IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

FinNews: Detecting Second-Order Supply-Chain Shocks Using Knowledge-Enriched Cross-Encoders

Ramya Bharathy[1], Abhilash D[2], Kamalesh Kumar K*[3], Deepashri[4]

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamil Nadu, India[1-4]

*Corresponding Author

Code: https://github.com/kamalesh55555/FinNews

Abstract: Investor news filters usually match company names, tickers or keywords, or compare embeddings of a query and a headline that were computed separately. These methods handle company announcements well but miss a different kind of risk: a problem at a supplier, a raw-material source or a transport route that affects a company without the headline ever naming it. We call these second-order shocks. FinNews detects them by describing each monitored company with a three-part text profile (identity, sector and upstream inputs) and scoring every headline–profile pair with a cross-encoder, which reads both texts together instead of relying on name overlap. The scorer is a DeBERTa-v3-small model initialized from a natural language inference (NLI) checkpoint and fine-tuned on 66,606 pairs covering 55 companies in 10 sectors, split chronologically into training, validation and holdout sets. On the 6,661-pair holdout set at a threshold of 0.40, the model reaches 74.06% accuracy, 71.67% precision and 79.16% recall; it retains 77.78% of second-order shocks and rejects 69.01% of irrelevant headlines. A 12-layer MiniLM passage-ranking baseline, scored on an earlier holdout without enriched profiles, retains 47.60% of second-order shocks. Because the backbone and the profile format changed together between these runs, the gain cannot yet be assigned to either. On a small supplementary live set collected on 10–12 October 2026 and scored at a threshold of 0.35, the model reaches 94.83% accuracy and retains 20 of 24 second-order shocks (83.33%). We also report an evaluation error in which a second sigmoid applied to already-normalized scores makes every pair pass the threshold, and we describe a CPU-only deployment that scores a pair in 46–66 ms.

Keywords: financial news filtering, supply-chain risk, second-order events, cross-encoder, DeBERTa-v3, natural language inference

I. INTRODUCTION

Financial news arrives faster than anyone can read it, and most of it has nothing to do with a given portfolio. Commercial trackers and news aggregators cope mainly with first-order filtering: a story is kept if it contains a company name, a ticker such as NVDA or HMC, or a sector tag.

This works for earnings releases, acquisitions and executive changes. It fails for a large class of supply-chain events in which something goes wrong upstream and the affected company is never named. A restriction on lithium brine pumping in the Atacama, or a fire at a Japanese plant that makes battery separator film, threatens production at electric-vehicle makers such as Honda and Tesla. Water rationing in Taiwan's Hsinchu Science Park, or a delay in extreme-ultraviolet optics at a European subcontractor, constrains TSMC and, through its advanced packaging capacity, Nvidia and the cloud providers that buy its accelerators. Headlines about such events read like mining, weather or regional industry news, and name or keyword filters have zero recall on them by construction.

We separate two kinds of news. A first-order event names the company or one of its products. A second-order shock reports a disruption to an input, component or piece of infrastructure the company depends on, and the company itself does not appear in the text. Finding the second kind requires knowing what each company depends on. One option is to keep that knowledge in a graph and traverse it for every incoming headline, which adds query-time cost and needs a maintained graph. Relation-extraction methods avoid the graph but need supplier and customer to be mentioned together [1], which an isolated commodity headline does not satisfy. FinNews instead writes the dependencies into each company's profile text and lets a cross-encoder read the profile and the headline together, so that the dependency lookup a graph traversal would perform in several hops is folded into a single forward pass.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

The contributions of this paper are:

1. A company-conditioned relevance formulation in which a three-part profile (identity, sector, upstream inputs) is used as the query, built for 55 companies in 10 sectors.
2. A comparison of two MS MARCO MiniLM cross-encoders with an NLI-initialized DeBERTa-v3-small cross-encoder, with recall reported separately for second-order events.
3. A description of a double-sigmoid error in evaluating single-label cross-encoders, a threshold sweep from 0.35 to 0.50 on the holdout set, and a supplementary evaluation on a live set from October 2026.
4. A full-stack implementation (RSS ingestion, deduplication, database, web interface) that runs on a consumer CPU.

We state at the outset that the comparison with the MiniLM baselines is not a controlled experiment; the reasons are given in Section VI.

II. RELATED WORK

Financial text classification. Malo et al. [2] released FinancialPhraseBank, a set of sentence-level sentiment annotations, and FinBERT [3] adapts BERT to financial sentiment with it. These models score the polarity of a text for every reader alike. Jacobs et al. [4] classify company-specific news into ten economic event types, and Pei et al. [5] build a risk-event taxonomy and detector for Bloomberg news; both assume the article is about the company in question. Yıldırım et al. [6] separate market-moving "hot news" from routine articles, and Del Corro and Hoffart [7] derive a relevance signal from the attention weights of stock-movement models. Neither is conditioned on a particular firm's inputs, and the latter depends on noisy price co-movement.

