What AI Can Tell Us About Hidden Agendas in the News

Originally published at https://www.unite.ai/what-ai-can-tell-us-about-hidden-agendas-in-the-news/

ChatGPT-style models are being trained to detect what a news article really *thinks about an issue – even when that stance is buried under quotes, framing, or (sometimes disingenuous) 'neutrality'. By breaking articles into segments such as headlines, leads, and quotations, a new system learns to spot bias even in long-form professional journalism.*

The ability to understand the true viewpoint of a writer or speaker – a pursuit known in the literature as *stance detection* – addresses one of the most difficult interpretive problems in language: gleaning the intent from content that may be designed to hide or obscure it.

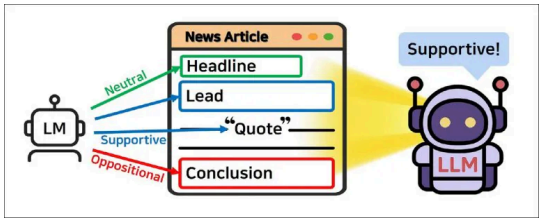
From Jonathan Swift's *A Modest Proposal*, to recent performances by political actors borrowing the polemics of their ideological opponents, the surface of a statement is no longer a reliable indicator of its intent; the rise of irony, trolling, disinformation and strategic ambiguity has made it harder than ever to pinpoint what side a text actually lands on, or whether it lands at all.

Often, what goes unsaid carries as much weight as what is stated, and simply choosing to cover a topic can signal the author's position.

That makes the task of automatic stance detection unusually challenging, since an effective detection system needs to do more than tag isolated sentences as 'supportive' or 'oppositional': instead it must iterate through layers of meaning, weighing small cues against the shape and drift of the whole article; and this is harder in long-form journalism, where tone may shift and where opinion may rarely be stated outright.

Agents For Change

To address some of these issues, researchers in South Korea have developed a new system called JOA-ICL (*Journalism-guided Agentic In-Context Learning*) for detecting the stance of long-form news articles.



The core idea behind JoA-ICL is that article-level stance is inferred by aggregating segment-level predictions produced by a separate language model agent. Source: https://arxiv.org/pdf/2507.11049

Instead of judging an article as a whole, JOA-ICL breaks it into structural parts (headline, lead, quotations, and conclusion) and assigns a smaller model to label each one. These local predictions are then passed to a larger model, which uses them to determine the article's overall stance.

The method was tested on a newly compiled Korean dataset containing 2,000 news articles annotated for both article-level and segment-level stance. Each article was labeled with input from a journalism expert, reflecting how stance is distributed across the structure of professional news-writing.

According to the paper, JOA-ICL outperforms both prompting-based and fine-tuned baselines, demonstrating particular strength in detecting supportive stances (which models with a similar ambit tend to miss). The method also proved effective when applied to a German dataset under matched conditions, indicating that its principles are potentially resilient to language forms.

The authors state:

> *'Experiments show that JOA-ICL outperforms existing stance detection methods, highlighting the benefits of segment-level agency in capturing the overall position of long-form news articles."*

The new paper is titled *Journalism-Guided Agentic In-Context Learning for News Stance Detection*, and comes from various faculties at Seoul's Soongsil University, as well as KAIST's Graduate School of Future Strategy.

Method

Part of the challenge of AI-augmented stance detection is logistical, and related to how much signal a machine learning system can retain and collate at one time, at the current state-of-the-art.

News articles tend to avoid direct statements of opinion, relying instead on an *implicit* or *assumed* stance, signaled through choices about which sources to quote, how the narrative is framed, and what details are left out, among many other considerations.

Even when an article does take a clear position, the signal is often scattered across the text, with different segments pointing in different directions. Since language models (LMs) still struggle with limited context windows, this can make it difficult for models to assess stance in the way they do with shorter content (such as tweets and other short-form social media), where the relationship between the text and the target is more explicit.

Therefore standard approaches often fall short when applied to full-length journalism; a case where ambiguity is a feature rather than a flaw.

