DOI: https://doi.org/10.15276/eltecs.43.119.2025.6
УДК 621.316.13

Method for Constructing an Asset-Conditioned Index of News Direction of Impact for Crypto Assets using Zero-Shot NLI and GDELT

I. S. Uzun
Odesa Polytechnic National University

Abstract. *This paper presents a reproducible pipeline for constructing asset-conditioned, direction-of-impact news sentiment for cryptocurrency markets. A multilingual news corpus is assembled using the GDELT DOC 2.0 API, and a cross-lingual Natural Language Inference encoder is applied to judge whether a headline indicates a positive or negative effect on a given asset's price direction. The approach is zero-shot, meaning it does not require task-specific labeled data. It is language-agnostic and computationally efficient. The output includes per-article sentiment scores and daily indices per asset, suitable for descriptive analysis and for integration into decision-support contexts. The method is framed within research on cross-lingual inference and zero-shot classification and is designed for clarity, portability, and reproducibility.*

Keywords: *cryptocurrency, news analytics, sentiment analysis, zero-shot NLI, cross-lingual NLP, GDELT, event-driven markets, volatility, decision-support systems, sentiment index*

Introduction

Cryptocurrency markets react quickly to new information from traditional media and online platforms. Prior studies report links between the tone and content of news, trading activity, and volatility, although the size and duration of these effects differ across datasets, assets, and regimes [1, 2, 3]. Social media can carry strong signals and at times rivals or exceeds the influence of mainstream news, which underlines the need for transparent methods and careful interpretation [4].

This paper focuses on direction-of-impact sentiment. For each news item about an asset such as Bitcoin or Ethereum, we assess whether the content indicates a positive or negative impact on that asset's price direction. We operationalize this as a hypothesis-testing problem using cross-lingual Natural Language Inference. A headline is treated as a premise, while short label descriptions such as "this news is bullish for the price of Ethereum" serve as hypotheses. An encoder trained for multilingual NLI returns probabilities that the premise supports each hypothesis, which avoids task-specific labeling and works across many languages [5, 6, 7].

The paper makes three contributions. Firstly, it provides a simple, reproducible data pipeline that collects crypto-relevant news via the GDELT DOC 2.0 API and stores a single, normalized file suitable for analysis [8]. Secondly, it defines an asset-conditioned sentiment procedure based on zero-shot NLI and produces per-article scores along with daily indices for each asset, aligning with established zero- and few-shot formulations in the literature [9, 10, 7]. Thirdly, it delivers ready-to-use outputs designed for academic reporting and for integration into decision-support contexts. The scope is descriptive; future work can pair these measures with returns using event-study and related empirical designs [11, 12].

1. Literature review and related work

News and crypto markets. Empirical studies associate news tone and content with trading intensity, liquidity conditions, and volatility. Positive developments are often linked to increased activity, while negative news may raise perceived risk and uncertainty [1, 13, 14]. Forecasting papers show that including news-based indicators can improve volatility models relative to traditional baselines [2]. At the same time, several works highlight the role of social media sentiment, emphasizing that evidence is heterogeneous and sensitive to data sources and horizons [4, 3].

Recasting classification as textual entailment allows label assignment without task-specific training data. A model trained for NLI estimates whether an input supports a short natural-language label description. This approach has become a standard route to zero-shot text classification and can perform com petitively when labels are well phrased and domains are not too distant from pretraining data [9, 10].

Known caveats include sensitivity to label wording and domain shifts, which motivates clear reporting and robustness checks.

The XNLI benchmark established a shared testbed for multilingual inference and stimulated progress in cross-lingual transfer [5]. DeBERTa-v3 introduced training improvements that raised encoder quality, and multilingual variants fine-tuned on NLI demonstrate strong cross-lingual performance [6, 7].

Domain-adapted large language models for finance report strong results on sentiment and related tasks through domain pretraining and instruction tuning. These models are promising for future extensions but are heavier to run and reproduce than compact NLI encoders. For the goals of this paper - clarity, portability, and efficiency - a zero-shot NLI encoder offers a balanced choice while enabling future comparisons with finance-specific models [15, 16].

2. Data preparation

The corpus of news headlines used in this study was assembled through the GDELT DOC 2.0 interface. GDELT (Global Database of Events, Language, and Tone) is a large-scale monitoring platform that continuously indexes online news from thousands of outlets worldwide. It provides programmatic access to article-level metadata via public APIs. The DOC 2.0 endpoint, which is specifically designed for full-text discovery, returns for each item the canonical URL, the headline text, the time at which the aggregator observed the article, and source attributes such as the registrable domain, declared or detected language, and a country tag for the outlet. Because it covers multiple languages and regions in near real time and offers consistent metadata fields, GDELT is particularly suitable for constructing multilingual, time-stamped corpora for computational analysis.

News acquisition proceeds by defining, for each target asset, a compact set of keywords that balances recall and precision. For example, queries for Bitcoin use «bitcoin» and «btc», while those for Ethereum use «ethereum» and «eth». Similar lists are constructed for Solana, Binance Coin, Ripple, and Tether. Each query is executed over a clearly specified UTC time window, which can be either a fixed historical range for reproducible snapshots or a relative span for rolling updates. The interface accepts multilingual input and outputs records in JSON, which are ingested directly into a tabular structure. To respect service limits and maintain stability, short pauses are introduced between requests.