Supply-chain text and shock propagation. Wichmann et al. [1] build supply-chain maps from news using named-entity recognition and a BiLSTM relation classifier, which requires both firms in one sentence. Caliari et al. [8] classify supply-chain disruption news and Vishnuthilak et al. [9] detect disruptions through sentiment, in both cases without reference to a specific holding. Sato and Mizuno [10] study how shocks propagate along supplier–customer links, using inter-firm databases rather than news streams.

Cross-encoders and calibration. Cross-encoders attend jointly over both inputs and tend to beat bi-encoders on pairwise tasks, at a higher cost per pair [11]. DeBERTa [12] and DeBERTaV3 [13] separate content and relative position in the attention computation and, in V3, use replaced-token-detection pretraining. Desai and Durrett [14] show that fine-tuned transformers can be poorly calibrated, particularly under distribution shift, which is why we do not treat our scores as probabilities.

We did not find prior work that combines a holding-specific profile listing upstream dependencies, a cross-encoder, and separate recall measurement for target-absent events.

III. METHODOLOGY

A. Problem Formulation

A portfolio contains M monitored companies, each with a text profile Q. For an incoming headline D the model produces a relevance score between 0 and 1:

$$0 \le s = f(Q, D) \le 1 \tag{1}$$

The headline is kept for that company if s is at least a threshold τ and discarded otherwise. Kept headlines are then routed by a plain string check $\psi(D)$, which is 1 when the company name or ticker appears in the headline and 0 otherwise:

$$\text{Category}(Q, D) = \begin{cases} \text{DIRECT_NEWS} & \text{if } s \ge \tau \text{ and } \psi(D) = 1 \\ \text{BUTTERFLY_SHOCK} & \text{if } s \ge \tau \text{ and } \psi(D) = 0 \\ \text{NOISE} & \text{if } s < \tau \end{cases} \tag{2}$$

The score comes from a regression head trained with mean squared error (MSE) on 0/1 labels. It works for ranking and thresholding, but it is not a calibrated probability.

B. Enriched Company Profiles

A bare ticker or company name gives a transformer nothing to match against a headline about battery separator film or EUV lithography. We therefore describe each company with three segments joined by a pipe character:

$$Q = E \mid S \mid U \tag{3}$$

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

E is the identity (company name and ticker). *S* is the operational sector and main revenue activities. *U* lists tier-1 and tier-2 upstream inputs: raw materials, specialty chemicals, equipment, logistics routes and hardware sub-assemblies. Table I shows four of the 55 profiles. Profiles are written by hand and limited to inputs that were established before the dates of the headlines in our data, to avoid leaking later knowledge into historical evaluation. Section VI discusses the limits of this approach.

TABLE I. FOUR OF THE 55 COMPANY PROFILES

Ticker	Identity (E)	Sector and activities (S)	Upstream inputs (U)
TSM	Taiwan Semiconductor (TSMC)	Silicon wafer fabrication, advanced foundry, logic nodes	EUV lithography, raw silicon wafers, photoresist chemicals, ultra-pure neon, argon gas, cleanroom filtration
NVDA	Nvidia	GPU acceleration, AI hardware, high-performance computing	TSMC CoWoS advanced packaging, HBM3e memory modules, ABF substrate, liquid-cooling heat sinks
HMC	Honda Motor	Passenger vehicles, motorcycles, hybrid engines, electric mobility	Lithium carbonate, cobalt, nickel cathode, automotive MCUs, battery separator film, sheet steel, synthetic rubber
ALK	Alaska Air Group	Regional and transcontinental passenger aviation, air cargo	Aviation jet fuel, commercial jet engines, aircraft maintenance tooling, de-icing glycols

C. Cross-Encoder

A bi-encoder encodes the query and the document separately and compares the two vectors, which discards token-level interaction. A cross-encoder instead feeds both texts to the transformer as one sequence, [CLS] profile [SEP] headline [SEP], truncated to 256 tokens, so every profile token can attend to every headline token (Fig. 1). This costs more per pair but lets the model notice that "jet fuel" in a profile matches "jet fuel crack spreads" in a headline.

We initialize from cross-encoder/nli-deberta-v3-small and replace its three-way NLI head (contradiction, entailment, neutral) with a single linear output. We chose this checkpoint for two reasons. First, DeBERTa keeps content and relative position as separate vectors inside attention, which may help where direction matters, for instance an export ban that cuts a company's supply versus a routine delivery. Second, NLI pretraining teaches a model to judge whether one text bears on another, which is closer to "does this event affect this company's operations" than the query–passage matching learned from MS MARCO, which tends to reward lexical overlap. We treat both as motivation only; Section VI explains why our experiments cannot confirm either. The backbone has about 44 million parameters, and with the SentencePiece embedding matrix the checkpoint has about 141 million (567 MB in 32-bit floats).