The paper states:

> *'To address these challenges, we propose a hierarchical modeling approach that first infers the stance at the level of smaller discourse units (e.g., paragraphs or sections), and subsequently integrates these local predictions to determine the overall stance of the article.*
>
> *'This framework is designed to retain local context and capture dispersed stance cues in assessing how different parts of a news story contribute to its overall position on an issue.'*

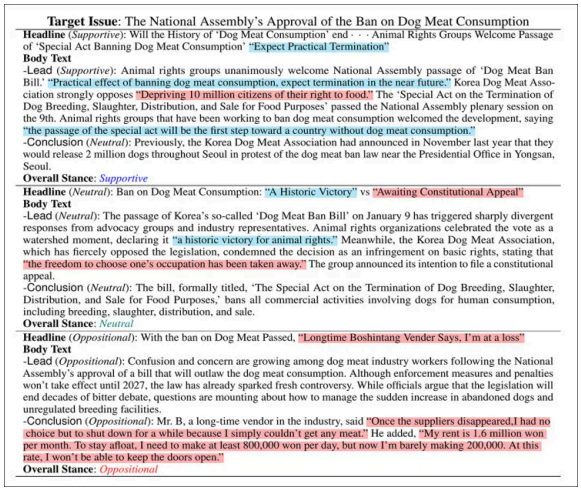
To this end, the authors compiled a novel dataset titled *K-NEWS-STANCE*, drawn from Korean news coverage between June 2022 and June 2024. Articles were first identified through BigKinds, a government-backed metadata service operated by the Korea Press Foundation, and full texts were retrieved using the Naver News aggregator API. The final dataset comprised 2,000 articles from 31 outlets, covering 47 nationally-relevant issues.

Each article was annotated twice: once for its overall stance toward a given issue, and again for individual segments; specifically the *headline*, *lead*, *conclusion*, and *direct quotations*.

The annotation was led by journalism expert Jiyoung Han, also the paper's third author, who guided the process through the use of established cues from media studies, such as source selection, lexical framing, and patterns of quotation. By these means a total of 19,650 segment-level stance labels were obtained.

To ensure the articles contained meaningful viewpoint signals, each was first classified by genre, and only those labeled as analysis or opinion (where subjective framing is more likely to be found) were used for stance annotation.

Two trained annotators labeled all articles, and were instructed to consult related articles in case the stance was unclear, with disagreements resolved through discussion and additional review.



Target Issue: The National Assembly's Approval of the Ban on Dog Meat Consumption

Headline (*Supportive*): Will the History of 'Dog Meat Consumption' end · · · Animal Rights Groups Welcome Passage of 'Special Act Banning Dog Meat Consumption' "Expect Practical Termination"

Body Text

-Lead (*Supportive*): Animal rights groups unanimously welcome National Assembly passage of 'Dog Meat Ban Bill.' "Practical effect of banning dog meat consumption, expect termination in the near future." Korea Dog Meat Association strongly opposes "Depriving 10 million citizens of their right to food." The 'Special Act on the Termination of Dog Breeding, Slaughter, Distribution, and Sale for Food Purposes' passed the National Assembly plenary session on the 9th. Animal rights groups that have been working to ban dog meat consumption welcomed the development, saying "the passage of the special act will be the first step toward a country without dog meat consumption."

-Conclusion (*Neutral*): Previously, the Korea Dog Meat Association had announced in November last year that they would release 2 million dogs throughout Seoul in protest of the dog meat ban law near the Presidential Office in Yongsan, Seoul.

Overall Stance: *Supportive*

Headline (*Neutral*): Ban on Dog Meat Consumption: "A Historic Victory" vs "Awaiting Constitutional Appeal"

Body Text

-Lead (*Neutral*): The passage of Korea's so-called 'Dog Meat Ban Bill' on January 9 has triggered sharply divergent responses from advocacy groups and industry representatives. Animal rights organizations celebrated the vote as a watershed moment, declaring it "a historic victory for animal rights." Meanwhile, the Korea Dog Meat Association, which has fiercely opposed the legislation, condemned the decision as an infringement on basic rights, stating that "the freedom to choose one's occupation has been taken away." The group announced its intention to file a constitutional appeal.

-Conclusion (*Neutral*): The bill, formally titled, 'The Special Act on the Termination of Dog Breeding, Slaughter, Distribution, and Sale for Food Purposes,' bans all commercial activities involving dogs for human consumption, including breeding, slaughter, distribution, and sale.

Overall Stance: *Neutral*

Headline (*Oppositional*): With the ban on Dog Meat Passed, "Longtime Boshintang Vender Says, I'm at a loss"

Body Text

-Lead (*Oppositional*): Confusion and concern are growing among dog meat industry workers following the National Assembly's approval of a bill that will outlaw the dog meat consumption. Although enforcement measures and penalties won't take effect until 2027, the law has already sparked fresh controversy. While officials argue that the legislation will end decades of bitter debate, questions are mounting about how to manage the sudden increase in abandoned dogs and unregulated breeding facilities.

-Conclusion (*Oppositional*): Mr. B, a long-time vendor in the industry, said "Once the suppliers disappeared,I had no choice but to shut down for a while because I simply couldn't get any meat." He added, "My rent is 1.6 million won per month. To stay afloat, I need to make at least 800,000 won per day, but now I'm barely making 200,000. At this rate, I won't be able to keep the doors open."

