Use of AI in classifying Misinformation

Whitepaper

Yudhanjaya Wijeratne, Isuru Samaratunga, Gayashi Jayasinghe, Dimuthu Attanayake

LIRNEasia, 12 Balcombe Place, Colombo 8, Sri Lanka
(john, jane)@lirneasia.net

LIRNEasia
Pro-poor. Pro-market.

LIRNEasia is a pro-poor, pro-market think tank whose mission is *catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology*.

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160.
info@lirneasia.net www.lirneasia.net

This work was carried out with the aid of a grant from The Asia Foundation.

An overview of the problem

Misinformation is often described as a serious threat to societies great and small. While lying and fabrication may be as old as language itself, and possibly even the human species[1], the flood of fabricated content circulating during the 2016 US presidential election seems to have attracted much attention to the erosion of notions of objectivity and balance in public discourse and the media ecosystem, supercharged by social media[2] [3]. As a result, this situation has brought about renewed interest in factchecking.

For factcheckers, however, this may be a case of too little, too late. Misinformation has become an endemic part of our digital sphere.

Two recent phenomena have brought the inadequacy of factchecking - both on platforms and in media - to the fore. The first is misinformation around COVID-19, regarded in the literature as a killer of public trust[4]; many public health care services across the world have themselves battling both a pandemic *and* rampant conspiracy theories and public mistrust.[5][6]

The second concerns the political sphere; former U.S. President Donald Trump, along with many elected representatives in the Republican Party of the United States of America, actively encouraged a repeatedly-debunked conspiracy alleging that the 2020 elections were fraudulent, culminating in a wave of armed protestors attempting an insurrection in Washington, D.C. It would be a reach to say that the politicians alone affected this act of local terrorism; evidence indicates that these people had been stewing in misinformation for years, with conspiracy theories migrating from the fringes of the Internet to the media[7] and to the highest offices of the most powerful nation on the planet[8].

Much of this has to do with the nature of social media platforms and the algorithmically curated agoras that they present us with[9]. Drawing from conversations[10] with factchecking initiatives and researchers, we find a consensus that factcheckers around the world are generally ill-equipped in the face of the sheer volume and velocity of information flow online, with nowhere near enough human or attention resources to inject enough truth into conversations.

General solutions have been proposed, but have yet to catch on. Wisdom-of-the-crowds approaches - a weaker form of *consensus gentium* - is a poor determinant of truth, as is authority (in some cases) and naive realism; tests invoking coherence and correspondence to generate the truth consume both time and effort, while generating an untruth takes no such effort.

---

[1] Dor, D. (2017). The role of the lie in the evolution of human language. *Language Sciences*, *63*, 44-59.
[2] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554-559.
[3] Lazer, David MJ, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger et al. "The science of fake news." *Science* 359, no. 6380 (2018): 1094-1096.
[4] https://interactives.lowyinstitute.org/features/covid19/issues/truth/
[5] Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., & Labrique, A. (2020). Building trust while influencing online COVID-19 content in the social media world. *The Lancet Digital Health*.
[6] https://factcheck.afp.com/public-distrust-hampers-africa-fight-against-coronavirus-misinformation
[7] https://www.nytimes.com/article/what-is-qanon.html
[8] https://www.nature.com/articles/d41586-021-00257-y
[9] https://foreignpolicy.com/2021/02/04/social-media-ban-trump-greek-agora-vote/
[10] As part of a series of interviews in the course of a much broader study by LIRNE*asia* into the information disorder, spanning 80+ key informant interviews across Asia with fact checkers, journalists, policymakers, academics, and other stakeholders; the study covered all of Asia except mainland China, North Korea, Timor Leste, Brunei, Armenia and Azerbaijan.

Therefore, at any given point in a crisis, the potential volume and velocity of misinformation far outstrips the ability of even large organizations to counter it.