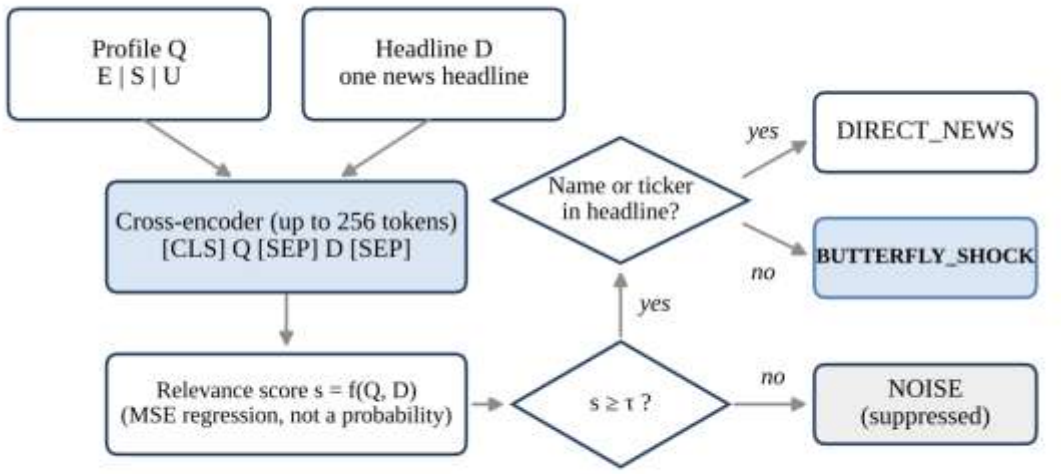


Fig. 1. Scoring and routing of one headline–profile pair. A score and a name check route every pair to one of three drawers.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

D. System Implementation

The model is deployed as a four-tier web application.

Tier 1, ingestion. An asynchronous worker polls nine RSS feeds covering equities, commodities, shipping and energy (including WSJ Markets and Business, MarketWatch, Google News Business and Technology, FreightWaves, SupplyChainBrain and OilPrice), about 295 articles per cycle.

Tier 2, deduplication and concurrency. The MD5 hash of each article URL, together with feed GUIDs, identifies articles already scored, which are skipped without running the model. A non-blocking lock returns HTTP 409 if two sync requests overlap.

Tier 3, scoring. The inference engine scores new (profile, headline) pairs in batches at $\tau = 0.40$ and assigns each to DIRECT_NEWS, BUTTERFLY_SHOCK or NOISE.

Tier 4, storage and interface. A FastAPI backend stores watchlist holdings, deduplicated articles and shock events (score ≥ 0.40) in Neon serverless PostgreSQL through SQLAlchemy 2.0. A React 19 front end built with Vite and Tailwind CSS shows three drawers: direct news, second-order shocks and suppressed noise.

IV. EXPERIMENTAL SETUP

A. Data

The data come from the Daily Financial News for 6000+ Stocks corpus (Kaggle, curated by Miguel Aenlle) and the FNSPID time-series news benchmark; FinancialPhraseBank [2] was used for financial vocabulary calibration. From 6,552 NYSE, NASDAQ and AMEX tickers we selected 55 high-volume companies in 10 sectors, chosen so that dependencies cross sector lines: semiconductors (e.g., TSM, NVDA, INTC), battery and critical minerals (ALB, FCX), automotive and clean energy (HMC, TSLA), energy and oil (XOM, SLB), logistics and transport (FDX, ALK), banking and finance (JPM, BLK), retail and consumer (KR, CLX), healthcare and pharmaceuticals (LLY, MDT), enterprise technology (MSFT, PANW) and utilities and grid (DUK, SO).

Every pair belongs to one of three types (Table II). A second-order positive differs from a direct positive in that the company is absent from the headline, which describes a problem with an input listed in the company's *U* field.

TABLE II. THE THREE PAIR TYPES

Pair type	Label	Rule	Example
Direct positive	y = 1	The headline names the target company.	AVGO paired with "31,377 Broadcom Inc. $AVGO Shares Bought by Oak Grove Capital LLC"
Second-order positive	y = 1	The company is absent from the headline, which describes an upstream supplier, material or infrastructure problem listed in the company's U field.	HMC paired with "Battery Separator Film Shortage Emerges in Asia as Key Japanese Polyethylene Membrane Extruder Catches Fire"; NVDA paired with "Advanced CoWoS Packaging Capacity Bottlenecks Expected to Tighten Through First Half of 2027"
Irrelevant negative	y = 0	A company profile paired with a real headline from a non-overlapping industry.	The grocer KR paired with a headline on TSMC's AI-accelerator revenue

Pairs were sorted by publication time and split 80/10/10 into training, validation and holdout sets, so no training headline is later than a test headline (Table III). The MiniLM baselines were scored on an earlier version of the holdout (v1) and the DeBERTa model on the final version (v2). Classes are close to balanced (49.95% positive over the whole corpus), but second-order shocks make up only 4.74% of the corpus (3,157 of 66,606 pairs), which is why we report recall for them separately.