Overall Stance: *Oppositional*

Sample entries from the K-NEWS-STANCE dataset, translated into English. Only the headline, lead, and quotations are shown; full body text is omitted. Highlighting indicates stance labels for quotations, with blue for supportive and red for oppositional. Please refer to the cited source PDF for clearer rendition.

JoA-ICL

Rather than treating an article as a single block of text, the authors' proposed system divides it into key structural parts: headline, lead, quotations, and conclusion, assigning each of these to a language model agent, which labels the segment as *supportive*, *oppositional*, or *neutral*.

These local predictions are passed to a second agent that decides the article's overall stance, with the two agents coordinated by a controller that prepares the prompts and gathers the results.

Thus JoA-ICL adapts in-context learning (where the model learns from examples in the prompt) to match the way that professional news stories are written, using segment-aware prompts instead of a single generic input.

(Please note that most of the examples and illustrations in the paper are lengthy and difficult to reproduce legibly in an online article. We therefore implore the reader to examine the original source PDF)

Data and Tests

In tests, the researchers used macro F1 and accuracy to evaluate performance, averaging results over ten runs with random seeds from 42 to 51 and reporting standard error. Training data was used to fine-tune baseline models and segment-level agents, with few-shot samples selected through similarity search using KLUE-RoBERTa-large.

Tests were run over three RTX A6000 GPUs (each with 48GB of VRAM), using Python 3.9.19, PyTorch 2.5.1, Transformers 4.52.0, and vLLM 0.8.5.

GPT-4o-mini, Claude 3 Haiku, and Gemini 2 Flash were utilized via API, at a temperature of 1.0 and with max tokens set to 1000 for chain-of-thought prompts, and 100 for others.

For full fine-tuning of Exaone-3.5-2.4B, the AdamW optimizer was used at a 5e-5 learning rate, with 0.01 weight decay, 100 warmup steps, and with the data trained for 10 epochs at a batch size of 6.

For baselines, the authors used RoBERTa, fine-tuned for article-level stance detection; Chain-of-Thought (CoT) Embeddings, an alternate tuning of RoBERTa for the assigned task; LKI-BART, an encoder-decoder model that adds contextual knowledge from a large language model by prompting it with both the input text and the intended stance label; and PT-HCL, a method that uses contrastive learning to separate general features from those specific to the target issue:

RoBERTa		CoT Embeddings		LKI-BART		PT-HCL	
Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
0.594±0.029	0.577±0.046	0.582±0.028	0.562±0.039	0.545±0.011	0.538±0.014	**0.617±0.007**	**0.618±0.008**

(a) Existing methods

Method	CoT	Few-shot	GPT-4o-mini		Gemini-2.0-flash		Claude-3-haiku		Exaone-2.4b (finetuned)	
			Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Baseline			0.594±0.002	0.577±0.002	0.636±0.001	0.628±0.002	**0.568±0.002**	**0.538±0.002**	**0.554±0.003**	**0.544±0.003**
	√		**0.597±0.002**	**0.581±0.003**	**0.661±0.002**	**0.657±0.002**	0.561±0.002	0.533±0.003	0.54±0.001	0.539±0.003
		√	0.565±0.003	0.54±0.003	0.635±0.003	0.631±0.004	0.545±0.004	0.523±0.005	0.445±0.004	0.431±0.003
	√	√	0.526±0.003	0.494±0.003	0.641±0.005	0.635±0.004	0.541±0.004	0.51±0.004	0.456±0.002	0.443±0.002
JoA-ICL (Oracle)			0.724±0.001	0.706±0.001	0.758±0.001	0.753±0.001	0.791±0.002	0.789±0.002	**0.837±0.004**	**0.837±0.004**
	√		0.716±0.002	0.692±0.002	**0.778±0.001**	**0.776±0.001**	0.74±0.001	0.732±0.001	0.813±0.003	0.821±0.002
		√	**0.796±0.003**	**0.798±0.003**	0.772±0.003	0.769±0.003	**0.815±0.001**	**0.816±0.001**	0.344±0.004	0.308±0.004
	√	√	0.747±0.001	0.74±0.002	0.777±0.003	0.773±0.003	0.794±0.002	0.797±0.002	0.338±0.005	0.3±0.006
JoA-ICL (RoBERTa)			0.571±0.002	0.53±0.001	0.633±0.003	0.619±0.004	0.591±0.004	0.577±0.004	0.584±0.002	0.581±0.001
	√		0.553±0.001	0.509±0.001	0.633±0.001	0.626±0.001	0.579±0.003	0.552±0.004	**0.601±0.006**	**0.599±0.006**
		√	**0.607±0.006**	**0.602±0.007**	0.662±0.001	0.657±0.001	**0.639±0.002**	**0.638±0.004**	0.354±0.003	0.331±0.003
	√	√	0.591±0.002	0.57±0.002	**0.678±0.002**	**0.672±0.002**	0.61±0.004	0.608±0.004	0.332±0.001	0.308±0.001

