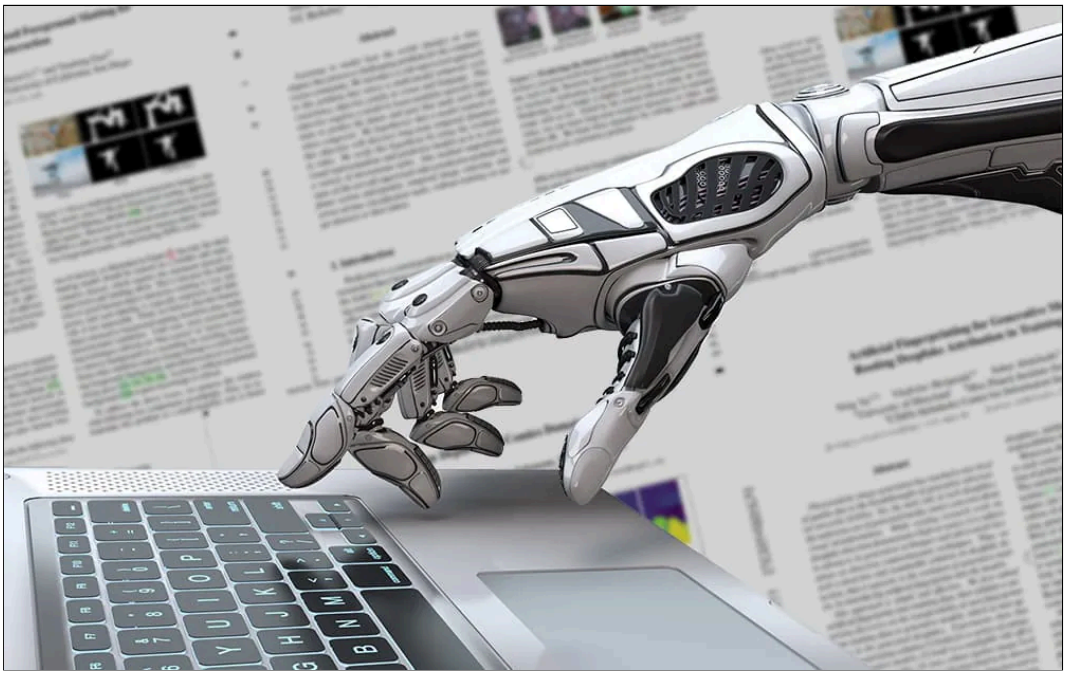
# Towards Automated Science Writing

Originally published at https://www.unite.ai/towards-automated-science-writing/



This morning, trawling the Computer Science sections of Arxiv, as I do most mornings, I came across a recent paper from the Federal University of Ceara in Brazil, offering a new Natural Language Processing framework to automate the summarization and extraction of core data from scientific papers.

Since this is more or less what I do every day, the paper brought to mind a comment on a Reddit writers' thread earlier this year – a prognostication to the effect that science writing will be among the earliest journalistic jobs to be taken over by machine learning.

Let me be clear – I *absolutely believe* that the automated science writer is coming, and that all the challenges I outline in this article are either solvable now, or eventually will be. Where possible, I give examples for this. Additionally, I am not addressing whether or not current or near-future science-writing AIs will be able to *write* cogently; based on the current level of interest in this sector of NLP, I'm presuming that this challenge will eventually be solved.

Rather, I'm asking if a science-writer AI will be able to *identify* relevant science stories in accord with the (highly varied) desired outcomes of publishers.

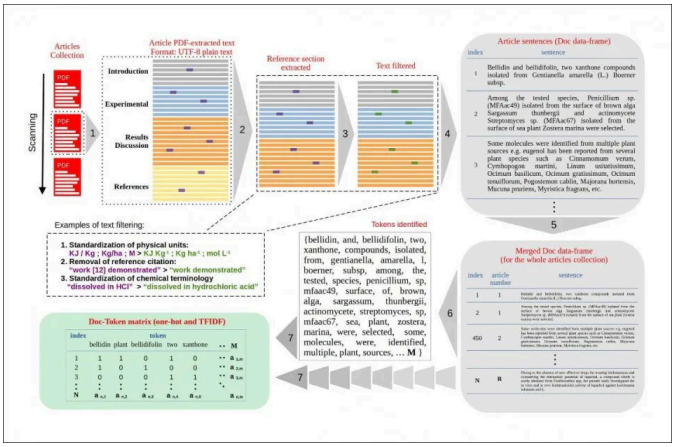
I don't think it's imminent; based on trawling through the headlines and/or copy of around 2000 new scientific papers on machine learning every week, I have a rather more cynical take on the extent to which academic submissions can be algorithmically broken down, either for the purposes of academic indexing or for scientific journalism. As usual, it's those damned *people* that are getting in the way.