A supplementary live set for 25 tickers was collected from news published on 10–12 October 2026. Its dataset description lists 115 pairs (64 positive, of which 40 direct and 24 second-order, or 20.87%, and 51 negative), whereas the evaluation log of the scoring run counts 116 pairs with 52 negatives; the positive counts agree. We report the live-set metrics exactly as logged and return to this discrepancy in Sections V-G and VI.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

TABLE III. DATASET PARTITIONS

Partition	Pairs	Positives	Negatives	Direct positives	Second-order positives	Use
Training	53,284	26,642	26,642	24,115	2,527	Fine-tuning
Validation	6,661	3,315	3,346	3,000	315	Checkpoint selection
Holdout v1	6,661	3,287	3,374	2,974	313	MiniLM baselines
Holdout v2	6,661	3,315	3,346	3,000	315	DeBERTa evaluation
Live set	115 (116 in log)	64	51 (52 in log)	40	24	Supplementary

B. Training and Baselines

We fine-tuned the DeBERTa model for 4 epochs on a single Google Colab NVIDIA T4 GPU with batch size 16, AdamW at a learning rate of 2×10^{-5}, 300 linear warmup steps and MSE loss against 0/1 targets, saving a checkpoint every 250 steps and selecting on the validation set. The baselines are the public passage-ranking cross-encoders cross-encoder/ms-marco-MiniLM-L-6-v2 (22.7M parameters) and cross-encoder/ms-marco-MiniLM-L-12-v2 (33.4M parameters). The experiment records used for this paper do not state whether the MiniLM baselines were fine-tuned on the same training data or used as released, so we treat the comparison as indicative only (Section VI).

C. Evaluation Metrics

We report accuracy, precision, recall, F1 and specificity (the share of irrelevant headlines rejected, called noise specificity below). Because direct positives outnumber second-order positives by roughly ten to one, we also report recall separately on each group. Second-order F1 combines second-order recall with the overall precision, since negative pairs are not tagged by type and a subset-specific precision cannot be computed:

$$F_{1,\mathrm{SO}} = 2P\,R_{\mathrm{SO}} / (P + R_{\mathrm{SO}}) \tag{4}$$

where P is the overall precision and R_{SO} is the second-order recall. Decision thresholds of 0.35, 0.40, 0.45 and 0.50 are applied directly to the model output.

V. RESULTS AND DISCUSSION

A. Comparison with Passage-Ranking Baselines

Table IV gives the holdout results and Fig. 2 sets the two models side by side at the same threshold. The 6-layer MiniLM at $\tau = 0.50$ has good precision (78.14%) and noise rejection (84.08%) but keeps only 118 of 313 second-order shocks (37.70%). Lowering the threshold to 0.40 and doubling the depth to 12 layers recovers 31 more (149, or 47.60%), after which the MiniLM results appear to level off, with 1,261 relevant headlines still missed. A plausible reason is that MS MARCO training rewards lexical overlap between query and passage, so a headline that shares no words with the profile receives a low score; we did not test this explanation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

TABLE IV. HOLDOUT RESULTS

Metric	MiniLM-L6 $\tau = 0.50$	MiniLM-L12 $\tau = 0.40$	DeBERTa $\tau = 0.35$	DeBERTa $\tau = 0.40$	DeBERTa $\tau = 0.45$	DeBERTa $\tau = 0.50$
Accuracy	—	72.60	72.93	74.06	75.11	75.78
Precision	78.14	78.22	69.31	71.67	74.84	79.45
Overall recall	—	61.64	81.87	79.16	75.29	69.26
Overall F1	—	0.6895	0.7507	0.7523	0.7507	0.7400
Direct recall	—	63.11	81.97	79.30	75.67	69.87
Second-order recall	37.70	47.60	80.95	77.78	71.75	63.49
Second-order caught	118/313	149/313	255/315	245/315	226/315	200/315
Second-order F1	0.5086	0.5919	0.7468	0.7460	0.7326	0.7058
Noise specificity	84.08	83.28	64.08	69.01	74.93	82.25
True positives	—	2,026	2,714	2,624	2,496	2,296
True negatives	—	2,810	2,144	2,309	2,507	2,752
False positives	—	564	1,202	1,037	839	594
False negatives	—	1,261	601	691	819	1,019

Percentages unless stated; "—" means not reported. The MiniLM columns use holdout v1 and the DeBERTa columns use holdout v2. The shaded column is the production operating point.