Artificial Intelligence, or AI, has been put forward in the zeitgeist as a potential solution. Much of the narrative around artificial intelligence is in its ability to automate and upscale work; what is usually considered a threat to jobs may, in this state of the world, be a relief to organizations that are drastically understaffed to face the challenge at hand.
Such systems and algorithms would, in theory, be able to process a significantly higher workload than a human, and could function at different points in the process of content generation and consumption. Use cases range from powering automated moderation on content platforms such as Facebook[11] to assisting in the process of review in scientific journals[12] to aiding businesses with some stake in monitoring digital discourse[13].

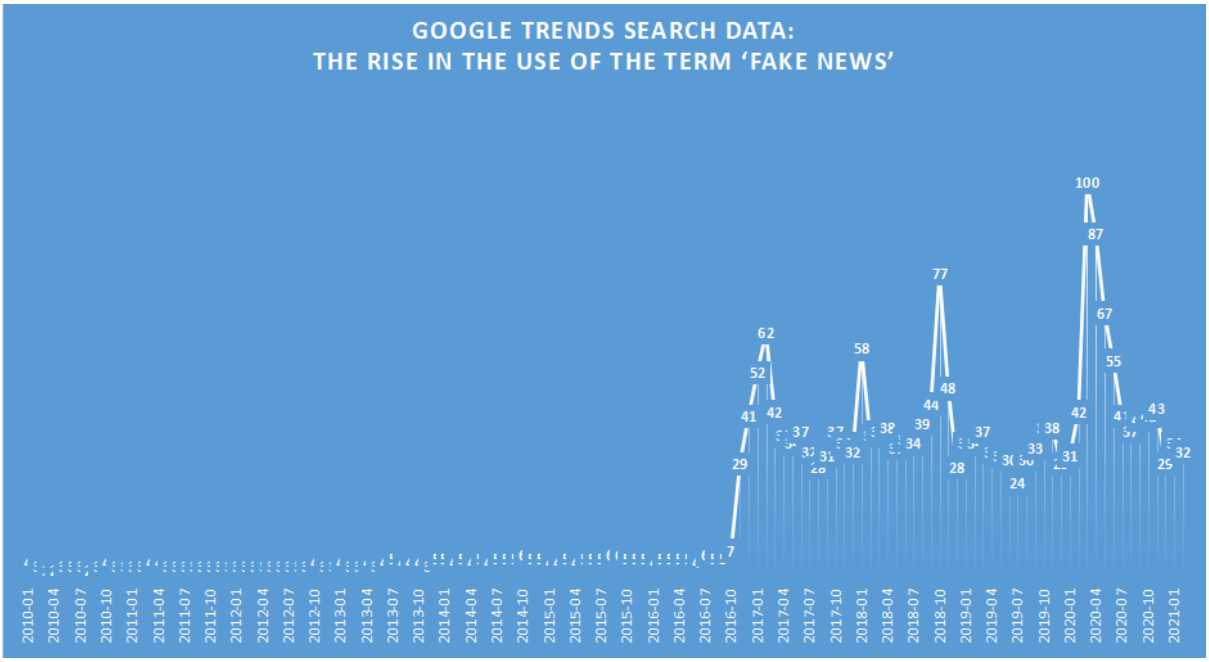
The State of the Art, and, therefore, our research questions

Computer science, particularly the field known as natural language processing, has long since spawned a number of ways of classifying text content through the use of AI. These methods have often been turned to the task of classifying misinformation. The general idea is that various AI algorithms train on large collections, or corpora, of *annotated* text, labelled by humans to signify which are truth, and which are misinformation. This training process creates models that can then mimic the kind of classification performed by humans. Recent methods have expanded on this to include images in sophisticated multi-media analysis.

There are epistemological boundaries: these techniques do not fact-check as a human agent does, but often rely on linguistic features - such as the co-occurence of words and their relation to each other. Anecdotally, this is one of the strongest offhand reasons for dismissing automated, corpus-based AI methods, since the process of search, triangulation, and *journalism* that human factcheckers go through simply does not happen here. Any patterns not visible to the algorithms from the training corpus would be increasingly difficult to classify, and therefore, as public discourse and misinformation trends change, these tools become obsolete unless retrained or remade with new data. Furthermore, an astute observer may point out that satire may be impossible to interpret using these methods.