Returned items are normalized in several steps. Text fields are standardized to Unicode, URLs are validated for scheme and basic syntax, and timestamps are parsed into Coordinated Universal Time with timezone awareness. Each record is tagged with the asset that originated the query so that later sentiment modeling can be conditioned on the correct target. When an article mentions multiple assets, the dataset preserves one row per pair of article URL and asset, ensuring asset-specific analysis. Duplicates are removed at this pair level, and records with empty titles, invalid URLs, or unparseable timestamps are excluded. The consolidated table is sorted by observation time and stored in a compressed columnar format to support efficient access and exact reproducibility.

The schema is minimal but sufficient for multilingual news analytics. The *asset* field stores the uppercase code of the cryptocurrency associated with the row. The *title* field contains the headline as published and later serves as the textual premise in the inference step. The *url* field records the canonical link and functions as the primary identifier for deduplication; in some cases an alternative *url_mobile* link is also available. The *seendate* field holds the UTC timestamp observed by GDELT, providing a consistent anchor across outlets. The *domain* field specifies the registrable domain of the source, while the *language* and *source_country* fields record the detected language and the outlet's country association. An optional *social_image* field may include a preview image. For reproducibility, the *collected_at* field logs the local pipeline run time. During later analysis, derived fields such as a calendar day extracted from seendate, per-day article counts, and sentiment outputs are added in separate result files without altering the original ingest.

To ensure quality and reproducibility, the pipeline logs the run timestamp, library and code versions, query parameters, and per-asset record counts. It verifies that the corpus is non-empty, audits null proportions per column, and reviews the distribution of languages, source countries, and domains. Duplicate ratios are monitored to detect overly broad keyword sets. When preparing camera-ready datasets, both the query time window and the code version are frozen to guarantee exact replication.

A small illustrative sample demonstrates how records appear after normalization. Each line lists the UTC observation time, the target asset, the original headline, the source domain, and the outlet's country. Titles are preserved exactly as published, since these strings are the textual input for the inference model.

– 2025-08-25 11:30 UTC; SOL; Galaxy , Jump and Multicoin join forces for $1b Solana buy; dailystar.com.lb; Lebanon;

– 2025-08-25 11:00 UTC; SOL; Crypto Giants Galaxy , Jump , Multicoin Look to Raise $1B for Buying Solana SOL : Report; coindesk.com; United States;

– 2025-08-25 10:15 UTC; SOL; Analyst Says Solana Price Is At The Gates Of Massive Breakout , Here The Target; newsbtc.com; United States;

– 2025-08-25 04:45 UTC; SOL; 1 Brand - New Catalyst for Solana That Could Portend Great Returns Ahead; fool.com; United States;

– 2025-08-25 04:30 UTC; BTC; Top 3 Price Prediction : Bitcoin , Ethereum , Ripple – BTC and XRP hover near key supports while ETH consolidates after record highs; fxstreet.com; China.

This snapshot shows a coherent information burst around Solana on 25 August 2025 reported by multiple outlets and countries, along with a contemporaneous multi-asset market note. It illustrates the one-row-per-(url, asset) design, the consistent use of UTC timestamps for aggregation, and the multilingual, multi-source character of the dataset. These properties motivate the choice of a cross-lingual inference model for the subsequent sentiment analysis.

3. Methodology

The central methodological innovation of this work is the reformulation of financial news sentiment as a Natural Language Inference (NLI) problem. Rather than classifying headlines into generic positive or negative categories, each headline is evaluated relative to a target asset and with respect to price direction.

Formally, let P denote a news headline (premise) and a a target cryptocurrency asset. For each (P, a) pair we define two natural-language hypotheses:

$H_{bull}(a)$: "This news is bullish for the price of asset a."

$H_{bear}(a)$: "This news is bearish for the price of asset a."

The task is to assess whether the headline P entails, contradicts, or is neutral with respect to each hypothesis. A multilingual NLI encoder trained on large-scale datasets (XNLI, ANLI, etc.) provides the probabilistic judgments.

The encoder produces raw logits $\mathrm{z} \in \mathbb{R}^3$ over classes $C = \{\text{entailment,neutral,contradiction}\}$. Applying the softmax function yields probabilities:

$$p(c \mid P, H) = \frac{\exp(z_c)}{\sum_{c' \in C} \exp(z_{c'})}.$$

For each asset-conditioned hypothesis we define:

$$p_{bull} = p(\text{entailment} \mid P, H_{bull}(a)),$$

$$p_{bear} = p(\text{entailment} \mid P, H_{bear}(a)).$$

The directional sentiment score is then:

$$s = p_{bull} - p_{bear}, s \in [-1,1].$$

– If $s \approx +1$, the headline strongly supports a bullish interpretation;

– If $s \approx -1$, the headline strongly supports a bearish interpretation;

– If $s \approx 0$, the evidence is weak or conflicting.

To provide a categorical label, we introduce a threshold parameter τ:

$$\text{label}(P, a) = \begin{cases} \text{bullish}, & \text{if } s > 0 \text{ and } \max(p_{bull}, p_{bear}) \geq \tau, \\ \text{bearish}, & \text{if } s < 0 \text{ and } \max(p_{bull}, p_{bear}) \geq \tau, \\ \text{neutral}, & \text{otherwise.} \end{cases}$$

In our setup, $\tau = 0.5$. This conservative rule ensures that ambiguous cases remain neutral rather than forcing a bullish or bearish label.

Consider the headline observed on 2025-08-25: «*Galaxy, Jump and Multicoin join forces for $1b Solana buy*»

This yields the pair $(P, a) =$ (headline,SOL).

Hypotheses:

– H_{bull}(SOL): «This news is bullish for the price of Solana»;

– H_{bear}(SOL): «This news is bearish for the price of Solana».