At the same threshold of 0.40, the DeBERTa model finds 598 more relevant headlines than the 12-layer MiniLM (2,624 against 2,026 true positives) and cuts missed headlines from 1,261 to 691. Overall recall rises by 17.52 points (61.64% to 79.16%) and second-order recall by 30.18 points (47.60% to 77.78%, or 245 of 315). The price is more false alerts: false positives rise from 564 to 1,037, precision falls from 78.22% to 71.67%, and noise specificity falls from 83.28% to 69.01%. For orientation, the 95% Wilson score intervals, computed by us from the reported counts, are 42.1–53.1% for the second-order recall of the 12-layer MiniLM (149/313) and 72.9–82.0% for that of DeBERTa at $\tau = 0.40$ (245/315). These intervals treat pairs as independent and say nothing about the confounds below: the comparisons are not controlled, because the two models were scored on different holdout versions and input formats (Section VI).

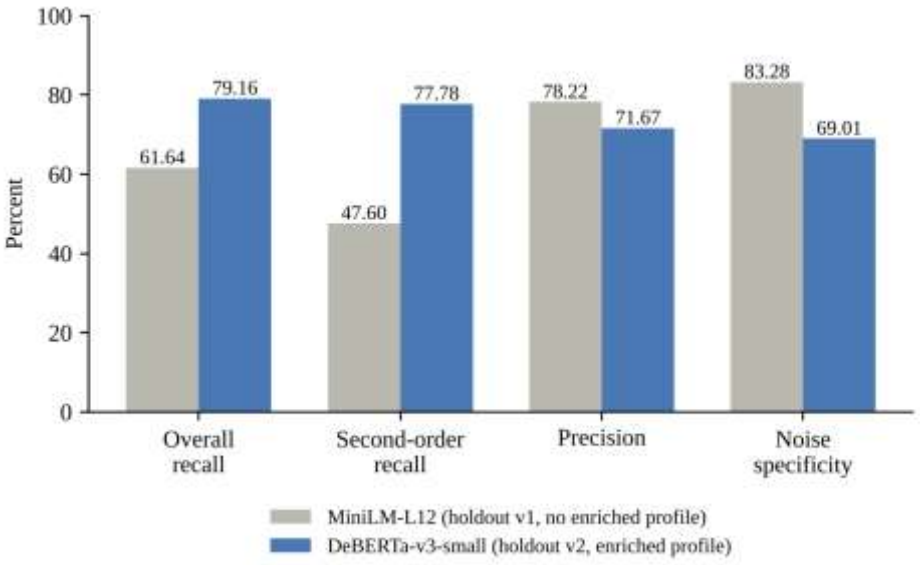


Fig. 2. MiniLM-L12 and DeBERTa columns of Table IV at $\tau = 0.40$. The two models were scored on different holdout versions, and only DeBERTa used enriched profiles.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

B. Production Operating Point

At $\tau = 0.40$ the 2,624 true positives split into 2,379 direct items (79.30% of 3,000) and 245 second-order shocks (77.78% of 315). The 691 missed positives are 621 direct items and 70 second-order shocks. Of the 3,346 negative pairs, 2,309 were discarded and 1,037 were kept. Accuracy is 74.06%, precision 71.67%, recall 79.16% and F1 0.7523. The evaluation run scored all 6,661 pairs in 433.49 seconds (65.08 ms per pair).

C. Effect of the Threshold

Raising τ from 0.35 to 0.50 trades recall for precision (Fig. 3). Recall falls from 81.87% to 69.26%, second-order recall from 80.95% to 63.49%, and precision rises from 69.31% to 79.45%. At $\tau = 0.50$ the model discards 1,019 relevant headlines, 115 of them second-order shocks; at $\tau = 0.35$ it passes 1,202 false positives. Overall F1 is almost flat across 0.35, 0.40 and 0.45 (0.7507, 0.7523, 0.7507), so the choice among them is mainly about the cost of errors, not about F1. We use 0.40, which has the highest F1 of the four, by a margin of 0.0016.