## Requisites for the Automated Science Writer

Let's consider the challenge of automating science reporting on the latest academic research. To keep it fair, we'll mostly limit it to the CS categories of the very popular non-paywalled Arxiv domain from Cornell University, which at least has a number of systematic, templated features that can be plugged into a data extraction pipeline.

Let's assume also that the task at hand, as with the new paper from Brazil, is to iterate through the titles, summaries, metadata and (if justified) the body content of new scientific papers in search of constants, reliable parameters, tokens and actionable, reducible domain information.

This is, after all, the principle on which highly successful new frameworks are gaining ground in the areas of earthquake reporting, sports writing, financial journalism and health coverage, and a reasonable departure point for the AI-powered science journalist.




*The workflow of the new Brazilian offering. The PDF science paper is converted to UTF-8 plain text (though this will remove italic emphases that may have semantic meaning), and article sections labeled and extracted before being passed through for text filtering. Deconstructed text is broken into sentences as data-frames, and the data-frames merged before token identification, and generation of two doc-token matrices* Source: https://arxiv.org/ftp/arxiv/papers/2107/2107.14638.pdf

## Complicating the Template

One encouraging layer of conformity and regularization is that Arxiv imposes a pretty well-enforced template for submissions, and provides detailed guidelines for submitting authors. Therefore, papers generally conform to whichever parts of the protocol apply to the work being described.

Thus the AI pre-processing system for the putative automated science writer can generally treat such sections as sub-domains: *abstract*, *introduction*, *related/prior work*, *methodology/data*, *results/findings*, *ablation studies*, *discussion*, *conclusion*.

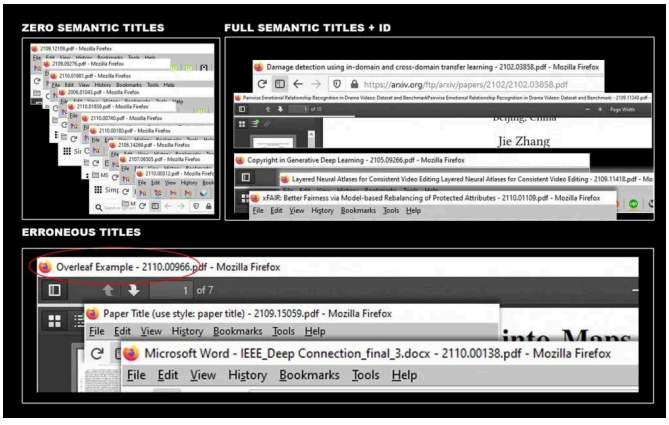
However, in practice, some of these sections may be missing, renamed, or contain content that, strictly speaking, belongs in a different section. Further, authors will naturally include headings and sub-headings that don't conform to the template. Thus it will fall to NLP/NLU to identify pertinent section-related content from context.