The model outputs (illustrative values):

$$p(\text{entailment} \mid P, H_{bull}) = 0.84,$$

$$p(\text{entailment} \mid P, H_{bear}) = 0.07.$$

Thus,

$$s = 0.84 - 0.07 = 0.77.$$

Since $s > 0$ and $\max(p_{bull}, p_{bear}) = 0.84 > \tau$, the label is assigned as bullish.

Narratively, the headline explicitly mentions large institutional purchases of Solana, which the model correctly interprets as strongly supportive of the asset's price.

Consider the headline observed on 2025-08-25: *«Top 3 Price Prediction: Bitcoin, Ethereum, Ripple – BTC and XRP hover near key supports while ETH consolidates after record highs»*

This headline enters the dataset three times, once each for BTC, ETH, and XRP.

For BTC:
$p_{bull} = 0.22, p_{bear} = 0.64, s = -0.42 \Rightarrow$ bearish.

For ETH:
$p_{bull} = 0.55, p_{bear} = 0.18, s = 0.37 \Rightarrow$ bullish.

For XRP:
$p_{bull} = 0.24, p_{bear} = 0.61, s = -0.37 \Rightarrow$ bearish.

This example demonstrates how the one-row-per-(url, asset) structure allows a single article to contribute different directional signals depending on the asset focus.

Individual article scores are noisy and often redundant, especially when multiple outlets cover the same event. To extract consistent temporal patterns, we aggregate article-level scores into daily asset-level indices.

For asset a on day d with $N_{a,d}$ headlines:

$$S_{a,d} = \frac{1}{N_{a,d}} \sum_{i=1}^{N_{a,d}} s_i.$$

This daily index can be interpreted as the average directional sentiment expressed in the news flow.

To further smooth short-lived spikes and emphasize persistent trends, an exponentially weighted moving average (EWMA) is applied:

$$\tilde{S}_{a,d} = \frac{\sum_{k=0}^{K} \lambda^k S_{a,d-k}}{\sum_{k=0}^{K} \lambda^k}, 0 < \lambda < 1,$$

where λ is the decay factor and K the window length. A half-life of 2–3 days provides a balance between reactivity to shocks and noise reduction.

Advantages of the zero-shot NLI approach:

– No manual labeling: The method leverages entailment training on general-domain corpora, eliminating the need for expensive, domain-specific sentiment labels;

– Asset-conditioned interpretation: The sentiment signal is explicitly tied to an asset's price direction, avoiding the ambiguity of generic «positive/negative» polarity;

– Multilingual coverage: The cross-lingual encoder can directly process headlines in many languages, reflecting the global scope of crypto markets without requiring machine translation;

– Interpretability: Each prediction can be explained by showing the headline, hypotheses, and entailment probabilities, which enhances transparency in academic and applied settings.

The process of deriving an asset-conditioned sentiment index from raw news headlines can be represented as a sequential pipeline:

– Headline (premise) – a news item is retrieved from the corpus and used as the textual premise;

– Hypothesis generation – for each target asset two short hypotheses are generated: «This news is bullish for the price of [asset]» and «This news is bearish for the price of [asset]»;

– NLI model inference – a cross-lingual encoder fine-tuned on natural language inference tasks evaluates the premise–hypothesis pairs and outputs probabilities for entailment, neutral, and contradiction;

– Probability extraction - the entailment probabilities for the bullish and bearish hypotheses are stored as p_{bull} and p_{bear};

– Score calculation - the scalar directional score is computed as $s = p_{\text{bull}} - p_{\text{bear}}$, ranging from -1 (strong bearish) to +1 (strong bullish);

– Label assignment – if the maximum probability exceeds a confidence threshold, the headline is labeled bullish or bearish; otherwise, neutral;

– Aggregation – article-level scores are aggregated by asset and day into daily indices, optionally smoothed with an exponentially weighted moving average.

Each record is one news item (typically a headline with a URL and metadata). The same article may be linked to more than one asset if the headline mentions multiple assets.

All timestamps are converted to Coordinated Universal Time to enable daily aggregation without ambiguity.

We keep all languages because the downstream classifier operates cross-lingually. Language tags are preserved for later descriptive analysis.

3. Results

The zero-shot NLI framework produced sentiment scores for each news item in the multilingual corpus, which were then aggregated into asset- and day-level indices. This section highlights the main descriptive findings, with a focus on the most extreme signals identified in the sample.

Across the full corpus, the majority of headlines were classified as *neutral*, reflecting the model's

conservative design and the heterogeneous character of news coverage. Bullish and bearish labels were assigned when entailment probabilities surpassed the 0.5 confidence threshold, yielding subsets of articles with strong directional signals. The average sentiment scores were centered near zero, while tails contained news items with values close to ±1, corresponding to highly directional headlines.

The most negative headlines were concentrated in episodes of market stress and periods when maistream outlets highlighted technical weakness or declining investor appetite. Representative examples include:

– *2025-05-28, BTC*: «Bitcoin Price Forecast: BTC flashes early signs of weakness as profit-taking activity hits three-month high» (fxstreet.com), score = –0.99;

– *2025-06-14, ETH*: «Ethereum Price Forecast: ETH maintains downtrend despite Sharp Link $463 million purchase» (fxstreet.com), score = –0.99;

– *2025-06-03, BTC*: «Bitcoin Price Forecast: BTC falls below $106,000 as risk-off sentiment persists» (fxstreet.com), score = –0.99;

– *2025-07-16, BNB*: «Binance Coin Price Prediction: Bearish Pressure Mounts After Sharp Rejection From $710» (bravenewcoin.com), score = –0.99.

