary in the way they code stories, and different evaluators might view the same news articles differently. But using a machine learning program can mitigate this risk. Software can be biased (especially when humans encode their bias into a model), but also consistent in how it applies that bias. Once it's properly trained, Salient will interpret and identify specific content the same way over time, devoid of human factors. This can improve the quality and integrity of media content analysis and evaluation efforts.

Limitations and Lessons

Limitations

This promising storyline comes with its fair share of limitations. Here is an overview of some worth considering, followed by more details for each one:

1. Our experiment tested only one of the 18 MCAT indicators of media quality.
2. The software can find opinions, but it can't yet distinguish whether those opinions belong to the author.
3. The results data presented here is a subset of all data.
4. This experiment did not eliminate bias.
5. The news that we evaluated was limited to content in written text format only.
6. Whether or not an article includes an opinion is a relatively minor indicator of impartiality.

Here are more details about each limitation:

1. **Our experiment tested only one of the 18 MCAT indicators of media quality**. Our MVP lacked the scope and resources to evaluate all 18 indicators, so only one—whether authors insert their own opinion into news content—was tested here. Some indicators require substantial contextual knowledge (such as "the author cites reliable sources") which, while potentially achievable using machine learning, would require far more time and effort to explore. Choosing only one of the 18 indicators was a difficult decision and limits our conclusions. We have not proven that machine learning can make evaluating the integrity or quality of media content more efficient, scalable, and consistent writ large. We have, however, demonstrated that the impartiality of news content—measured by whether journalists insert their opinions into news articles that should

be impartial—can be detected using machine learning software.

2. **The software can find opinions, but it can't yet distinguish whether those opinions belong to the author**. The MCAT indicator we measured in this MVP is whether an author inserts his or her own opinion into a news story that is supposed to be unbiased. The evaluators trained the software to identify opinions, but more training is needed to help the software distinguish between whether those opinions are valid (such as when the journalist quotes a political commentator) or invalid (such as when the journalist inserts their own opinion). Although we made progress in this regard during the MVP—such as upgrading the software to recognize when an opinion that it's predicted is embedded within quotation marks (meaning it belongs to someone else and not to the journalist)—more training is needed for the software to recognize when an opinion is a legitimate part of a fact-based news article or when it reflects the journalist's own opinion and bias. Software improvements could also mitigate this, such as telling the software to ignore opinions that occur within italicized or indented passages of text.

More training is needed before the software can recognize whether an opinion belongs to the journalist.

3. **The results data presented here is a subset of all data**. About halfway through our training process, we learned that human evaluators had unintentionally been sending mixed signals to Salient's algorithms (see the Lessons discussion below for more details). This caused the accuracy to fluctuate, despite a clear trend towards increased accuracy before and after these incidents. We are confident that this fluctuation in accuracy is a consequence of manageable and addressable human error, so that data has been excluded in the Results section, in order to offer a more digestible analysis (regardless, all the raw data can be found in Annex 3).
4. **This experiment did not eliminate bias**. A common misconception is that artificial intelligence offers opportunities to analyze something in a neutral and unbiased way. But in reality, through training the software to look for opinions, our team actually encoded its existing bias into the software.[20] This reaffirms the importance of recognizing that machine learning algorithms often codify human bias, rather than reduce it. On the other hand, machine learning codifies not just one but multiple humans biases. As a consequence, if enough varied perspectives are mixed, the resulting net bias can be diluted, unless there is a systemic, similar bias in all of the individual humans who are training the system (which is quite possible).

A common misconception is that artificial intelligence offers opportunities to analyze something in a neutral and unbiased way. But in reality, through training the software to look for opinions, our team actually encoded its existing bias into the software.

