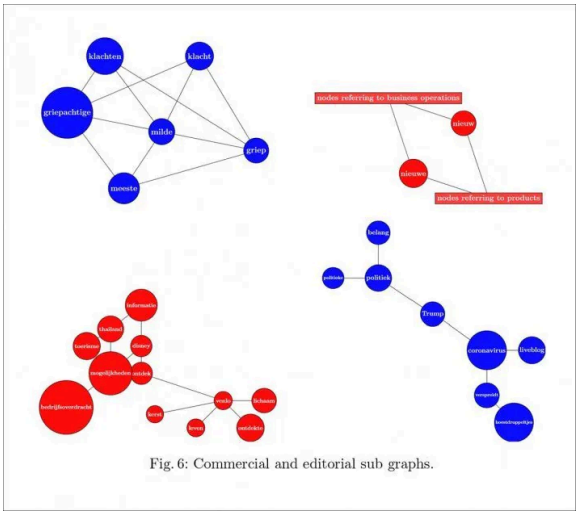
Identifying Sponsored Content in News Sites With Machine Learning

Originally published at https://www.unite.ai/identifying-sponsored-content-in-news-sites-with-machine-learning/

Researchers from the Netherlands have developed a new machine learning method that's capable of distinguishing sponsored or otherwise paid content within news platforms, to an accuracy of more than 90%, in response to growing interest from advertisers in 'native' advertising formats that are difficult to distinguish from 'real' journalistic output.

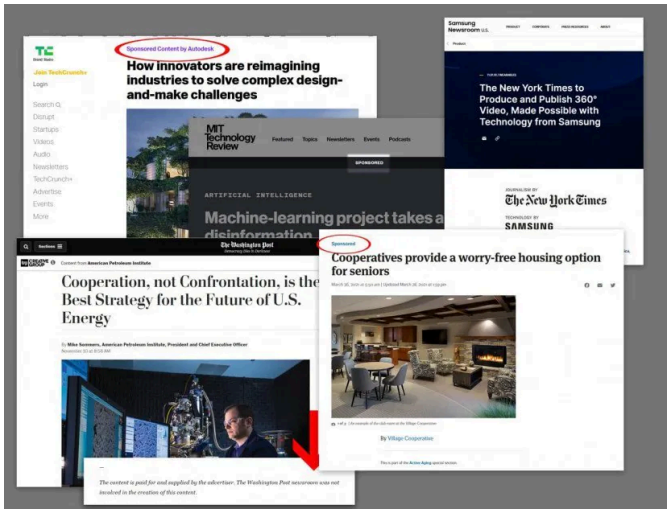
The new paper, titled *Distinguishing Commercial from Editorial Content in News*, comes from researchers at Leiden University.



Commercial (red) and editorial (blue) sub-graphs emerging from analysis of the data. Source: https://arxiv.org/pdf/2111.03916.pdf

The authors observe that though more serious publications, which can more easily dictate terms to advertisers, will make a reasonable effort to distinguish 'partner content' from the general run of news and analysis, the standards are slowly but inexorably shifting to increased integration between editorial and commercial teams on an outlet, which they consider an alarming and negative trend.

> *'The ability to disguise content, willingly or unwillingly, and the probability that advertorials are not recognized as such even if properly labelled is significant. Marketers call it native [advertising] for a reason.'*



Some current examples of native advertising, variously called 'partner content', 'brand content', and many other appellations designed to subtly obscure the distinction between native and commercially-placed content in journalistic platforms.

The work was carried out as part of a broader investigation into networked news culture at the ACED Reverb Channel, based in Amsterdam, which concentrates on data-driven analysis of evolving journalistic trends.

Acquiring Data

To develop source data for the project, the authors used 1,000 articles and 1,000 advertorials from four Dutch news outlets and classified them based on their textual features. Since the dataset was relatively modest in size, the authors avoided high-scale approaches such as BERT, and instead evaluated the effectiveness of more classical machine learning frameworks, including Support Vector Machine (SVM), LinearSVC, Decision Tree, Random Forest, K-Nearest Neighbor (K-NN), Stochastic Gradient Descent (SGD) and Naïve Bayes.

The Reverb Channel corpus was able to furnish the 1,000 necessary 'straight' articles, but the authors had to scrape advertorials directly from the four Dutch websites featured. The obtained data is available in limited form (due to copyright concerns) at GitHub, together with some of the Python code used to obtain and evaluate the data.

The four publications studied were the politically conservative Nu.nl, the more progressive Telegraaf, NRC, and the business journal De Ondernemer. Each publication was equally represented in the data.