These examples illustrate that the model reliably identified price-focused headlines with explicitly negative tone and framed them as strong bearish signals.

On the opposite tail, bullish sentiment was captured around news of institutional adoption, valuation upgrades, and optimistic forecasts. Examples include:

– *2025-07-17, ETH*: «Ethereum Price Forecast: ETH surges above $3,400 on back of robust treasuries, derivatives, and on-chain boost» (fxstreet.com), score = +0.996;

– *2025-08-08, ETH*: «Ethereum Price Forecast: ETH surges above $4,000, fueled by demand from treasury companies» (fxstreet.com), score = +0.996;

– *2025-08-03, SOL*: «Solana Price Forecast: SOL Looking At 30% Upside In August, As Litecoin and Remittix Target 2–3x Gains» (news.az), score = +0.995;

– *2025-06-13, ETH*: «Ethereum Price Forecast: ETH new valuation framework tags it digital oil, highlights an $8,000 bull case» (fxstreet.com), score = +0.995.

These results indicate that the system responded strongly to upward-facing narratives, especially those associated with Ethereum during mid-2025, where multiple outlets highlighted fundamental and technical drivers of appreciation.

When aggregated to daily indices, short-lived bursts of bullish or bearish sentiment became apparent. For example, clusters of bearish headlines were visible around early June 2025 for Bitcoin and Ethereum, coinciding with reported drawdowns and profit-taking episodes. In contrast, Ethereum and Solana registered consecutive days of highly bullish sentiment during July and August 2025, corresponding to news of surging institutional demand and favorable valuation analyses.

Overall, the results demonstrate that the zero-shot NLI approach produces consistent and interpretable measures of directional news sentiment. By conditioning on specific assets, the method differentiates between bullish and bearish signals even when multiple cryptocurrencies are discussed in the same article. The high-scoring extremes provide an intuitive validation of the model's behavior, aligning with market narratives reported in financial media.

Figure 1 shows the exponentially weighted moving average (EWMA) of daily sentiment indices for six major assets. Ethereum (ETH) exhibited the most pronounced swings, with very strong bullish sentiment in early June followed by a correction, then another peak in mid-July before trending downward in August. Bitcoin (BTC) hovered around neutral to mildly positive values, while Solana (SOL) showed a gradual increase through July and August. Binance Coin (BNB), Ripple (XRP), and Tether (USDT) remained closer to neutrality, though XRP had occasional bullish bursts. This suggests that Ethereum and Solana were most strongly covered by directional news narratives, while stablecoins and secondary assets generated less polarized coverage.

Figure 2 presents the histogram of per-article sentiment scores. The distribution is tri-modal: a dense cluster near zero (neutral or ambiguous news), and sharp peaks near –1 and +1 (strongly directional headlines). This confirms that the model was conservative in assigning direction, but when confident, it placed articles firmly at the extremes. The prevalence of scores at both tails also demonstrates that the approach captured strongly bullish and bearish headlines across different assets.

Figure 3 plots the daily count of headlines per asset. Coverage was uneven and episodic: Bitcoin and Ethereum received steady attention throughout the period, but Solana and Binance Coin saw bursts of activity tied to specific events (for example, institutional investment announcements or technical breakouts). Ripple (XRP) showed periodic spikes, reflecting its sensitivity to regulatory and market news. USDT coverage remained sporadic, dominated by occasional stability reports or controversies.