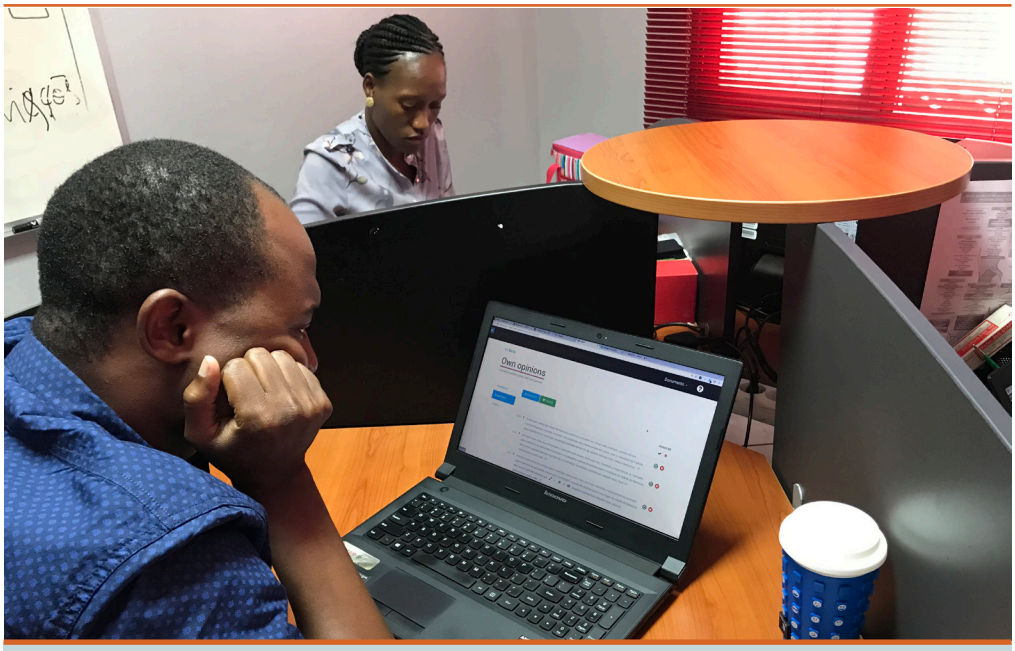
5. **The news that we evaluated was limited to content in written text format only.** All 1,200 articles that contributed to this experiment were originally in text format, either print or online. This means that we have not tested the feasibility of using machine learning to measure impartiality

[20] See https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf p. 36 for a more detailed discussion.

of other types of content, such as podcasts, radio, or television. Although one solution is to convert these audiovisual media into text before ingesting into Salient, this can be time consuming, distort the original meaning, and lose valuable nuance. More experimentation is needed to apply machine learning to analyze non-text news media.

6. **Whether or not an article includes an opinion is a relatively minor indicator of impartiality**. Impartiality is one of four dimensions in the MCAT, and it is defined by three indicators: whether the reporter uses incorrect language, whether all sides of an issue are fairly represented, and whether the reporter inserts their opinion into the story (which is the indicator tested in this experiment). In Mozambique, 7,412 news articles analyzed by the MCAT team fail at least one of these three tests of impartiality. However, nearly 96% of these articles did so because they failed the third test: whether all sides of an issue are fairly represented. Only about 4% of the articles that failed these impartiality tests did so because of the the first two indicators. This means that, even if a machine learning tool were able to automatically identify every article that contained an opinion with 100% accuracy, it would still only be detecting a small fraction of all articles that are not impartial. This simply underscores the need to use a tool like this in concert with other mechanisms, including traditional offline evaluation. But it also highlights the efficiency value of a tool like this: isolating just the 4% of 7,412 articles that contain opinions would be incredibly time consuming using traditional evaluation methods, but nearly instantaneous using machine learning software.

Even if a machine learning tool were able to automatically identify every article that contained an opinion with 100% accuracy, it would still only be detecting a small fraction of all articles that are not impartial, since the presence of the author's own opinions is only one of three indicators of impartiality.



The MCAT evaluators gained first-hand experience in the promises and limitations of ML tools.

Lessons

The entire experiment, which ran for roughly six months, offers several lessons for other NGOs, media organizations, global development practitioners, and monitoring, evaluation, and learning (MEL) specialists who are interested in using machine learning to amplify their work. We have generalized these lessons to apply to the broader global development context, rather than to the media support sector specifically. An overview of each one is shared below, followed by more details.

1. Manage expectations about what machine learning can and can't do.
2. Start with a minimum viable prototype.
3. Define a clear and specific problem statement with the help of subject matter experts.
4. Be mindful of and document human error.
5. Be aware of the risks of vendor lock-in.
6. Build a small team of complementary skills.
7. Choose the right partner.

Machine learning doesn't replace our work, but it can help us work more intelligently and use our limited resources more efficiently.