(b) LLM in-context learning methods

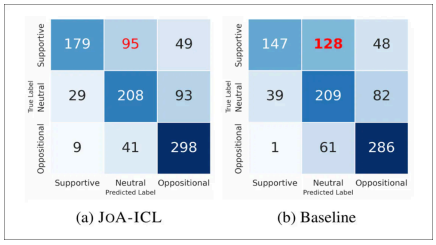
Performance of each model on the K-NEWS-STANCE test set for overall stance prediction. Results are shown as macro F1 and accuracy, with the top score in each group in bold.

JOA-ICL achieved the best overall performance across both accuracy and macro F1, an advantage evident across all three model backbones tested: GPT-4o-mini, Claude 3 Haiku, and Gemini 2 Flash.

The segment-based method consistently outperformed all other approaches, with, the authors observe, a notable edge in detecting supportive stances, a common weakness in similar models.

Baseline models performed worse overall. RoBERTa and Chain-of-Thought variants struggled with nuanced cases, while PT-HCL and LKI-BART fared better, while still trailing JOA-ICL across most categories. The most accurate single result came from JOA-ICL (Claude), with 64.8% macro F1 and 66.1% accuracy.

The image below shows how often the models got each label right or wrong:



(a) JoA-ICL (b) Baseline

Confusion matrices comparing the baseline and JoA-ICL, showing that both methods struggle most with detecting 'supportive' stances.

JoA-ICL did better overall than the baseline, getting more labels correct in every category. However, both models struggled most with supportive articles, and the baseline misclassified nearly half, often mistaking these for neutral.

JoA-ICL made fewer mistakes but showed the same pattern, reinforcing that 'positive' stances are harder for models to spot.

To test whether JoA-ICL works beyond the confines of the Korean language, the authors ran it on CheeSE, a German dataset for article-level stance detection. Since CheeSE lacks segment-level labels, the researchers used *distant supervision*, wherein every segment was assigned the same stance label as the full article.

Method	Accuracy	F1
RoBERTa	0.526 ± 0.014	0.442 ± 0.034
CoT Embeddings	0.424 ± 0.01	0.198 ± 0.003
LKI-BART	0.466 ± 0.01	0.34 ± 0.011
PT-HCL	0.517 ± 0.021	0.39 ± 0.048

(a) Fine-tuned models

LLM	Method	Accuracy	F1
GPT-4o-mini	Zero-shot	0.549 ± 0.002	0.518 ± 0.002
	JoA-ICL	0.593 ± 0.002	0.538 ± 0.002
Gemini-2.0-flash	Zero-shot	0.566 ± 0.001	0.54 ± 0.001
	JoA-ICL	$\mathbf{0.614 \pm 0.001}$	$\mathbf{0.579 \pm 0.001}$
Claude-3-haiku	Zero-shot	0.487 ± 0.002	0.444 ± 0.002
	JoA-ICL	0.59 ± 0.001	0.527 ± 0.001

(b) In-context learning

Stance detection results on the German-language CheeSE dataset. JoA-ICL consistently improves over zero-shot prompting across all three LLMs and outperforms fine-tuned baselines, with Gemini-2.0-flash yielding the strongest overall performance.

Even under these 'noisy' conditions, JoA-ICL outperformed both fine-tuned models and zero-shot prompting. Of the three backbones tested, Gemini-2.0-flash gave the best results.

Conclusion

Few tasks in machine learning are more politically charged than stance prediction; yet it is often handled in cold, mechanical terms, while more attention is given to less complex issues in generative AI, such as video and image creation, which trigger far louder headlines.

The most encouraging development in the new Korean work is that it offers a significant contribution to analysis of *full-length* content, rather than tweets and short-form social media, whose incendiary effects are more quickly forgotten than a treatise, essay or other significant work.

One notable omission in the new work and (as far as I can tell) in the stance prediction corpus in general is the lack of consideration given to *hyperlinks*, which frequently stand in for quotes as optional resources for readers to learn more about a subject; yet it must be clear that the choice of such URLs is potentially very subjective and even political.

That said, the more prestigious the publication, the less likely that it will include *any links at all* that guide the viewer away from the host domain; this, together with diverse other SEO uses and abuses of hyperlinks, makes them more difficult to quantify than explicit quotes, titles, or other parts of an article that may seek, consciously or not, to influence the reader's opinion.

First published Wednesday, July 16, 2025

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

Discover More