Nevertheless, sophisticated attempts have been made. Our survey of the state-of-the-art[14] reveals a rich history of text classification methods originating from the need for detecting spam and fake reviews on commercial websites. 2016-2017 seems to have been a watershed period; the 2016 US election brought with it a multi-disciplinary, mass awareness of misinformation and its effects, spilling over from journalism to fields as far removed as economics (Allcott & Matthew, 2017). This general uptick is reflected in Google search traffic worldwide for 'fake news', the layman term for various types of misinformation:

[11] https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/
[12] https://www.technologyreview.com/2020/05/29/1002349/ai-coronavirus-scientific-fact-checking/
[13] https://www.logically.ai/
[14] Wijeratne, Y., & Attanayake, D. C. (2021). Artificial intelligence for factchecking: Observations on the state and practicality of the art.



This appears to have led to a knock-on effect in fields where misinformation detection met AI automation, bringing with it an explosion of both research and citations thereof. This loosely coincided with the rise of *deep learning* (LeCun et al., 2015), the burgeoning availability of increasingly more powerful processing via the utilization of GPUs, or graphical processing units; and the spread of software libraries such as Keras, Torch and Tensorflow that, in turn, enabled more researchers to build ever-more sophisticated models for the task.

As a result, there is now a readily deployable "classical machine learning stack" with a variety of popular algorithms, and a "deep learning stack" with its own set of algorithmic favourites. The former boasts the company of methods such as logistic regression, naive Bayes, random forests, support vector machines, and gradient boosting methods; the latter has a rogue's gallery of recurrent neural networks (RNNs), convolutional neural networks (CNNs); as well as a scattering of more esoteric approaches: 3HAN, Ti-CNNs for image+text, etc. For a while now, classical machine learning has been capable of reaching over 90% accuracy in tasks where an AI model was asked to classify a piece of content as either true or false; recent deep learning methods have upped the ante to as high as 98%.

The trouble with AI

But despite artificial intelligence being put forward as a panacea for almost all our ills, there are problems that we must address. At the end of the day, artificial intelligence is the training of an algorithmic model on vast amounts of data so that it may infer patterns represented within the data. If there is no data, there is no artificial intelligence.

This therefore lends itself to regional disadvantages. Whenever artificial intelligence comes into the conversation, it seems to be anchored in examples based on the Anglosphere – that is to say, the United States of America or the United Kingdom - or, in limited cases, the European Union. Very rarely are these conversations had about countries or contexts in South Asia.
From prior work in language processing, we knew that this has much to do with the nature of English as a lingua franca, particularly within the computer science community, and with the global

arrangement of computational language resources[15]. As a result of the evolution of the field and the structural setup of the production of knowledge in such, many languages in South Asia simply do not have enough data for such artificial intelligence to be built.

This first problem, therefore, is one of **data availability**. Without the data being available, despite proof of concept in English, it is difficult to assess whether the same methods can be applied to other languages.

The second problem is one of **computational complexity and resource availability.** Many fact checkers and related outfits in South Asia operate in environments that are already resource–scarce. Much of the cutting-edge research in the field is computationally complex, and therefore requires expertise, computational capability, and therefore financial resources that may not be available in many contexts.

As expected, our survey shows that almost all work is in English. A scan of underlying datasets used by the majority of well cited examples show that even Romance languages are underrepresented, leave alone languages throughout Asia. Furthermore, Very few of the results are directly comparable, as different researchers often build datasets from scratch and publish results formulated from very different datasets. This makes it difficult to compare methods; one set of results might be from a corpus of less than 500 news articles, while yet another set of results might be derived from 400,000 tweets. This also makes it difficult to understand how much data these algorithms require to achieve a given baseline level of competency or accuracy. Furthermore, computer science academia remains obsessed with accuracy, often neglect other practical criteria for deployment, such as computational resource requirements.