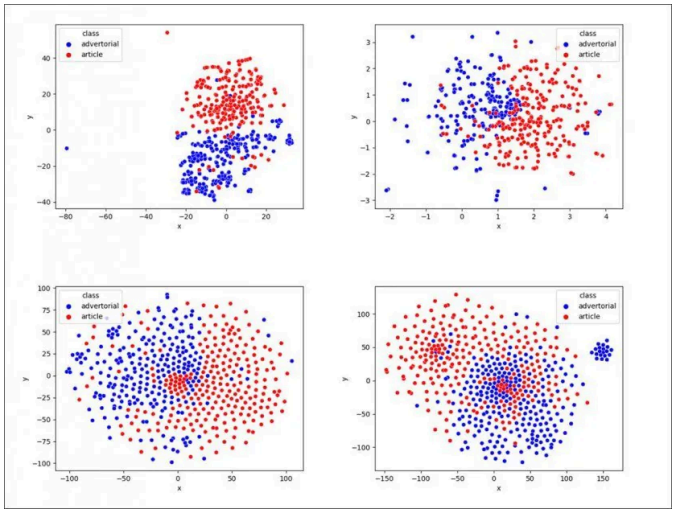
It was necessary to identify and discount potential 'leakers' in the lexicon formed by the research – words which might appear in both types of content with little distinction between their frequency and usage, in order to establish clear patterns for genuinely native and sponsored content.

Results

Across the methods tested for identification, the best results were obtained by SVM, linearSVC, Random Forest and SGD. Therefore the researchers proceeded to use SVM in further analysis.

	SVM	Linear SVC	Decision Tree	Random Forest	k-NN	SGD	Naive Bayes	
Nu.nl	0.84	0.84	0.72	0.83	0.52	0.84	0.72	0.76 ± 0.12
NRC	0.93	0.95	0.75	0.82	0.52	0.95	0.81	0.82 ± 0.15
Ondernemer	0.76	0.68	0.54	0.55	0.51	0.65	0.56	0.61 ± 0.09
Telegraaf	0.85	0.84	0.66	0.84	0.46	0.83	0.73	0.74 ± 0.14
	0.85 ±0.06	0.83 ±0.10	0.67 ±0.08	0.76 ±0.12	0.5 ±0.02	0.82 ±0.11	0.71 ±0.09	

The best model approach for extracting classification across the corpus exceeded 90% accuracy, though the researchers note that obtaining a clear classification becomes more difficult when dealing with B2B-oriented publications, where the lexical overlap between perceived 'real' and 'sponsored' content is excessive – perhaps because the native style of business language is already more subjective than the general run of reporting and analysis conventions, and can more easily conceal an agenda.



t-Distributed Stochastic Neighbor Embedding (t-SNE) plots for separation of real and sponsored content across the four publications.

Is Sponsored Content 'Fake News'?

The authors' research suggests that their project is novel in the field of news content analysis. Frameworks capable of identifying sponsored content could pave the way to developing year-on-year monitoring of the balance between objective journalism and the growing tranche of 'native advertising' which sits in almost the same context in most publications, using the same visual cues (CSS stylesheets and other formatting) as general content.

In a certain sense, the frequent lack of obvious context for sponsored content is emerging as a sub-field of the study of 'fake news'. Though most publishers recognize the need for separation of 'church and state', and the obligation to provide readers with clear divisions between paid and organically-generated content, the realities of the post-print journalistic scene, and increased dependence on advertisers, have turned the de-emphasis of sponsored indicators into a fine art in UI psychology. Sometimes the rewards of running sponsored content are tempting enough to risk a major optical disaster.

In 2015 the social media and competitive benchmarking platform Quintly offered an AI-based detection method to determine if a post on Facebook is sponsored, claiming an accuracy rate of 96%. The following year, a study from the University of Georgia contended that the way publishers handle the declaration of sponsored content could be 'complicit with deception'.

In 2017 MediaShift, an organization that examines the intersection between media and technology, observed the growing extent to which the New York Times monetizes its operations through its branded content studio, T Brand Studio, claiming diminishing levels of transparency around sponsored content, with the tacitly intentional result that readers cannot easily tell whether or not content is organically generated.

In 2020, another research initiative from the Netherlands developed machine learning classifiers to automatically identify Russian state-funded news appearing in Serbian news platforms. Further, it was estimated in 2019 that Forbes' 'media content solutions' account for 40% of its total revenue through BrandVoice, the content studio launched by the publisher in 2010.

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