## Heading for Trouble

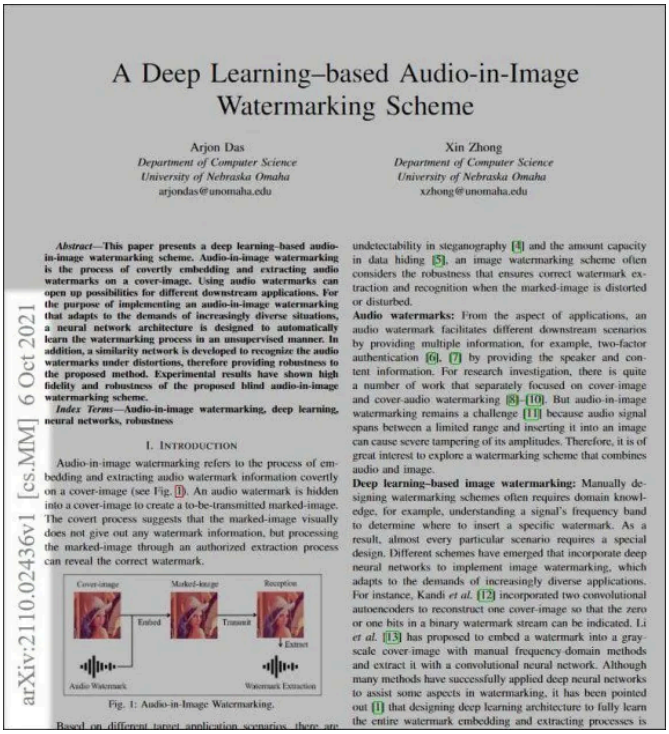
A header hierarchy is an easy way for NLP systems to initially categorize blocks of content. A lot of Arxiv submissions are exported from Microsoft Word (as evidenced in the mishandled Arxiv PDFs that leave 'Microsoft Word' in the title header – see image below). If you use proper section headings in Word, an export to PDF will recreate them as hierarchical headings that are useful to the data extraction processes of a machine reporter.

However, this assumes that authors are actually using such features in Word, or other document creation frameworks, such as TeX and derivatives (rarely provided as native alternative formats in Arxiv submissions, with most offerings limited to PDF and, occasionally, the even more opaque PostScript).

Based on years of reading Arxiv papers, I've noted that the vast majority of them do not contain *any* interpretable structural metadata, with the title reported in the reader (i.e. a web browser or a PDF reader) as the full title (including extension), of the document itself.




In this case, the paper's semantic interpretability is limited, and an AI-based science writer system will need to programmatically relink it to its associated metadata at the Arxiv domain. Arxiv convention dictates that basic metadata is also inserted laterally in large grey type on page 1 of a submitted PDF (see image below). Sadly – not least because this is the only reliable place you can find a publication date or version number – it's often excluded.



A Deep Learning–based Audio-in-Image Watermarking Scheme

Arjon Das
*Department of Computer Science*
*University of Nebraska Omaha*
arjondas@unomaha.edu

Xin Zhong
*Department of Computer Science*
*University of Nebraska Omaha*
xzhong@unomaha.edu



***Abstract*—This paper presents a deep learning–based audio-in-image watermarking scheme. Audio-in-image watermarking is the process of covertly embedding and extracting audio watermarks on a cover-image. Using audio watermarks can open up possibilities for different downstream applications. For the purpose of implementing an audio-in-image watermarking that adapts to the demands of increasingly diverse situations, a neural network architecture is designed to automatically learn the watermarking process in an unsupervised manner. In addition, a similarity network is developed to recognize the audio watermarks under distortions, therefore providing robustness to the proposed method. Experimental results have shown high fidelity and robustness of the proposed blind audio-in-image watermarking scheme.**

***Index Terms*—Audio-in-image watermarking, deep learning, neural networks, robustness**

## I. Introduction

Audio-in-image watermarking refers to the process of embedding and extracting audio watermark information covertly on a cover-image (see Fig. 1). An audio watermark is hidden into a cover-image to create a to-be-transmitted marked-image. The covert process suggests that the marked-image visually does not give out any watermark information, but processing the marked-image through an authorized extraction process can reveal the correct watermark.




Fig. 1: Audio-in-Image Watermarking.

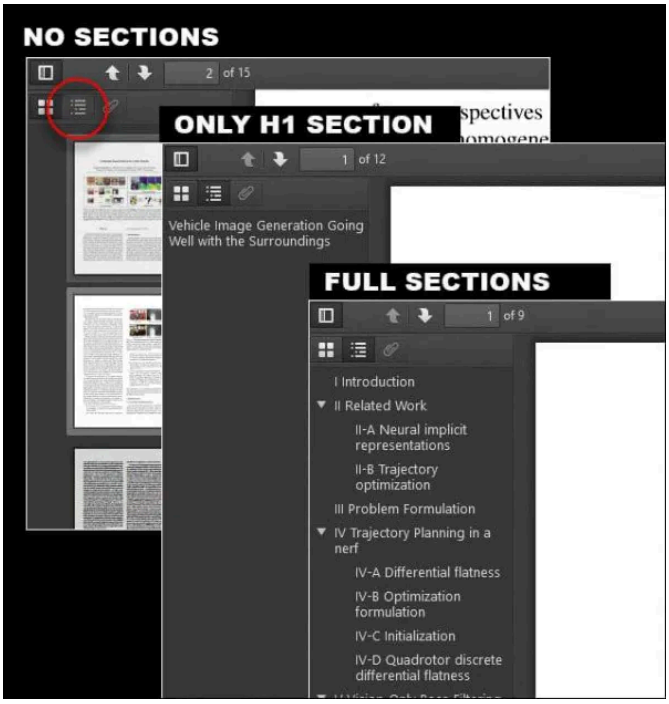
Based on different target application scenarios, there are undetectability in steganography [4] and the amount capacity in data hiding [5], an image watermarking scheme often considers the robustness that ensures correct watermark extraction and recognition when the marked-image is distorted or disturbed.