Where we come in

In May 2020, we applied for a grant from the Asia foundation. The overall question was: Given what we know about the limitations and the structural disadvantages that exist in this subject, can countries in South Asia make use of technologies for faster and more efficient fact checking?
In order to answer the overarching question of whether South Asian countries can make use of these technologies, we must drill down into three research questions:

- What commonly used, well understood AI algorithms are likely to be most accurate in resource context?
- Of these, how much computational time and effort is required to achieve a relatively high degree of accuracy?
- How much data do we need for these AI to be ported into any other language in South Asia, and, deriving from that, what kind of effort might be required to make AI readily usable and available for languages in South Asia?
- If it turns out that such technology can very well be implemented in languages in South Asia, what is the operating context of fact checkers in the region, and does it allow or accomodate the use of such technology?

[15] Wijeratne, Y., de Silva, N., & Shanmugarajah, Y. (2019). Natural language processing for government: Problems and potential. *International Development Research Centre (Canada)*.

Research: Quantitative

First, upon careful examination of the range of technologies in the field, we decided to confine ourselves to "the classical machine learning stack". There are two reasons for this: the first is that this set of AI algorithms are of less complexity, both in theory and in practice, as compared to the "deep learning stack". The second is that the set of algorithms requires less data as compared to the alternatives. Both these attributes suit conditions of resource scarcity.

To compare and contrast algorithms, and to understand the data requirements, we narrowed on the most major contenders for popularity in the field. These algorithms were tested on four different English datasets, comprised of different types of text, from news articles to fact checks on snippets from political speeches[16]. We extended these tests from data ranges ranging from just above 300 samples of data to 400,000 samples, randomising in such a manner as to better iron out the impact of unexpected patterns in the arrangement of the data. For each dataset, we measured the accuracy at a given amount of data, and the training time; the latter being a proxy for compute resources that might be required to run this algorithm[17].

This experiment, conducted in English, allowed us to benchmark these algorithms and generate expectations for how much data and effort might reasonably go into the core of AI that could perform The task of reading text and assigning it a binary classification (of true or false, or of credible or not) at an artificial threshold of 90% or above.
From this exercise, we ascertained that support vector machines, gradient boost methods, and random forests – the latter two being tree-based methods – are typically more than adequate for the task; at the top end of data, they were able to generate AI models that demonstrated up to 97% accuracy in this sort of fact checking in the context of binary categories.
In all cases, roughly 1500 items of data were sufficient to teach a model up to and past the 90% accuracy threshold. These results are highly compatible across different types of data. Also important is the fact that these results can be achieved with readily available programming libraries, using common off-the-shelf techniques that are highly portable between languages because of their lack of language specific preprocessing.

---

[16] 1) **A class-balanced dataset comprised of 500,000 news articles extracted from the Fake News Corpus by Szpakowski (2020)**, which in turn is a collection of English news articles extracted using a typology from the (now-defunct) Open Sources project2.
Our reduced version, hereafter referred to as **FNC500k**, is labelled reliable and fake: the reliable data is drawn directly from the Credible category presented in Szpakowski's source dataset, the fake category drawing equally from the Fake News and Conspiracy Theory categories. Both this source and the format of a binary classification is in use in the Kaggle Fake News Detection Challenge KDD 2020.
2) **The dataset from the Kaggle Fake News Detection Challenge KDD 2020**, which includes 20,800 news articles curated by Kai Shu. This dataset (often referred to as the Kaggle dataset) is part of fake news detection shared task for the Second International TrueFact Workshop: Making a Credible Web for Tomorrow. The data is labelled as fake (1) or real (0) respectively.
3) **The dataset by George McIntire, commonly cited as the KDNuggets Fake News dataset**,consisting of news articles labelled as real and fake. The fake component here draws from a previously released Kaggle dataset assembled from sources flagged by BS Detector; the real is derived from AllSides.com. This dataset is mentioned as having 10,558 articles, but public forks of it, duplicated from McIntire's Github repository, contain only 6335 entries.
4) **The LIAR dataset by Wang (2017)**: 12,800 short statements from Politifact.com, among them excerpts from news releases, TV/radio interviews, campaign speeches, TV advertisements, social media posts, and statements issued in political debate; these are labelled as pants-fire, false, barely-true, half-true, mostly-true, and true.
[17] Wijeratne, Y. (2021). How Much Bullshit Do We Need? Benchmarking Classical Machine Learning for Fake News Classification.