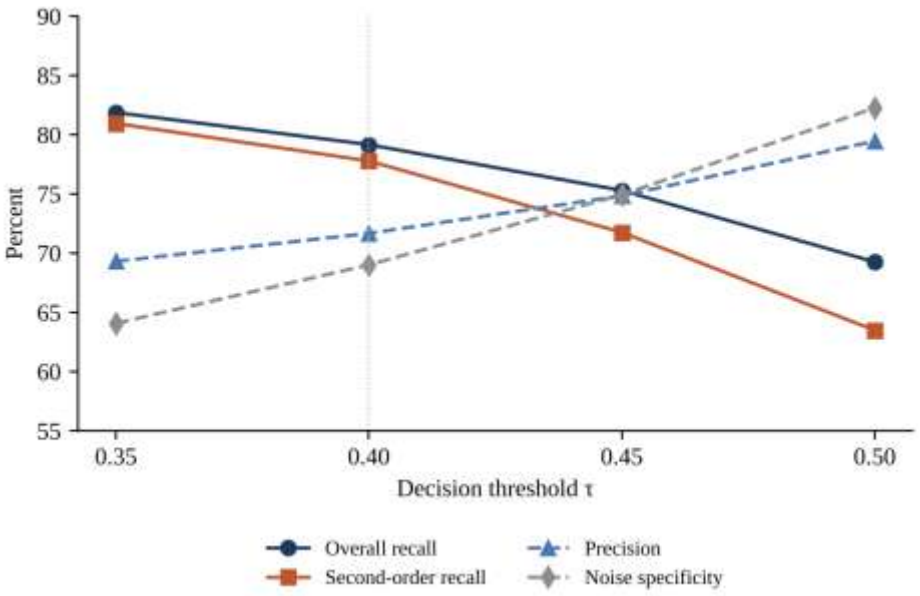


Fig. 3. Effect of the decision threshold on the DeBERTa model (holdout v2). The dotted line marks the production threshold.

D. A Pitfall: Applying the Sigmoid Twice

When we first evaluated the fine-tuned single-output model, we applied a logistic sigmoid, $\sigma(x) = 1 / (1 + e^{-x})$, to the output of CrossEncoder.predict(). The result was a complete collapse to the positive class at both $\tau = 0.35$ and $\tau = 0.50$: all 3,315 positives and all 3,346 negatives were kept, giving 100% recall and 0% specificity. The cause is that predict() already returns scores in [0, 1] for a single-label cross-encoder trained on 0/1 targets. A second sigmoid maps that range into [0.5, 0.7311], so every pair, however irrelevant, lands at or above 0.50:

$$\sigma(s) \geq 0.5 \text{ for all } s \geq 0; \;\; 0.5000 \leq \sigma(s) \leq 0.7311 \text{ for } 0 \leq s \leq 1 \tag{5}$$

Fig. 4 draws the mapping. A raw score of 0.0018, such as an Intel profile paired with an irrelevant headline, becomes about 0.50045 and passes both cutoffs. Because the sigmoid is monotonic, the ranking of pairs is unchanged; only the thresholding is broken. Removing the extra sigmoid and thresholding the raw score restored discrimination (2,144 true negatives at $\tau = 0.35$ and 2,309 at $\tau = 0.40$). Anyone using a single-label cross-encoder should check the output range before adding an activation function.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

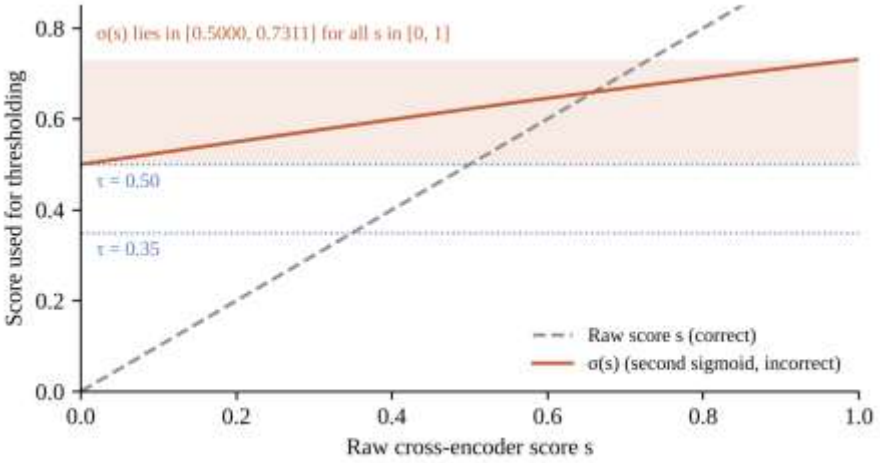


Fig. 4. A second sigmoid compresses scores in [0, 1] into [0.5000, 0.7311], above both thresholds tested.

E. Inference Cost

Benchmarks ran on a consumer laptop with a 13th-generation Intel Core i5 U-series CPU and 8 GB of RAM, under Windows with Python 3.10+ and CPU-only PyTorch 2.2+ (batch size 64 for the holdout runs, 32 for live RSS streaming). Across six holdout runs, the DeBERTa model took between 303.43 and 440.14 seconds for 6,661 pairs, depending on background load. That is 45.55 to 66.08 ms per pair, or 15.1 to 22.0 pairs per second, several times slower than the 12-layer MiniLM (Table V). Articles already seen are skipped by URL hash and cost no model time.