**Audio watermarks:** From the aspect of applications, an audio watermark facilitates different downstream scenarios by providing multiple information, for example, two-factor authentication [6], [7] by providing the speaker and content information. For research investigation, there is quite a number of work that separately focused on cover-image and cover-audio watermarking [8]–[10]. But audio-in-image watermarking remains a challenge [11] because audio signal spans between a limited range and inserting it into an image can cause severe tampering of its amplitudes. Therefore, it is of great interest to explore a watermarking scheme that combines audio and image.

**Deep learning–based image watermarking:** Manually designing watermarking schemes often requires domain knowledge, for example, understanding a signal's frequency band to determine where to insert a specific watermark. As a result, almost every particular scenario requires a special design. Different schemes have emerged that incorporate deep neural networks to implement image watermarking, which adapts to the demands of increasingly diverse applications. For instance, Kandi *et al.* [12] incorporated two convolutional autoencoders to reconstruct one cover-image so that the zero or one bits in a binary watermark stream can be indicated. Li *et al.* [13] has proposed to embed a watermark into a gray-scale cover-image with manual frequency-domain methods and extract it with a convolutional neural network. Although many methods have successfully applied deep neural networks to assist some aspects in watermarking, it has been pointed out [1] that designing deep learning architecture to fully learn the entire watermark embedding and extracting processes is

Many authors either use no styles at all, or only the H1 (highest header/title) style, leaving NLU to once again extract headings either from context (probably not so difficult), or by parsing the reference number that comprises the title in the document route (i.e. https://arxiv.org/pdf/**2110.00168**.pdf) and availing itself of net-based (rather than local) metadata for the submission.

Though the latter will not solve absent headings, it will at least establish which section of Computer Science the submission applies to, and provide date and version information.




## GluedText at ParagraphReturns

With PDF and postscript the most common available Arxiv formats submitted by authors, the NLP system will need a routine to split end-of-line words from the start-of-subsequent-line words that get 'attached' to them under PDF format's unfortunate default optimization methods.

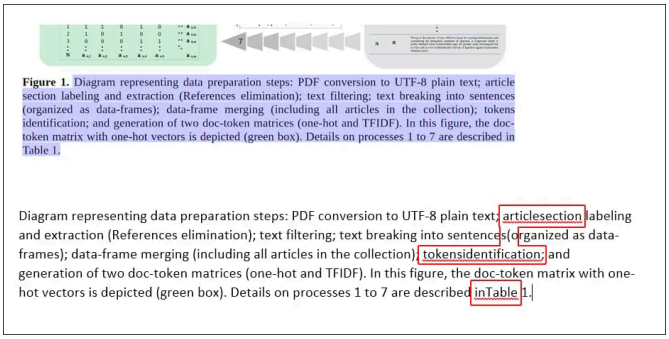


**Figure 1.** Diagram representing data preparation steps: PDF conversion to UTF-8 plain text; article section labeling and extraction (References elimination); text filtering; text breaking into sentences (organized as data-frames); data-frame merging (including all articles in the collection); tokens identification; and generation of two doc-token matrices (one-hot and TFIDF). In this figure, the doc-token matrix with one-hot vectors is depicted (green box). Details on processes 1 to 7 are described in Table 1.