Here are more details about each of these lessons:

1. **Manage expectations about what machine learning can and can't do**. The hype around artificial intelligence and machine learning can hinder experiments like this. People sometimes expect that machine learning will lead to earth-shattering transformations or new solutions in a project, and while this might be true in some contexts, our experiment suggests that we should approach these technologies with more humility. Some stakeholders, for example, expected that our experiment would automate the entire MCAT—an unrealistic goal given that some of the MCAT's indicators require deep contextual knowledge that even advanced machine learning techniques can not have. Machine learning doesn't replace our work, but it can help us work more intelligently and use our limited resources more efficiently—especially our time which is often a critical and scarce resource. Communicating these messages clearly to key stakeholders, as well as framing this experiment as an MVP (see below) helped to mitigate inflated expectations.
2. **Start with a minimum viable prototype**. The key to getting this pilot off the ground was being cautious about what we were testing. Starting with an MVP—to simply prove the technology is effective—afforded our team leeway to focus on validating the tool before making bold claims about how it might transform a media support program. Approaching the technology with humility and validating an MVP equipped us with the evidence we need to present a stronger case to advocate for more resources to apply this tool to a media support program.

3. **Define a clear and specific problem statement with the help of subject matter experts**. Arriving at a clear problem statement was one of the most challenging and time-consuming aspects of this experiment. This involved taking time to understand the technology and sectoral context (such as interviewing MCAT evaluators), to brainstorm potential problems, and to prioritize and narrow them down. To do this, build a team that includes people who are as "problem-literate" as they are technologically or sectorally proficient. Defining a clear problem statement is a valuable skill set when applying any technology to global development contexts.

Arriving at a clear problem statement was one of the most challenging and time-consuming aspects of this experiment.

4. **Be mindful of and document human error**. After a few rounds of training, we noticed a drop in the accuracy of Salient's predictions. We soon realized a critical issue: sometimes, Salient would present a sentence to the evaluators that it believed to contain an opinion—and it was correct. However, the evaluators noticed that this opinion was actually part of a quote or passage referenced by the article's author (such as a citation by political commentator), and so they rejected Salient's suggestion.[21] This essentially confused Salient (because it was being told by the evaluators that the opinion was not an opinion), leading to a drop in accuracy, precision, and recall. The impact of this mistake was significant and required us to later omit the results associated with these errors from the data presented here (see more details in Annex 3). Regular check-ins between the technology partner and the field office (sometimes several times a week) helped to identify these instances of human error quickly.

5. **Be aware of the risks of vendor lock-in**. Salient is a powerful tool, and the Mozambique Media Strengthening Program, which incubated this experiment, now has a handful of trained specialists who can use the software. But this also means that the program must commit to continued partnership with this specific technology partner if they want to continue applying its machine learning technologies (which include opaque models whose algorithms are not published openly) to their program. This could limit scalability, reuse, and local sustainability, which are risks we considered as part of our commitment to the Principles of Digital Development.[22] By framing this is an MVP to prove the value proposition, we are now better equipped to advocate for more resources to continue this collaboration. At the same time, this limitation is a reminder that further dialog is needed to establish good practices for applying machine learning in global development contexts to determine how algorithms contribute to the public good and can be repurposed to reduce duplication

To succeed in working with machine learning, we need people who are skilled in defining actionable problems. "Problem literacy" is as important as tech or subject matter expertise.

[21] For example, one audit found that, out of 63 sentences that Salient believed to contain opinions, 65% of them (41) contained opinions as part of quotes or citations within a story. Stories that contain quotes or citations should, in reality, not be considered opinionated since it is perfectly sound journalistic practice to reference quotes and citations in objective and impartial material.

[22] The Principles of Digital Development are a set of nine considerations that help to ensure that technologies support global development programs in effective and lasting ways. See more here: www.digitalprinciples.org

of investments. Echoing recommendations put forth by USAID, we also recognize the need for more investment in local technical capacity and indigenous talent who can apply machine learning technologies in the future.[23]

6. **Build a small team of complementary skills**. For this experiment, IREX and Lore assembled a small team that included an M&E officer from the Mozambique program office and a project manager from IREX, and an AI specialist and MEL and technology for global development expert from Lore. IREX's technical advisors for media and M&E also provided valuable insight. Having a committed team member from the field office—who was comfortable and willing to learn about technical and digital concepts—was critical to this experiment's success, since most of the leg work of training and testing the software was conducted by his team of evaluators in Mozambique. Being comfortable learning technology concepts was an asset as well; although it's tempting to relegate the software to a "black box", taking the time to try and understand what's happening under the hood—at least at a basic level—can drastically help to make more informed decisions and to communicate the experiment effectively to diverse stakeholders.