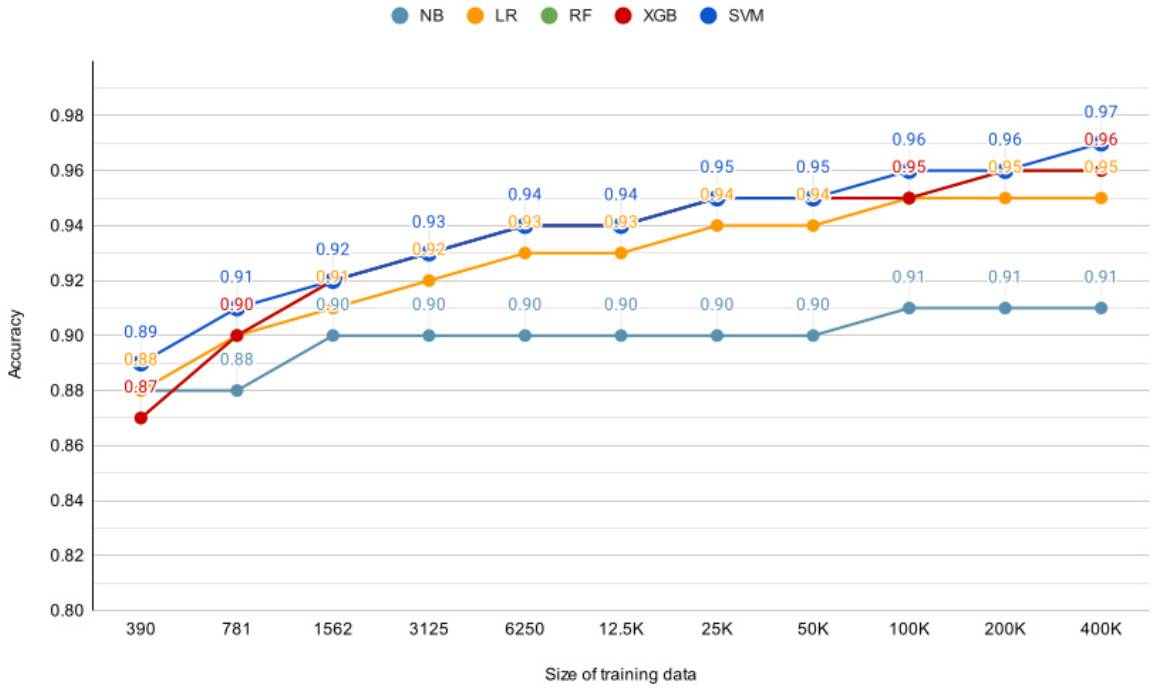


Figure 1: *Top-line accuracy figures from benchmarks on the FNC500K dataset*

With this figure of 1500 items of training data, we were able to establish practical bounding boxes on the amount of effort that would be required to replicate such success in other languages.
We then set out to test whether these algorithms could work in a compatible fashion in South Asian languages. For this exercise we selected Sinhala and Bengali, two languages that are typically considered computationally resource-poor. Working with local journalists and fact checkers, we assembled and trained two teams across both countries to create the brand-new datasets required for the task.