Diagram representing data preparation steps: PDF conversion to UTF-8 plain text; articlesection labeling and extraction (References elimination); text filtering; text breaking into sentences(organized as data-frames); data-frame merging (including all articles in the collection); tokensidentification; and generation of two doc-token matrices (one-hot and TFIDF). In this figure, the doc-token matrix with one-hot vectors is depicted (green box). Details on processes 1 to 7 are described inTable 1.

De-concatenating (and de-hyphenizing) words can be accomplished in Perl and many other simple recursive routines, though a Python-based approach might be less time-consuming and more adapted to an ML framework. Adobe , the originator of the PDF format, has also developed an AI-enabled conversion system called Liquid Mode, capable of 'reflowing' baked text in PDFs, though its roll-out beyond the mobile space has proved slow.

## Poor English

English remains the global scientific standard for submitting scientific papers, even though this is controversial. Therefore, interesting and newsworthy papers sometimes contain appalling standards of English, from non-English researchers. If adroit use of English is included as a metric of value when a machine system evaluates the work, then not only will good stories often be lost, but pedantic lower-value output will be rated higher simply because it says very little very well.

NLP systems that are inflexible in this regard are likely to experience an additional layer of obstacles in data extraction, except in the most rigid and parameterized sciences, such as chemistry and theoretical physics, where graphs and charts conform more uniformly across global science communities. Though machine learning papers frequently feature formulae, these may not represent the defining value of the submission in the absence of the fully-established scientific consensus on methodology that older sciences enjoy.

## Selection: Determining Audience Requirements

We'll return to the many problems of decomposing eccentric science papers into discrete data points shortly. Now, let's consider our audience and aims, since these will be essential to help the science writer AI sift through thousands of papers per week. Predicting the success of potential news stories is already an active area in machine learning.

If, for instance, high volume 'science traffic' is the sole objective at a website where science-writing is just one plank of a broader journalistic offering (as is the case with the UK's *Daily Mail* science section), an AI may be required to determine the highest-grossing topics in terms of traffic, and optimize its selection towards that. This process will probably prioritize (relatively) low-hanging fruit such as *robots*, *drones*, *deepfakes*, *privacy* and *security vulnerabilities*.

In line with the current state of the art in recommender systems, this high-level harvesting is likely to lead to 'filter bubble' issues for our science writer AI, as the algorithm gives increased attention to a slew of more spurious science papers that feature 'desirable' high-frequency keywords and phrases on these topics (again, because there's money to be had in them, both in terms of traffic, for news outlets, and funding, for academic departments), while ignoring some of the much more writeable 'Easter eggs' (see below) that can be found in many of the less-frequented corners of Arxiv.

## One and Done!

Good science news fodder can come from strange and unexpected places, and from previously unfruitful sectors and topics. To further confound our AI science writer, which was hoping to create a productive index of 'fruitful' news sources, the source of an off-beat 'hit' (such as a Discord server, an academic research department or a tech startup) will often *never again produce actionable material*, while continuing to output a voluminous and noisy information stream of lesser value.

What can an iterative machine learning architecture deduce from this? That the many thousands of previous 'outlier' news sources that it once identified and excluded are suddenly to be prioritized (even though doing so would create an ungovernable signal-to-noise ratio, considering the high volume of papers released every year)? That the topic itself is worthier of an activation layer than the news-source it came from (which, in the case of a popular topic, is a redundant action)..?

More usefully, the system might learn that it has to move up or down the data-dimensionality hierarchy in search of patterns – if there really are any – that constitute what my late journalist grandfather called 'a nose for news', and define the feature *newsworthy* as an itinerant and abstract quality that can't be accurately predicted based on provenance alone, and which can be expected to mutate on a daily basis.

## Identifying Hypothesis Failure

Due to quota pressure, academic departments will sometimes publish works where the central hypothesis has failed completely (or almost completely) in testing, even if the project's methods and findings are nonetheless worth a little interest in their own right.

Such disappointments are often not signaled in summaries; in the worst cases, disproved hypotheses are discernible only by reading the results graphs. This not only entails inferring a detailed understanding of the methodology from the highly select and limited information the paper may provide, but would require adept graph interpretation algorithms that can meaningfully interpret everything from a pie-chart to a scatter-plot, in context.

An NLP-based system that places faith in the summaries but can't interpret the graphs and tables might get quite excited over a new paper, at first reading. Unfortunately, prior examples of 'hidden failure' in academic papers are (for training purposes) difficult to generalize into patterns, since this 'academic crime' is primarily one of omission or under-emphasis, and therefore elusive.