TABLE V. MODEL SIZE, MEMORY AND CPU INFERENCE COST ON THE HOLDOUT SET

Model	Parameters	Disk	Active RAM	Time per pair	Throughput
MiniLM-L6	22.7M	80–91 MB	~250 MB	5.95 ms	~168 pairs/s
MiniLM-L12	33.4M	~130 MB	~350 MB	12.09–12.38 ms	~81 pairs/s
DeBERTa-v3-small	44M backbone, 141M total	~170 MB compressed, ~567 MB uncompressed	~400 MB	45.55–66.08 ms	15.1–22.0 pairs/s

F. Example

The live set contains the headline “West Coast Aviation Jet Fuel Wholesale Crack Spreads Jump 9% After Northern California Refinery Outage”. Alaska Air Group (ALK) is not mentioned, but its profile lists aviation jet fuel among its upstream inputs, so the headline can be matched to it. The same headline paired with Kroger (KR) or Palo Alto Networks (PANW), whose profiles contain no fuel dependency, scores below 0.05. This is a single illustrative case, not a systematic test.

G. Supplementary Live Set

The supplementary live set was scored with the same fine-tuned model at $\tau = 0.35$, the threshold recorded in its evaluation log. The model kept 59 of 64 relevant headlines (92.19% recall), including 39 of 40 direct items (97.50%) and 20 of 24 second-order shocks (83.33%), and blocked 51 of 52 irrelevant headlines (98.08% specificity). Accuracy was 94.83%, precision 98.33% and F1 0.9516; second-order F1, computed with the overall precision as in Section IV-C, was 0.9021. The run took 7.01 seconds, or 60.42 ms per pair. Table VI compares these results with the holdout at the same threshold.

 This work is licensed under a Creative Commons Attribution 4.0 International License

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151164

TABLE VI. HOLDOUT V2 AND LIVE SET AT THE SAME THRESHOLD ($\tau = 0.35$)

Metric	Holdout v2	Live set
Pairs	6,661	116 (evaluation log)
Accuracy (%)	72.93	94.83
Precision (%)	69.31	98.33
Overall recall (%)	81.87	92.19
Overall F1	0.7507	0.9516
Direct recall (%)	81.97	97.50
Second-order recall (%)	80.95 (255/315)	83.33 (20/24)
Second-order F1	0.7468	0.9021
Noise specificity (%)	64.08	98.08
TP / TN / FP / FN	2,714 / 2,144 / 1,202 / 601	59 / 51 / 1 / 5

Precision, noise specificity and accuracy are much higher on the live set, while second-order recall is close to the holdout value (83.33% against 80.95%). We have not analysed the cause of the gap. The live set is small, has a much higher share of second-order pairs than the holdout (20.87% against 4.74%), and is scored at 0.35 rather than the production threshold of 0.40, so it is a supplementary check and not a replacement for the holdout evaluation. The 95% Wilson interval for its second-order recall, computed by us, is 64.1–93.3% and therefore includes the holdout value. The pair-count discrepancy (116 in the evaluation log against 115 in the dataset description, with 52 against 51 negatives) is unresolved; the reported metrics are those of the evaluation log.

VI. DISCUSSION AND LIMITATIONS

In portfolio monitoring the two error types are not equally costly. A false positive means a reader glances at a borderline headline. A missed second-order shock, such as a change in Indonesian nickel quotas or a bottleneck in Taiwanese wafer allocation, leaves the reader unaware of an upstream disruption. This asymmetry is why we operate at 0.40, even though 0.50 gives higher accuracy (75.78%) and precision (79.45%). The following limitations apply to the results above.

1. **Backbone and profile format changed together.** The MiniLM baselines were scored on the earlier holdout without the enriched profile, and the DeBERTa model on the final holdout with it. The gain therefore cannot be attributed to profile enrichment, NLI pretraining or disentangled attention separately. It is also not recorded whether the MiniLM baselines were fine-tuned on the same training data (Section IV-B). A controlled ablation on one backbone and one split, using the identity only (E), identity plus sector ($E \mid S$) and the full profile ($E \mid S \mid U$), is needed.
2. **Scores are not calibrated.** The output is an MSE regression score. Reading it as a probability would require temperature scaling or isotonic regression fitted on held-out data and checked with expected calibration error, Brier score and reliability diagrams [14].
3. **Threshold chosen on the test set.** The sweep in Table IV was run on the holdout set. A cleaner protocol fixes τ on the validation set and uses the holdout once. The F1 differences between 0.35 and 0.45 are small enough that this may not change the choice, but it should be shown.
4. **Hand-written profiles.** The 55 profiles were curated manually. If a listed dependency arose after the date of a historical headline, the profile leaks information (point-in-time leakage). Manual profiles also do not scale to thousands of companies or to tier-3 and tier-4 suppliers.
5. **Event-level leakage.** The split is chronological and exact duplicate URLs and headlines were removed, but wire services publish many follow-up stories on one event. Near-duplicates of the same event may fall on both sides of a split. Event clustering and a time gap between splits would rule this out.
6. **Short headlines and easy negatives.** Headlines are 8–20 words, so ambiguous ones (for instance on automotive microcontrollers or interbank liquidity) match broad profile terms and contribute to the 1,037 false positives at $\tau = 0.40$. Negatives come from non-overlapping industries, so the test is easier than a live feed in which the same commodity appears in unrelated contexts.
7. **Label construction and partitions.** Second-order positives are defined by pair construction (company absent from the headline, topic listed in the company's U field); no independent annotation or agreement study is reported. The validation set and holdout v2 are designated for checkpoint selection and final evaluation respectively, and they have identical class counts (Table III).