Although it's tempting to relegate the software to a "black box", taking the time to try and understand what's happening under the hood—at least at a basic level—can help non-technical sector specialists use this technology responsibly and effectively.

7. **Choose the right partner**. A strong and trusting working relationship between IREX and Lore was critical to this experiment's success. Find a technology partner who is genuinely willing and able to interact with subject matter experts (such as a media NGO), provide the necessary training, and navigate some of the technical complexities, pivots, and uncertainties that come with an MVP. Afterwards, the NGO can replicate the training and knowledge internally to operate independently, but for the initial interaction, having a strong relationship with the technology partner is imperative. At its essence, a strong partnership is one where both partners are willing to learn together. For example, in addition to the substantial insight and advice that IREX gained from Lore, Lore also turned some requests that surfaced during their training sessions with the Mozambique team into new features of their Salient platform. This symbiotic relationship strengthened the partnership, reaffirmed mutual trust, and ensured that both partners gained new value and insights from this experiment.

A symbiotic relationship between IREX and the tech company strengthened the partnership and reaffirmed mutual trust.

[23] See https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf, p. 74

What's next?

We know that it's possible to use machine learning to help us monitor the impartiality of news articles. But what does this mean for supporting and strengthening media, and what comes next?

A fundamental next step is to transition away from using this tool for monitoring and towards learning about the effectiveness of our work. To strengthen our commitment to learning and adapting as we implement programs in a variety of sectors, machine learning can help us to more effectively evaluate what's working and what is not, and to adjust more nimbly. Here are some illustrative examples, some of which are currently being explored as a next evolution of the MVP:

- A **journalism school** trains 200 budding journalists to improve how they reference and use sources in their news coverage. Over the course of the training program, the school tracks thousands of articles they publish online, automatically detects the number and reliability of diverse sources in each article, and uses that information to more efficiently target additional mentorships to students who need more support on this skill. This helps to ensure that programming is adapted to tailor to the real-time needs of students.
- A **media watchdog** partners with a **policy think-tank** to analyze thousands of news articles across different media platforms, using machine learning tools to discover which topics receive more opinionated news coverage. Over time, they unveil trends of increasing opinionated and impartial writing about certain topics—knowledge that can help to inform training programs, advocacy, and policy recommendations.
- Insights from analysis of thousands of online news articles can help to measure the **vibrancy of information systems** by contributing to indices

Mainstreaming a tool like this would have a huge impact on the other activities we perform as a team. We'd evolve from simply generating reports about media quality to analyzing and discussing deeper insights with stakeholders.

ALEXANDRE GAVAZA, M&E MANAGER

like IREX's Media Sustainability Index, Freedom House's Freedom of the Press report, and Reporters Without Borders' World Press Freedom Index. These barometers don't currently incorporate media quality in their calculations—likely due to the complexity of scale and time requirement—but these could be resolved using machine learning tools like Salient.

Trained media professionals in Mozambique after working with the Media Strengthening Program.

- A **civic technologist** develops a browser add-on that scans news articles as users browse them and challenges them to highlight content that is biased or opinionated. It compares users' highlights to the opinions that it automatically detects in the article, and gives them a score that they can compare against peers. A school district installs the software on their library computers to reach hundreds of students a year, as part of a media literacy program to help young people discern facts from opinions.
- An **NGO** trains 30 reporters who apply for an investigative journalism program. After the training program, the organization tracks the news articles their trainees publish, as well as hundreds more produced by applicants who weren't accepted into the program. After automatically measuring the impartiality of articles published by both groups, the NGO sees a marked improvement in the trainee group and conducts key informant interviews to learn about what elements of the training might have led to this improvement.
- A **donor**, designing its new information and media strengthening strategy in the Middle East, tracks thousands of online news articles from news media in countries across that region. Using machine learning to scan these stories across different measures of journalistic integrity and quality,

the donor identifies two specific areas where the gaps in quality are the greatest and uses this information to inform its new investments. And by running this test with updated articles every six months, it can follow the return on their investments to learn about which programs are most effective.