In Sinhala, the resultant AI models reached over 80% accuracy, compared to trained researchers[18]; in Bengali, we were able to achieve results over the 90% threshold that was the target of our work in English. As a bonus, we were also able to publish this data for free under open access terms for other research and potentially for more sophisticated AI work.
As expected, the largest share of the work was in identifying thresholds for particular algorithms and in creating data in a manner that is ontologically consistent. In our research, we noted that due to the lack of factchecking in media sources, the classical true/false binary may not perhaps apply as well in the media landscape South Asia; we suggested instead that machine learning models could be used to identify news that was typically uncertain in nature – as in, belonging to a class of information that is difficult to corroborate from other sources – and passed it on to humans to complete the fact checking process with ground truth. Accordingly, we were able to build and demonstrate AI models that functioned as “uncertainty detectors” in this manner, with comparable, if somewhat variable accuracy.

[18] Jayawickrama, V., Ranasinghe, A., Attanayake, D. C., & Wijeratne, Y. A Corpus and Machine Learning Models for Fake News Classification in Sinhala.

Research: Qualitative

In answering the questions above, we consistently happened on the question of operating contexts and capacity. While proof of concept can be demonstrated, there is typically another layer of processes before any form of technology can be used by people other than programmers.
We therefore conducted interviews with fact checkers across Sri Lanka and Bangladesh to understand whether their business models and operations would allow them to use such tools, and to understand what else they might require from such tooling.
All in all, 8/11 fact checking organisations expressed willingness to try out such tooling. Factcheckers typically have limited human resources (number of employees vary from 1 to 15), and multiple skills (content writing, translation and graphic designing) are required for operations, leaving operations perpetually needing more time, funding and human resources. Bearing in mind that a this assessment is heavily reliant on subjective understandings of the concept of AI, and off the function of the machine learning models built by LIRNEasia, key informants noted that the ability to scan through content and perform repetitive analysis tasks would be quite useful in augmenting their workflow.

"There are tens of thousands or millions of content in a single language circulating on social media platforms in a day. It is totally impossible for fact checkers or researchers to manually go through all those content"

- *AFP Fact Check, Bangladesh*

AI can be the ultimate assistant fact-checkers need against their fight with the most crucial nemesis - time and the pile of information they need to dig into. AI can help report possible misinformation at the early stage when it gets viral, by monitoring news websites, known misinformation sources, and public social media platforms ... a tool that gives fact-checkers the ability to monitor news websites - viral contents could be a great tool in identifying misinformation quicker, before it can create a serious impact…

- Jachai, Bangladesh's oldest factchecking operation

A couple of operations had definite ideas on where such tooling could be used - to create shortlists and essentially act as spam filters for factcheckers (AFP Bangladesh and Fact Watch). The amount of work that could be automated thus varied widely in estimate, ranging from 30-40% (Fact Research) to 70-80% (Fact Watch).

Some expressed caution while agreeing to the general utility of such tools:
Local agencies that do fact checking, you know, there are community-based organizations like they will be interested in it, because they don't have the same kind of tools that we have. A lot of provincial media outlets, etc, who are interested in these kinds of things, it would be helpful for them...
There is a mechanism with Facebook where certain content is suggested to us, and some of it is through an AI, because of things like mass shares. And some of most of it is now people actually manually reporting content. And we find that you know, with certain things, mass shares don't necessarily work because sometimes the narrative is not false. But then it will still get flagged, because there's been a lot of shares. So it can help, it can be helpful. And it cannot be so helpful... A story may be potentially, you know, false or misleading...it's not a uniform application that you can use across the board.

- A Sri Lankan fact checking organisation

We note that affordability and human resource capability are consistently put forward the biggest barriers preventing factcheckers from adopting such tooling, especially since any tool will need to be evaluated before put into use in daily operations. Prior exposure to AI tools developed for fact

checking and content moderation had an impact on how fact checkers perceived a technology: both AFP and Fact Watch had access to Facebook's AI tools; Citizen Fact Check had been in discussions to using AI for detecting misleading images.