In an extreme case, our AI writer may need to locate and test repository data (i.e. from GitHub), or parse any available supplementary materials, in order to understand what the results signify in terms of the aims of the authors. Thus a machine learning system would need to traverse the multiple unmapped sources and formats involved in this, making automation of verification processes a bit of an architectural challenge.

## 'White Box' Scenarios

Some of the most outrageous claims made in AI-centered security papers turn out to require extraordinary and very unlikely levels of access to the source code or source infrastructure – 'white box' attacks. While this is useful for extrapolating previously unknown quirks in the architectures of AI systems, it almost never represents a realistically exploitable attack surface. Therefore the AI science writer is going to need a pretty good bullshit detector to decompose claims around security into probabilities for effective deployment.

The automated science writer will need a capable NLU routine to isolate 'white box' mentions into a meaningful context (i.e. to distinguish mentions from core implications for the paper), and the capability to deduce white box methodology in cases where the phrase never appears in the paper.

## Other 'Gotchas'

Other places where infeasibility and hypothesis failure can end up quite buried are in the *ablation studies*, which systematically strip away key elements of a new formula or method to see if the results are negatively affected, or if a 'core' discovery is resilient. In practice, papers that include ablation studies are usually quite confident of their findings, though a careful read can often unearth a 'bluff'. In AI research, that bluff frequently amounts to *overfitting*, where a machine learning system performs admirably on the original research data, but fails to generalize to new data, or else operates under other non-reproducible constraints.

Another useful section heading for potential systematic extraction is *Limitations*. This is the very first section any science writer (AI or human) should skip down to, since it can contain information that nullifies the paper's entire hypothesis, and jumping forward to it can save lost hours of work (at least, for the human). A worse-case scenario here is that a paper actually has a *Limitations* section, but the 'compromising' facts are included *elsewhere* in the work, and not here (or are underplayed here).

Next is *Prior Work*. This occurs early on in the Arxiv template, and frequently reveals that the current paper represents only a minor advance on a much more innovative project, usually from the previous 12-18 months. At this stage, the AI writer is going to need the capability to establish whether the prior work attained traction; is there still a story here? Did the earlier work undeservedly slip past public notice at the time of publication? Or is the new paper just a perfunctory postscript to a well-covered previous project?

## Evaluating Re-Treads and 'Freshness'

Besides correcting errata in an earlier version, very often V.2 of a paper represents little more than the authors clamoring for the attention they didn't get when V.1 was published. Frequently, however, a paper actually deserves a second bite at the cherry, as media attention may have been diverted elsewhere at time of original publication, or the work was obscured by high traffic of submissions in overcrowded 'symposium' and conference periods (such as autumn and late winter).

One useful feature at Arxiv to distinguish a re-run is the [UPDATED] tag appended to submission titles. Our AI writer's internal 'recommender system' will need to consider carefully whether or not *[UPDATED]=='Played Out'*, particularly since it can (presumably) evaluate the re-warmed paper *much faster* than a hard-pressed science hack. In this respect, it has a notable advantage over humans, thanks to a naming convention that's likely to endure, at least at Arxiv.

Arxiv also provides information in the summary page about whether the paper has been identified as having 'significant cross-over' of text with another paper (often by the same authors), and this can also potentially be parsed into a 'duplicate/retread' status by an AI writer system in the absence of the [UPDATED] tag.

## Determining Diffusion

Like most journalists, our projected AI science writer is looking for unreported or under-reported news, in order to add value to the content stream it supports. In most cases, re-reporting science breakthroughs first featured in major outlets such as TechCrunch, The Verge and EurekaAlert *et al* is pointless, since such large platforms support their content with exhaustive publicity machines, virtually guaranteeing media saturation for the paper.

Therefore our AI writer must determine if the story is fresh enough to be worth pursuing.

The easiest way, in theory, would be to identify recent inbound links to the core research pages (summary, PDF, academic department website news section, etc.). In general, frameworks that can provide up-to-date inbound link information are not open source or low cost, but major publishers could presumably bear the SaaS expense as part of a newsworthiness-evaluation framework.