Technology itself played only one part of an array of complementary components that led to the experiment's successs.

This experiment simply tested the technical feasibility of using software to monitor and evaluate the integrity of online news articles in Mozambique. With an evidence base that proves this is possible, with a team that has invaluable practice actually working with artificial intelligence tools, and with ever-growing pressures to be more efficient and smart with how projects invest their resources, we expect to apply machine learning to more programs, both within media support contexts and others. Doing so will help surface more frequent, consistent, and robust insights that can make our global development efforts more effective.

Perhaps more importantly, this experiment has surfaced important lessons about how best to apply frontier and emerging technologies to global development contexts. The technology itself played only one part of an array of complementary components that led to the experiment's success. Carefully defining a problem statement, starting small, investing in training on skills to use a new tool, understanding the limitations and advantages of the technology, and communicating effectively and realistically about its potential impact were each an invaluable asset to the successful implementation of this machine learning program. These lessons are agnostic to any specific technology and can guide future experiments into leveraging these promising tools to amplify our work in global development.

Annex 1: Mozambique Media Content Analysis Tool (MCAT)

The 18 indicators of the Mozambique MCAT are explained below. Emphasis has been added to **indicator #16**, which was the focus on this experiment.

In the future, we will explore how machine learning can help to more efficiently and consistently automate other MCAT dimensions as well.

Category	#	Indicator
Sources	1	Does the story have at least 3 sources? (Can be people and/or documents). If it's an INTERVIEW, score as YES (no more than the guest is required to be the source).
Sources	2	Are sources credible, qualified and relevant? (Person has privileged information, someone with authority on subject matter, a recognized expert)
Sources	3	Are the statements supported by evidence? (Evidence can mean direct observation by reporter, in addition to other witnesses or documentation)
Sources	4	Are individual sources properly attributed? (Who is this person and/or what is the document?
Sources	5	Does the story include a diversity of sources (i.e. different perspectives, sides of the story). If it's an INTERVIEW, score as YES (no more than the guest is required).
Relevance	6	Is the story timely (i.e. relates to "news stories" as opposed to feature writing - recent event
Relevance	7	Is the story fresh (i.e. new information or a new angle or follow up on previously reported stories)
Relevance	8	Is the story interesting (impacts your life or community and not just a small group of people; does it have "news value" - does it have consequence, or "so what"?)
Structure	9	A headline that reflects content of story
Structure	10	Paragraphs that are short and clear (meaning).
Structure	11	5 W's and H (who, what, where, when, why and how) - should be scored as a "package". This is foundation of good story. If one element is not there, score "no".
Structure	12	Inverted pyramid structure (important information at top of story, less important as the story progresses)?
Structure	13	A good lead (does it make you want to read rest of story, short with active voice, most important aspect(s) of story)?
Structure	14	Proper use of grammar and punctuation (language use)?
Structure	15	Use of Active Voice?
Impartiality	**16**	**The reporter does NOT insert his/her opinion into the story**
Impartiality	17	The reporter does NOT use incorrect language (bias and/or opinions)
Impartiality	18	The reporter represents all significant sides of the story in a fair and balanced manner. ("significant sides" means key players who must be included to create balance). If it's an INTERVIEW, score as YES.

Annex 2: The dataset

For this project we collected articles from online news sources that were selected by the MCAT Mozambique team. The existing process relied on collecting and reviewing print media but for the purposes of this project we decided to test with digital articles since this saves the time and effort of digitizing newspapers.

The Mozambique team selected eight different online sources and a set of keywords to search for. The Lore team helped automate searching for and downloading news articles based on those search terms and then showed results to the Mozambique team for further refinement of the search terms or sources. The following table shows the number of articles downloaded from each source and for each keyword.