Our Summary and Findings

On a technical level, we can demonstrate that the actual task of building AI technology for factchecking is, by and large, not only achievable; it can be done so in an open manner that promotes the growth of research in the field in languages in South Asia. Given, at a minimum, the following:

1) Trained annotators, embodying practical experience in a given information domain, capable of classifying at *least* 1500 examples of fake news and credible content in such a manner that there is even representation of both types
2) A robust review process to examine data annotated thusly according to a pre-set schema
3) Computer hardware (under $2000) and typically a single programmer, fluent in programming languages that have extensive support for machine learning (in this case, Python)

Under these conditions, the basic underlying AI model can be generated with today's technology without relying on proprietary tooling or large amounts of funding. It appears that tree-based methods provide the best mix of accuracy while minimising resource overhead.

Of course, there are caveats that we must note here. In order to examine this thesis, we have restricted ourselves to the analysis of texts. Images are a more computationally sophisticated use case; although the basic tenets remain the same – we need data, and typically a single programmer using existing software libraries can approximate, if not outright beat, the state-of-the-art. Given the high degree of accuracy that we were able to achieve here, we can posit that the bleeding edge of the field is now in pursuit of incremental accuracy gains, typically within a 2% margin.

However, if nothing else, South Asia has a way of making even simple truths complicated, and this likewise applies for the field in which we find ourselves. Given resource constraints, fact checkers in Sri Lanka and Bangladesh remain unwilling or unable to implement these technologies on their own. Any AI-based solution must necessarily be wrapped in user interfaces and a software experience that allows these factcheckers use it without necessarily being technically fluent on its inner workings.

There are also problems with Anglicization, reflected here as it is in AI in general:

The problem with misinformation in India is that the same piece of misinformation will throb up in multiple languages so the moment you've dealt with it in as of means it probably popped up in bengali, the moment you dealt with it bengali its popped up in punjabi, moment you've dealt with it in punjabi its popped up in malayalam. I mean you know that's the problem right so that's the first problem the second of course is fact checkers to begin with the journalism industry is heavily anglicized so we are not only now the last two years maybe begun to do it in Hindi in a serious way.

- BOOM, which operates in India, Bangladesh and Myanmar

This suggests that AI deployed in South Asia would have to be built on multilingual datasets; preferably using translations of material so as to maintain parity and the baseline competence in the major languages being used in the field.

Another observation is that most fact checkers in South Asia rarely use the kind of binary distinctions that most computer science research in the field operates on. The closest this sort of simplicity would be Watchdog and AFP, both of which operate in Sri Lanka; the former uses true / fake designations and the latter uses false/misleading/missing context. Jachai, for example, uses a system

of 7 classifications: False, True, Misleading, Mostly-false, Mostly-true, Unproven, and Mixed. The use of these classifications mirror the observations from our quantitative research.
Such complexity in classifications comes with a problem: AI accuracy tends to drop significantly as the number of labels or types of data increase. Therefore, AI created has a very specific place in the workflow; either to serve as a filter between human fact checkers and the incoming flood of information, or to serve as rudimentary truth filters to a general public; these tools would ideally not make the final judgement on the nuanced classification of a piece of content, and in fact several key informants noted that the human touch with that be critical regardless of how much technology is in the mix.

Some fact checkers also note that the combination of text and images is a significant form of misinformation (BOOM estimates that 80% of misinformation is such a mix). This suggests to us that hybrid deep learning architectures such as 3HAN will be essential for any deployment that has to deal with this type of multimedia.

Lastly, there is the elephant in the room: misinformation is a dynamic field where the discourse is in a constant state of change. This means that any AI trained with today's data must eventually be retrained with fresh data, and we expect accuracy loss in the predictions of any AI model between these periods of retraining. Therefore, it is not enough to put in the effort once; this effort must be put in at regular intervals, and the datasets constantly updated to keep up with the changing world. However, there is also a silver lining: there appears to be at least some willingness towards collaboration, as opposed to the competition that is typically present in journalism. This, coupled with the need for multilingual solutions and the practical experience from building a eye for the quantitative component of the study, leads us to postulate that it should be possible to bring together fact checker expertise in assembling and updating underlying datasets that can be used to build robust solutions that can be deployed widely in South Asia, either as part of fact check operations, or as tools for the general public to use in their interactions with the Internet and social media.