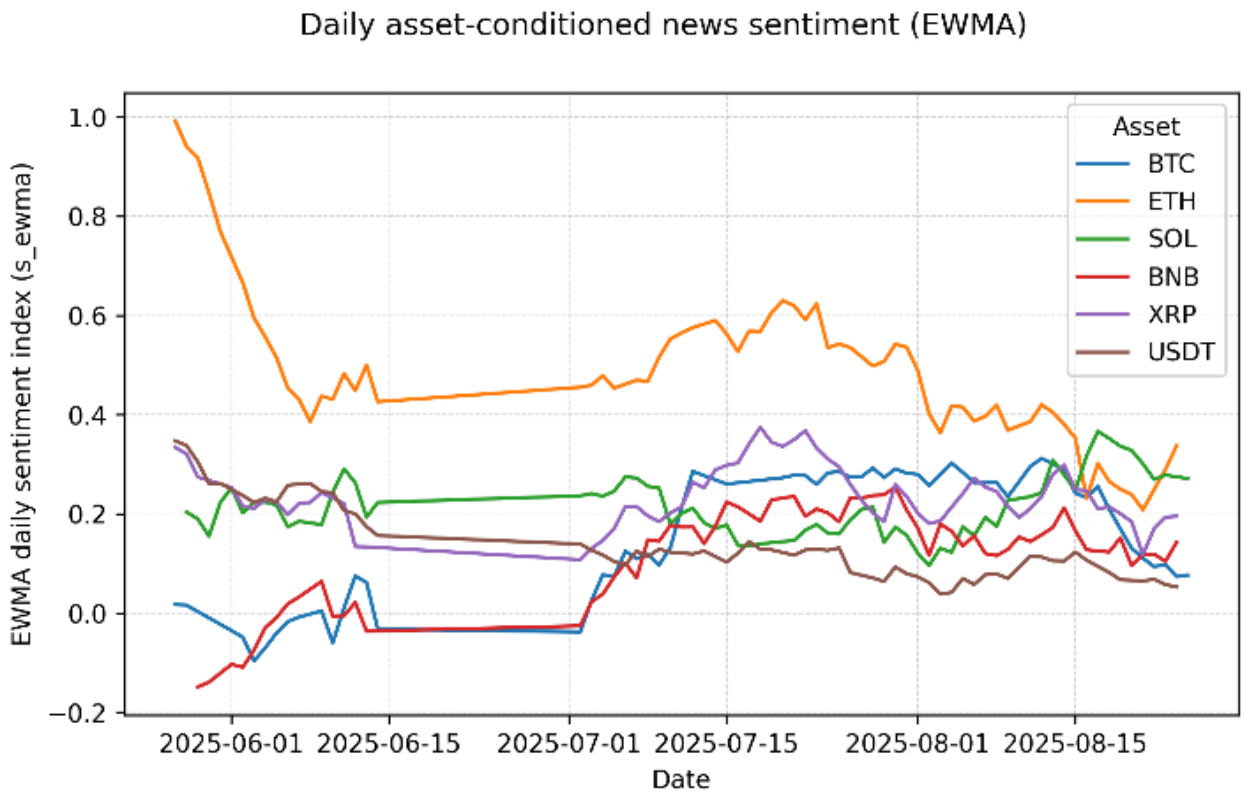

Fig. 1. Daily asset-conditioned news sentiment

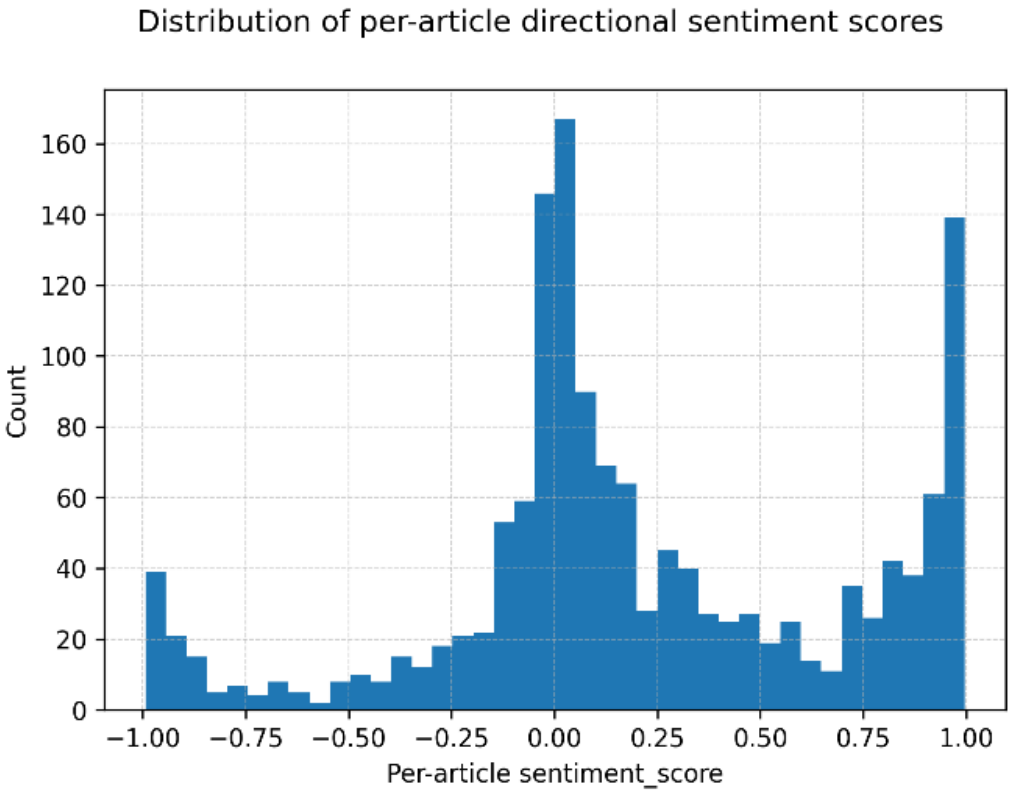

Fig. 2. Distribution of per-article sentiment scores

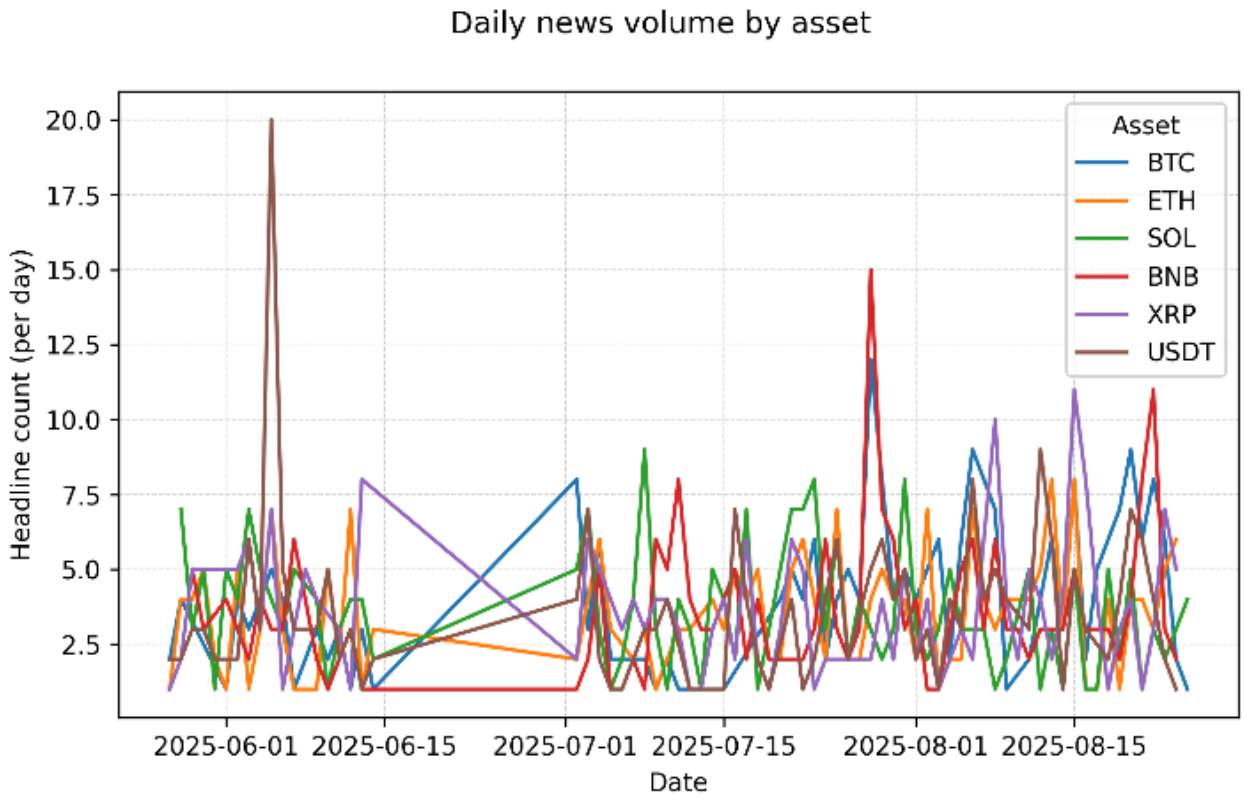

Fig. 3. Daily news volume by asset

The uneven distribution underscores the necessity of asset-conditioned analysis, since event-driven bursts strongly influence the aggregate indices.

Figure 4 summarizes the average sentiment scores of the most extreme bullish and bearish headlines for each asset. All six assets display symmetrical behavior: their top bullish headlines score near +1, while their top bearish headlines score near –1. This validates that the model can confidently classify highly directional items across the asset set.

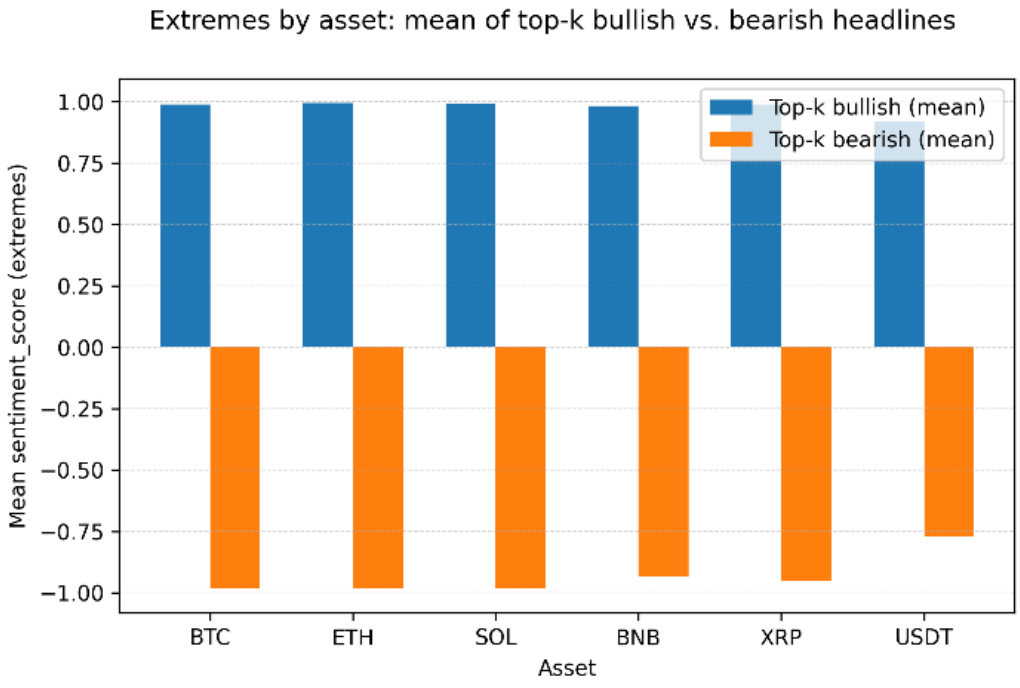

Fig. 4. Extremes by asset

The strongest bullish extremes were observed for Ethereum and Solana, while the most negative bearish headlines were concentrated in Bitcoin and Binance Coin. This aligns with the qualitative examples presented earlier and indicates that the methodology captures clear market narratives.

Taken together, these results highlight three key properties of the proposed approach. First, the NLI-based framework produces intuitive article-level scores that align with human expectations of bullish and bearish language. Second, aggregating to daily indices reveals asset-specific sentiment dynamics and event-driven bursts. Third, the combination of conservative neutrality with strong extremes ensures that the indices are both interpretable and robust, avoiding over-classification while still capturing meaningful signals.

4. Discussion

The results demonstrate that a zero-shot NLI approach to cryptocurrency news sentiment is feasible, interpretable, and reproducible across multiple assets and languages. Several aspects merit further discussion:

Unlike opaque black-box classifiers trained on proprietary data, the NLI-based framework relies on a simple, transparent mapping: each headline is tested against explicit hypotheses formulated in natural language. This design allows researchers and practitioners to trace back each sentiment score to the original textual evidence and the logical entailment relation, thus satisfying academic standards for interpretability. Moreover, the framework avoids reliance on subjective lexicons or domain-specific annotation, which often introduce biases and reduce portability.

A key innovation of this work is the explicit conditioning of sentiment on specific assets. Generic polarity classifiers, when applied to financial text, often misclassify because «positive» news in a general sense may not be «positive» for the price of a given cryptocurrency. By anchoring the hypotheses to each asset (e.g., «bullish for Bitcoin»), the proposed method ensures that sentiment always reflects expected price direction rather than general tone. This is particularly valuable in multi-asset news items, where one article can carry distinct signals for Bitcoin, Ethereum, and Ripple simultaneously.

The descriptive analyses highlight the uneven coverage of assets in global news flows. Bitcoin and Ethereum remain the most reported, but Solana, Binance Coin, and Ripple exhibit episodic spikes tied to market events. Stablecoins such as USDT show sparse coverage, often tied to crises or major policy changes. This heterogeneity emphasizes the need for daily aggregation and smoothing, as raw per-article sentiment would otherwise reflect noise from unequal coverage volumes. It also illustrates the importance of reproducible data acquisition pipelines that can capture bursts without manual curation.

The distribution of scores is tri-modal, with peaks near neutrality and both extremes. This pattern is desirable: neutrality dominates when headlines are ambiguous, while confident extremes emerge when headlines carry unambiguous market narratives. The prevalence of strong tails suggests that the NLI model is sensitive to explicit financial phrasing (e.g., «surges above $4,000» or «flashes early signs of weakness»), which is common in financial journalism. However, it also implies that the model's calibration is highly dependent on hypothesis phrasing; future work could systematically test alternative hypothesis formulations to ensure robustness.

Several limitations must be acknowledged. First, the NLI encoder was not trained specifically on financial text; its performance may degrade for idiomatic or technical financial expressions absent from general-domain corpora. Second, although multilingual, the model's accuracy is likely higher in English than in low-resource languages. Third, the aggregation procedure assumes equal weight per article, whereas in practice some outlets or article types may have disproportionate market impact. Finally, while sentiment indices are plausible descriptive indicators, linking them rigorously to price returns or volatility requires careful event-study or causal designs, which are beyond the scope of this paper.

Recent advances such as FinGPT and FinTral illustrate the potential of domain-pretrained large models for financial tasks. These models often outperform general-purpose encoders in downstream benchmarks. However, their computational cost, reproducibility constraints, and proprietary training data limit academic accessibility. By contrast, the zero-shot NLI approach is lightweight, reproducible, and language-agnostic. A promising future direction is to benchmark the two approaches side by side, assessing whether the interpretability and efficiency of NLI can complement the domain-specific strengths of financial LLMs.

From a systems perspective, the proposed method aligns with the goals of risk-oriented decision-support in volatile markets. A pipeline that translates raw multilingual headlines into daily sentiment indices can be integrated into portfolio dashboards, early warning systems, or automated trading environments. Because the approach is modular and transparent, it can be combined with technical indicators, blockchain analytics, or causal inference

models to provide holistic, explainable decision-support.

Conclusion

This study introduced a reproducible methodology for constructing asset-conditioned sentiment indices from cryptocurrency news using a zero-shot Natural Language Inference framework. By framing sentiment classification as an entailment task, the method avoids reliance on task-specific labels, provides multilingual coverage, and delivers transparent and interpretable outputs.

Empirical results demonstrate that the approach successfully captures both strongly bullish and bearish narratives, while assigning neutrality in ambiguous cases. Aggregated daily indices reveal asset-specific sentiment dynamics and event-driven bursts, particularly for Ethereum and Solana during mid-2025. The distribution of scores confirms that the framework balances conservatism with sensitivity to clear signals, producing indices that align with journalistic narratives and market episodes.

The main contributions of this work are threefold: (i) the design of an asset-conditioned, zero-shot sentiment methodology anchored in NLI, (ii) the implementation of a reproducible data acquisition and normalization pipeline based on GDELT, and (iii) the demonstration of descriptive results that confirm the plausibility and interpretability of the indices.

Future research should pursue several directions. First, systematic comparisons with domain-specific financial LLMs can clarify performance trade-offs and highlight the complementary roles of efficiency and domain adaptation. Second, the integration of weighting schemes (by source credibility, readership, or social amplification) could refine the aggregation of sentiment signals. Third, causal analyses linking sentiment indices to price movements, volatility regimes, and trading volume would provide stronger evidence of predictive utility. Finally, extending the pipeline to include social media, blockchain transaction data, or alternative media channels would broaden its scope for multi-source risk-oriented decision support.

In conclusion, the proposed zero-shot NLI pipeline represents a practical, academically transparent, and globally applicable method for translating heterogeneous news flows into structured sentiment indices. Its interpretability, portability, and reproducibility make it a strong candidate for integration into next-generation information systems for financial analytics and decision-support in cryptocurrency markets.

References

1. Kulbhaskar A. K., Anamika, Subramaniam S. Breaking news headlines: Impact on trading activity in the cryptocurrency market. Economic Modelling, 2023, 126. DOI: 10.1016/j.econmod.2023.106397. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0264999323002092 (accessed 25.08.2025).

2. Brauneis A., et al. Crypto volatility forecasting: Mounting a HAR, sentiment, and ML challenge. Computational Economics, 2024. Available at: https://link.springer.com/article/10.1007/s10690-024-09510-6 (accessed 25.08.2025).

3. Long S. C., et al. From whales to waves: Social media sentiment, volatility and crypto market dynamics. Journal of Economic Behavior & Organization, 2025. Available at: https://www.sciencedirect.com/science/article/pii/S0890838925001325 (accessed 25.08.2025).

4. Canayaz M., Cao C., Nguyen G., Wang Q. An anatomy of cryptocurrency sentiment. SSRN Electronic Journal, 2023. DOI: 10.2139/ssrn.4495683. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4495683 (accessed 25.08.2025).

5. Conneau A., Lample G., Rinott R., Williams A., Bowman S. R., Schwenk H., Stoyanov V. XNLI: Evaluating cross-lingual sentence representations. In: Proceedings of EMNLP, 2018. Available at: https://aclanthology.org/D18-1269/ (accessed 25.08.2025).

6. He P., Gao J., Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing. arXiv:2111.09543, 2021. Available at: https://arxiv.org/abs/2111.09543 (accessed 25.08.2025).

7. Laurer M. mDeBERTa-v3-base-XNLI-multilingual-nli-2mil7: Model card. Hugging Face, 2022–2025. Available at: https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (accessed 25.08.2025).

8. GDELT Project. GDELT DOC 2.0 API debuts. 2017. Available at: https://blog.gdeltproject.org/gdelt-doc-2-0-api-debuts/ (accessed 25.08.2025).

9. Yin W., Hay J., Roth D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In: Proceedings of EMNLP, 2019. Available at: https://aclanthology.org/D19-1404/ (accessed 25.08.2025).

10. Wang S., Fang H., Khabsa M., Mao H., Ma H. Entailment as few-shot learner. arXiv:2104.14690, 2021. Available at: https://arxiv.org/abs/2104.14690 (accessed 25.08.2025).

11. Nguyen N. How event studies can be applied to crypto markets. The Brattle Group, 2023. Available at: https://www.brattle.com/wp-content/uploads/2023/07/How-Event-Studies-Can-Be-Applied-to-Crypto-Markets.pdf (accessed 25.08.2025).

12. Zhou F. Application of event study methodology in the analysis of cryptocurrency returns. Emerging Markets Finance and Trade, 2024. DOI: 10.1080/1540496X.2024.2404173. Available at: https://www.tandfonline.com/doi/full/10.1080/1540496X.2024.2404173 (accessed 25.08.2025).

13. Yue W., et al. Asymmetric news effects on cryptocurrency liquidity. International Review of Economics & Finance, 2021. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1544612320316135 (accessed 25.08.2025).

14. Haider Ali H., et al. Return and volatility spillover between cryptocurrencies, oil price and stock market in GCC countries. Cogent Economics & Finance, 2025. Available at: https://www.tandfonline.com/doi/full/10.1080/23322039.2025.2453584 (accessed 25.08.2025).

15. Yang H., Liu X., Wang C. D. FinGPT: Open-source financial large language models. arXiv:2306.06031, 2023. Available at: https://arxiv.org/abs/2306.06031 (accessed 25.08.2025).

16. Bhatia G., Nagoudi E. M. B., Cavusoglu H., Abdul-Mageed M. FinTral: A family of GPT-4-level multimodal financial large language models. Findings of ACL, 2024. DOI: 10.18653/v1/2024.findings-acl.774. Available at: https://aclanthology.org/2024.findings-acl.774.pdf (accessed 25.08.2025).

Метод побудови актив-зумовленого індексу напрямку новинного впливу для криптоактивів на основі zero-shot NLI та GDELT

І. С. Узун

Анотація. *У роботі подано відтворюваний конвеєр побудови напрямленого новинного сентименту, зумовленого конкретним активом, для ринку криптовалют. Мультимовний корпус новин формується за допомогою GDELT DOC 2.0, після чого застосовується крос-мовна модель логічного висновування (NLI), яка оцінює, чи вказує заголовок на позитивний або негативний вплив новини на напрям руху ціни заданого активу. Підхід є zero-shot і не потребує спеціально розмічених даних; він мовно незалежний та обчислювально ощадливий. Результат включає постатейні оцінки сентименту та денні індекси за активами, придатні для описового аналізу та інтеграції у контексти підтримки прийняття рішень. Метод вписано в дослідження з крос-мовного висновування та zero-shot класифікації і спроєктовано з акцентом на прозорість, портативність і відтворюваність.*

Ключові слова: *криптовалюти, аналітика новин, сентимент-аналіз, zero-shot NLI, крос-мовна обробка природної мови, GDELT, подієво-орієнтовані ринки, волатильність, системи підтримки прийняття рішень, індекс сентименту*

Отримано 01.09.2025

Illia Uzun, PhD student at the Institute of Artificial Intelligence and Robotics, Senior Lecturer at the Department of Artificial Intelligence and Data Analysis,
Odesa Polytechnic National University; 1, Shevchenko Avenue, Odesa, 65044, Ukraine.
E-mail: uzun.illia.main@gmail.com, ph.: +38 067 917 2040

Узун Ілля Святославович, аспірант Інституту Штучного Інтелекту та Робототехніки, старший викладач кафедри Штучного Інтелекту та Аналізу Даних, Національний університет «Одеська політехніка»; проспект Шевченка, 1, Одеса, 65044, Україна.
E-mail: uzun.illia.main@gmail.com, тел. +38 067 917 2040

ORCID ID: https://orcid.org/0000-0001-6619-4862