Media Source	Biodiversity	Good governance	Gender	Nutrition	Health	Health & nutrition	Human trafficking	Transparency	Transparency & good governance	
@verdade	51	6	99	26	94	50	54	65		**445**
Calawo	3		7	1	8	3	3			**25**
Desafio		1	27				2	8		**38**
Domingo	13	1	17		8	3	5	1	3	**51**
Magazine Independente	13	27	46	4	25	1	3	4	8	**131**
MMO	3	2	57	1	85	2	5	2		**157**
Publico	2		11	2	37	2		5		**59**
Txopela	3	13	41	2	109	3	1	18	3	**193**
Wamphula Fax		7	35	5	84	4	4	7	2	**148**
	88	**57**	**340**	**41**	**450**	**68**	**77**	**110**	**16**	**1247**

Along with the above articles, the team decided to scan and include 21 print articles in the project. Those were hand selected by the MCAT Mozambique team.

The 1,247 articles in this table don't match exactly with the 1,226 that were ultimately included in Salient. This is either because they were blank, or there was an error when trying to load them into Salient. In the future, the team hopes to do a better job of tracking these issues.

Annex 3: Results data

During the course of the project, multiple training rounds were run either automatically by the software or manually by users. Each round of training helps the software refine its models. The key to understanding the statistics about F1 scores, precision, and recall that are presented in the graphs in the report is that Salient's software has an ensemble of models and algorithms. Like many machine learning tools, the inner workings about how exactly this ensemble works, which models are favored under which conditions, and other details are proprietary. But this means that the data presented in the graphs in the report are a selection or simplification of the actual results data. Specifically:

- The F1 scores presented earlier are from the **specific model that yielded the best F1 score** after all the rounds of training (even if other metrics of this model, such as precision and recall, were not the best of all of models).
- The precision and recall values presented earlier are an **average of precision and reacall values** from all models in the ensemble, rather than being from one specific model like the F1 scores are.
- Additionally, for F1, precision, and recall values, the data presented here **only pertain to positive samples**—that is, training and test results based on samples of sentences that do contain opinions. As mentioned in Footnote 16, Salient also learned from negative samples (sentences that do not contain opinions). Results for negative samples are not included in this report, but the table below articulates the average of positive and negative samples for reference.

The following table shows the evolution of the models' F1 score, precision, and recall as more positive samples were provided by coders. The "total samples" column is a sum of the positive samples and negative samples that were used to train the software in that round of training. Rows highlighted in orange are the training rounds that are omitted from the F1 score, precision, and recall charts in this report, based on our belief that significant human error distorted the results (see the Lessons section of the report).

Contact the Lore team at info@lore.ai with any questions or for more information.

		F1 Score		Precision		Recall	
Positive Samples	Total Samples	Validation Average	Training Average	Validation Avg	Training Avg	Validation Avg	Training Avg
19	69	0.78	1.00	0.72	0.95	0.87	0.97
22	72	0.56	0.99	0.28	0.43	0.52	0.67
25	75	1.00	1.00	0.60	0.91	0.72	0.89
30	80	0.86	1.00	0.85	0.95	0.76	0.95
33	83	0.89	1.00	0.81	0.96	0.79	0.93
37	87	0.47	0.83	0.36	0.79	0.54	0.70
41	91	0.89	0.99	0.45	0.83	0.64	0.70
47	97	0.79	1.00	0.80	0.88	0.71	0.79
53	103	0.52	0.98	0.43	0.86	0.54	0.74
54	104	0.70	0.99	0.58	0.85	0.60	0.71
56	108	0.34	0.85	0.25	0.84	0.50	0.74
57	114	0.79	0.88	0.63	0.83	0.64	0.73
57	116	0.52	0.61	0.43	0.78	0.54	0.61
57	116	0.63	0.75	0.51	0.77	0.54	0.62
60	121	0.76	0.82	0.62	0.82	0.61	0.72
61	127	0.74	0.67	0.64	0.77	0.67	0.60
63	140	0.71	0.95	0.66	0.85	0.59	0.75
64	145	0.65	0.96	0.50	0.83	0.60	0.75
64	148	0.67	0.87	0.63	0.84	0.57	0.80
61	145	0.43	0.58	0.56	0.74	0.54	0.59
93	177	0.95	0.96	0.86	0.94	0.84	0.93

Can machine learning help us measure the trustworthiness of news?

IREX
Center for Applied Learning and Impact
1275 K Street, NW, Suite 600
Washington DC, 20036
www.irex.org