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151164

8. **Small live set.** The supplementary live set has 64 positive pairs, is scored at 0.35 rather than 0.40, has an unresolved negative count (115 or 116 pairs, Section V-G) and was run without confidence intervals. Its results are not directly comparable with the holdout beyond the same-threshold view in Table VI.

VII. CONCLUSION AND FUTURE WORK

FinNews targets news that matters to a company without naming it. It describes each company by identity, sector and upstream inputs, and scores headline–profile pairs with an NLI-initialized DeBERTa-v3-small cross-encoder. On a chronological holdout of 6,661 pairs from 55 companies, the model at $\tau = 0.40$ reached 74.06% accuracy, 71.67% precision, 79.16% recall and an F1 of 0.7523, with 79.30% recall on direct news and 77.78% on second-order shocks. A 12-layer MiniLM baseline retained 47.60% of second-order shocks, but the baseline was scored on a different holdout without enriched profiles, and the new model produces more false alerts. On a small supplementary live set scored at $\tau = 0.35$, the model retained 20 of 24 second-order shocks and reached 94.83% accuracy. Separating the effect of profile enrichment from that of the backbone and calibrating the scores are the main steps still needed. The full system runs on an ordinary laptop CPU, which suggests that continuous supply-chain monitoring is feasible without a knowledge-graph service.

Future work is planned in five directions:

1. **Factorial ablation.** A 3 × 4 design crossing three profile schemas (E, $E \mid S$, $E \mid S \mid U$) with four models (BM25/TF-IDF, the all-MiniLM-L6-v2 bi-encoder, MiniLM-L12 and DeBERTa-v3-small) on the same split, with 95% bootstrap confidence intervals.
2. **Calibration.** Calibration fitted on the validation set and reported before and after.
3. **Automated profiles.** Profile construction from point-in-time Item 1 (Business) and Item 1A (Risk Factors) text in 10-K filings, using relation extraction in the style of Wichmann et al. [1].
4. **Less ambiguous input.** Adding the first two sentences of each article and training on hard negatives, such as articles that mention the same commodity in an unrelated industry, should reduce headline ambiguity.
5. **Faster inference.** Exporting the model to ONNX Runtime with dynamic INT8 quantization is expected to bring the checkpoint under 150 MB and latency below 15 ms per pair; we have not yet measured this.

REFERENCES

[1] Wichmann, P., Brintrup, A., Baker, S., Woodall, P. and McFarlane, D. (2020). Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58(17), 5320–5336.
[2] Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796.
[3] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
[4] Jacobs, G., Lefever, E. and Hoste, V. (2018). Economic event detection in company-specific news text. *Proceedings of the 1st Workshop on Economics and Natural Language Processing (ECONLP), ACL 2018*, pp. 1–10.
[5] Pei, J. et al. (2024). Modeling and detecting company risks from news: A case study in Bloomberg news. *Proceedings of NAACL 2024, Industry Track*, pp. 60–68.
[6] Yıldırım et al. (2018). Classification of "hot news" for financial forecast using NLP techniques. *2018 IEEE International Conference on Big Data*, pp. 4719–4722.
[7] Del Corro, L. and Hoffart, J. (2021). From stock prediction to financial relevance: Repurposing attention weights to assess news relevance without manual annotations. *Proceedings of the 3rd Workshop on Economics and Natural Language Processing (ECONLP)*, pp. 45–49.
[8] Caliari, de Menezes and Vieira (2024). Supply chain news classification using deep learning. *Anais do Encontro Nacional de Engenharia de Produção (ENEGEP 2024)*.
[9] Vishnuthilak, Rolf, Reggelin and Lang (2024). Using sentiment analysis to detect disruptive events in supply chains. *IFAC-PapersOnLine*, 58(19), 864–869.
[10] Sato, K. and Mizuno, T. (2022). Propagation of shocks in individual firms through supplier–customer relationships. *The Review of Socionetwork Strategies*, 16, 1–19.
[11] Thakur, N., Reimers, N., Daxenberger, J. and Gurevych, I. (2021). Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *Proceedings of NAACL-HLT 2021*, pp. 296–310.
[12] He, P., Liu, X., Gao, J. and Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *International Conference on Learning Representations (ICLR)*.

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151164

[13] He, P., Gao, J. and Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *International Conference on Learning Representations (ICLR)*.

[14] Desai, S. and Durrett, G. (2020). Calibration of pre-trained transformers. *Proceedings of EMNLP 2020*, pp. 295–302.