Assuming such access, our science writer AI is then faced with the problem that a great number of science-reporting outlets do not cite the papers they're writing about, even in cases where that information is freely available. After all, an outlet wants secondary reporting to link to them, rather than the source. Since, in many cases, they actually have obtained privileged or semi-privileged access to a research paper (see *The 'Social' Science Writer* below), they have a disingenuous pretext for this.

Thus our AI writer will need to extract actionable keywords from a paper and perform time-restricted searches to establish where, if anywhere, the story has already broken – and then evaluate whether any prior diffusion can be discounted, or whether the story is played out.

Sometimes papers provide supplementary video material on YouTube, where the 'view count' can serve as an index of diffusion. Additionally, our AI can extract images from the paper and perform systematic image-based searches, to establish if, where and when any of the images have been republished.

## Easter Eggs

Sometimes a 'dry' paper reveals findings that have profound and newsworthy implications, but which are underplayed (or even overlooked or discounted) by the authors, and will only be revealed by reading the entire paper and doing the math.

In rare cases, I believe, this is because the authors are far more concerned with reception in academia than the general public, perhaps because they feel (not always incorrectly) that the core concepts involved simply cannot be simplified enough for general consumption, despite the often hyperbolic efforts of their institutions' PR departments.

But about as often, the authors may discount or otherwise fail to see or to acknowledge the implications of their work, operating officially under 'scientific remove'. Sometimes these 'Easter eggs' are not positive indicators for the work, as mentioned above, and may be cynically obscured in complex tables of findings.

## Beyond Arxiv

It should be considered that parametrizing papers about computer science into discrete tokens and entities is going to be much easier at a domain such as Arxiv, which provides a number of consistent and templated 'hooks' to analyze, and does not require logins for most functionality.

Not all science publication access is open source, and it remains to be seen whether (from a practical or legal standpoint) our AI science writer can or will resort to evading paywalls through Sci-Hub; to using archiving sites to obviate paywalls; and whether it is practicable to construct similar domain-mining architectures for a wide variety of other science publishing platforms, many of which are structurally resistant to systematic probing.

It should be further considered that even Arxiv has rate limits which are likely to slow an AI writer's news evaluation routines down to a more 'human' speed.

## The 'Social' AI Science Writer

Beyond the open and accessible realm of Arxiv and similar 'open' science publishing platforms, even obtaining access to an interesting new paper can be a challenge, involving locating a contact channel for an author and approaching them to request to read the work, and even to obtain quotes (where pressure of time is not an overriding factor – a rare case for human science reporters these days).

This may entail automated traversing of science domains and the creation of accounts (you need to be logged in to reveal the email address of a paper's author, even on Arxiv). Most of the time, LinkedIn is the quickest way to obtain a response, but AI systems are currently prohibited from contacting members.

As to how researchers would receive email solicitations from a science writer AI – well, as with the meatware science-writing world, it probably depends on the influence of the outlet. If a putative AI-based writer from *Wired* contacted an author who was eager to disseminate their work, it's reasonable to assume that it might not meet a hostile response.

In most cases, one can imagine that the author would be hoping that these semi-automated exchanges might eventually summon a human into the loop, but it's not beyond the realm of possibility that follow-up VOIP interviews could be facilitated by an AI, at least where the viability of the article is forecasted to be below a certain threshold, and where the publication has enough traction to attract human participation in a conversation with an 'AI researcher'.

## Identifying News with AI

Many of the principles and challenges outlined here apply to the potential of automation across other sectors of journalism, and, as it ever was, identifying a potential story is the core challenge. Most human journalists will concede that actually writing the story is merely the last 10% of the effort, and that by the time the keyboard is clattering, the work is mostly over.

The major challenge, then, is to develop AI systems that can spot, investigate and authenticate a story, based on the many arcane vicissitudes of the news game, and traversing a huge range of platforms that are already hardened against probing and exfiltration, human or otherwise.

In the case of science reporting, the authors of new papers have as deep a self-serving agenda as any other potential primary source of a news story,and deconstructing their output will entail embedding prior knowledge about sociological, psychological and economic motivations. Therefore a putative automated science writer will need more than reductive NLP routines to establish where the news is today, unless the news domain is particularly stratified, as is the case with stocks, pandemic figures, sports results, seismic activity and other purely statistical news sources.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai