KOLIntel: Making Crypto Influence Transparent, One Tweet at a Time

Anonymous ACL submission

Abstract

KOLIntel is a modular pipeline for end-to-end assessment of crypto-focused accounts on Twitter (X) with a single, auditable methodology. The pipeline comprises four stages: (1) *KOL discovery and identification* via social-graph expansion from curated seeds and hop-wise quality filters; (2) *verifiable-prediction* extraction and truth checking, where only claims containing sufficient structure (asset, target type/value or direction/range/event, and a time window) are retained and evaluated against unified ground-truth checkpoints (Hit/Miss/Partial/data-void); (3) *bot/astroturf risk* screening from multiple signals (template/clone text, timing regularity, homogeneous interactions, incentive links, structural imbalance), used as downweighting rather than definitive judgments; and (4) *influence and content quality* scoring, including bot-adjusted engagement, secondary spread, and six-dimensional content assessment (completeness, clarity, attractiveness, accuracy, marketing effectiveness, originality). Category scores are aggregated into a final score and grade (A/B/C/D) with transparent thresholds and hard constraints (e.g., minimum sample size, calibration/stability gates, high bot-risk caps). The system outputs an interactive HTML report—ranked overview, account pages, and structured evidence snippets—so conclusions are explainable and independently verifiable. As an optional downstream output, the report can also include a budget-allocation view that proportionally assigns campaign funds based on evidence-backed grades and caps.

1 Introduction

Influencer marketing plays a prominent role in cryptocurrency and Web3 information flows, yet assessing the reliability of Key Opinion Leaders (KOLs) remains challenging. Surface metrics such as follower counts or raw engagement can be gamed and often fail to reflect semantic value or accountability (Shukla et al., 2023). Feature-based bot detectors are susceptible to evasion and domain shift (Orabi et al., 2020). Meanwhile, crypto discourse frequently interacts with market sentiment and event timing, which motivates domain-specific evaluation grounded in auditable evidence (Huang et al., 2021).

This work targets four practical questions: *(i)* is an account a KOL within a specific topic; *(ii)* are its public "predictions" verifiable against objective data and stable over time; *(iii)* is there evidence of bot/astroturf amplification; and *(iv)* after removing suspected inorganic effects, what are the account's influence and content quality. Answering these requires a consistent methodology with verifiable evidence rather than opaque scores.

We present **KOLIntel**, a modular pipeline that enforces a single, auditable methodology across four stages: (1) *KOL discovery and identification* via social-graph expansion from curated seeds with hop-wise quality filters; (2) *verifiable-prediction* extraction and truth checking, retaining only claims that specify asset/target/time so they can be evaluated at predefined checkpoints (Hit/Miss/Partial/data-void); (3) *bot/astroturf risk* screening using multiple independent signals (template/clone text, timing regularity, homogeneous interactions, incentive links, structural imbalance), treated as *screening heuristics* for downweighting rather than definitive judgments; and (4) *influence and content quality* under bot-adjusted engagement, including secondary spread and six-dimensional content scoring (completeness, clarity, attractiveness, accuracy, marketing effectiveness, originality).

Design choices emphasize transparency and comparability. Only *verifiable* predictions are scored, with evaluation tied to a clearly defined time origin (post time) and ground-truth sources; ambiguous claims are excluded. Where appropriate, author-specific historical baselines provide

context, and intermediate signals (e.g., commenter uniqueness, duplication, temporal clustering) are surfaced for inspection. Outputs are delivered as interactive HTML reports—ranked overviews, account pages, and structured evidence snippets—to support independent verification.

Our contributions are threefold:

- **A unified, auditable pipeline** for crypto KOL assessment that couples social-graph discovery with verifiable-prediction checking, multi-signal bot-risk screening, and bot-adjusted influence/content scoring.
- **Operational definitions and decision rules** (e.g., what qualifies as a verifiable prediction; Hit/Miss/Partial/data-void outcomes; A–G bot signals; hard constraints such as minimum sample size and stability/calibration gates) that align scoring with reproducible evidence.
- **An interpretable scoring and grading scheme** that aggregates category scores with fixed weights and applies hard caps for borderline or high-risk cases, producing A/B/C/D grades accompanied by traceable evidence.

The remainder of the paper details the system overview (Section 2), methodology including verifiable-prediction checking, bot-risk signals, and scoring (Section 3), implementation and qualitative results (Section 4), related work (Section 5), and conclusions (Section 6).

2 System Overview

KOLIntel is organized as a transparent, end-to-end pipeline that couples social-graph discovery with verifiable-prediction checking, multi-signal bot/astroturf screening, and bot-adjusted influence/content scoring. The system is designed to produce explainable, independently verifiable outputs rather than opaque aggregate scores.

2.1 Architecture Overview

The framework comprises three primary components that work in concert:

- **KOL Discovery.** Starting from curated seed accounts on Twitter (X), the system performs multi-hop expansion over *Following*-based edges and applies hop-wise quality filters (e.g., activity, topical relevance, engagement sanity checks) to build a candidate pool and identify likely KOLs.

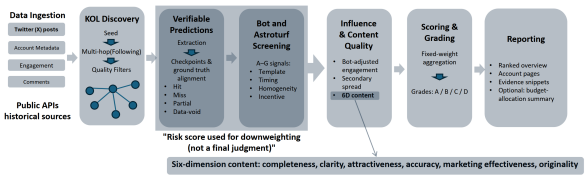

Figure 1: High-level architecture (feedback components removed). From left to right: *(i) KOL Discovery* expands from curated seeds via hop-wise quality filters to form a candidate pool; *(ii) Multi-Signal Evaluation* performs (a) verifiable-prediction extraction & truth checking, (b) bot/astroturf risk screening using template/clone text, timing regularity, homogeneous interactions, incentive links, and structural imbalance, and (c) influence & six-dimensional content scoring under bot-adjusted engagement; *(iii) Reporting* renders ranked overviews, account pages, and structured evidence snippets for independent verification.

- **Multi-Signal Evaluation.** For each candidate, the engine executes:
  1. *Verifiable-Prediction Extraction & Truth Checking*: only claims with sufficient structure (asset/target/time) are retained and evaluated at predefined checkpoints (Hit/Miss/Partial/data-void);
  2. *Bot/Astroturf Risk Screening*: aggregates independent signals (template/clone text, timing regularity, homogeneous interactions, incentive links, structural imbalance) into a risk score used for down-weighting rather than definitive judgments;
  3. *Influence & Content Quality*: computes bot-adjusted engagement and secondary spread, and scores content along six dimensions (completeness, clarity, attractiveness, accuracy, marketing effectiveness, originality).
- **Reporting and Visualization.** The pipeline generates an interactive HTML report with a ranked overview, account-level pages, and structured evidence snippets (prediction verdicts, bot-signal summaries, content-quality rationales), enabling independent verification. An optional view summarizes *budget allocation* that maps final grades to fund shares under simple caps and exclusion rules.

2.2 Design Principles and Technology Stack

Design Principles. *Modularity* (decoupled discovery, evaluation, and reporting); *Explainability*

(all intermediate signals and decisions surfaced for inspection); *Comparability* (single methodology and thresholds applied consistently); *Scalability* (batch processing for historical backfills and periodic refresh).

Technology Stack. The implementation uses Python with: (i) Twitter (X) APIs to collect posts, metadata, and engagement; (ii) an LLM-assisted layer for content quality scoring and prediction parsing (limited to verifiable claims); (iii) local storage (e.g., JSON/SQLite) for reproducible processing and audit trails; and (iv) templated HTML generation for interpretable reports. All thresholds and weights are configuration-driven to support ablations and sensitivity analyses. We rely on the official Twitter (X) Developer Platform and the Python language and documentation, and we cite their maintainers accordingly (Twitter, Inc., 2025; Python Software Foundation, 2025).

3 Methodology

We present a single, auditable methodology that turns raw social data into explainable, comparable assessments. The pipeline comprises four parts: (i) *verifiable-prediction* extraction and truth checking; (ii) multi-signal *bot/astroturf* screening; (iii) *influence* and *content-quality* measurement; and (iv) fixed-weight *scoring and grading* with light, stability-oriented constraints.

3.1 Verifiable Predictions: Extraction and Truth Checking

We evaluate only *verifiable* claims, i.e., posts that specify at least two of {asset, target (price/direction/range/event), deadline/window}, so that objective verification is possible. Each qualified post is structured as

$$\begin{aligned}&(\text{asset, target_type} \in \{\text{price, direction, range, event}\},\\ &\quad \text{target_value/dir, deadline/window,}\\ &\quad \text{optional: confidence, basis})\end{aligned}$$

Evaluation checkpoints (e.g., 6h /24h /3d /7d /14d /30d) are inferred from textual cues and author style. For price targets we replay historical prices from post time T_0 to each checkpoint; for event targets we align to authoritative confirmations before the deadline. Each claim receives a verdict in {HIT, MISS, PARTIAL, DATA-VOID}. A post-level prediction score $q_i \in [0, 1]$ is then computed (Hit=1, Miss=0, Partial in $[0.5, 0.9]$), optionally calibrated by declared confidence; extreme pre-hit drawdowns/volatility may lightly downweight q_i. At the author level, we report rolling and weighted hit rates, calibration error, and stability metrics; when sample size is small, summaries are retained with a visible *low-sample* caveat and later caps in grading. Atypical windows (e.g., rare exogenous shocks) are treated with modest robustness so short-lived extremes do not dominate long-horizon conclusions.

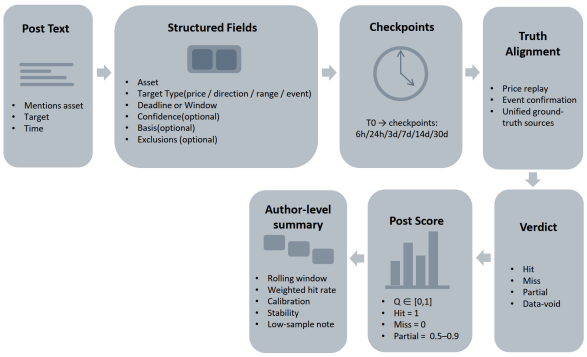

Figure 2: Verifiable-prediction flow: text → structured fields (asset/target/time) → checkpoints → ground-truth alignment → verdicts (Hit/Miss/Partial/data-void) → post score → author-level summaries (rolling/weighted hit rate, calibration, stability, sample-size note).

3.2 Bot/Astroturf Risk (Multi-Signal Screening)

We screen for inorganic amplification using independent indicators. Let $U_c(t_i)$ be the set of unique commenters for post t_i, c_i its total comments, $\rho_{\text{dup}}(t_i)$ the duplicate-content proportion, and $\tau_{\text{anomaly}}(t_i)$ a temporal-clustering indicator. The enhanced authenticity score (**EAS**) is

$$\begin{aligned}\text{EAS}(t_i) = \gamma_1 \frac{|U_c(t_i)|}{c_i} + \gamma_2\big(1 - \rho_{\text{dup}}(t_i)\big)\\ + \gamma_3\big(1 - \tau_{\text{anomaly}}(t_i)\big),\\ \text{with } \gamma_1 + \gamma_2 + \gamma_3 = 1.\end{aligned} \tag{1}$$

To capture short templated bursts and behavioral regularities, we define a complementary score

$$\begin{aligned}\text{SpamScore}(t_i) = \alpha\, \rho_{\text{short}}(t_i) + \beta\, \rho_{\text{dup}}(t_i)\\ + \gamma\, \tau_{\text{anomaly}}(t_i),\\ \text{with } \alpha + \beta + \gamma = 1.\end{aligned} \tag{2}$$

where $\rho_{\text{short}}(t_i)$ is the short-comment ratio (e.g., < 5 characters). Seven signal families (A–G: template/clone text; timing regularity/bursts;

homogeneous interactions; incentive links; self-identification; high-quality natural-interaction counter-signal; structural imbalance) are summarized into an account-level risk $Q_{\text{bot}} \in [0, 1]$ and a tag from the set {`human_like`, `ai_like`, `unclear`, `not_applicable`}. These are *screening heuristics* used for downweighting or capping, not definitive judgments; institutional/official accounts typically receive `not_applicable` to avoid mixing propagation regimes.

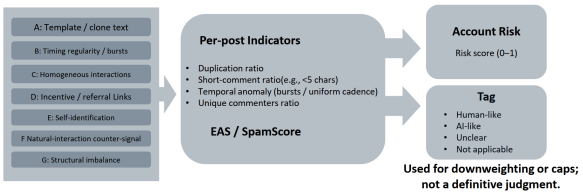

Figure 3: Bot/astroturf signals map: seven signal families (A–G) → per-post indicators (duplication, timing bursts, etc.) → account-level risk Q_{bot} and tag; high-risk cases are downweighted or capped in grading.

3.3 Content Quality and Context

Each post t_i is scored along six dimensions $\mathcal{D}$ with weights $\{w_j\}_{j=1}^{6}$:

$$\text{CQS}(t_i) = \sum_{j=1}^{6} w_j \, s_{i,j}, \qquad s_{i,j} \in [1, 5], \qquad \sum_{j=1}^{6} w_j = 1. \tag{3}$$

Dimensions cover completeness, clarity, attractiveness, accuracy, marketing effectiveness, and originality; rationales are surfaced so each score maps to concrete evidence (e.g., missing key details penalize completeness). Where informative, author-specific historical context is reported via

$$\text{CQS}'(t_i) = \frac{\text{CQS}(t_i) - \mu_{\text{KOL}}}{\sigma_{\text{KOL}} + \epsilon}, \tag{4}$$

with μ_{KOL} and σ_{KOL} the author's historical mean and standard deviation. Cross-lingual posts are translated to English before analysis to reduce semantic drift.

3.4 Influence Under Bot-Adjusted Engagement

Influence is measured after accounting for suspected inorganic effects (via Q_{bot} and post-level indicators), combining adjusted engagement, secondary spread (e.g., credible re-shares), and audience quality. The result is a normalized score $Q_{\text{infl}} \in [0, 1]$ comparable across candidates; windows dominated by rare exogenous events are handled with modest robustness to preserve comparability.

3.5 Scoring and Grading (Fixed Weights + Light Constraints)

We aggregate four categories with fixed weights: prediction performance Q_{pred}, bot risk Q_{bot}, influence Q_{infl}, and content quality Q_{cont}:

$$S = 0.35\, Q_{\text{pred}} + 0.20\,(1 - Q_{\text{bot}}) + 0.20\, Q_{\text{infl}} + 0.15\, Q_{\text{cont}}. \tag{5}$$

To guard against overfitting to luck or noise, we apply light stability-oriented constraints: a minimum sample-size note and cap for sparse predictors; gentle caps for poor calibration/instability; and caps or grey-listing for high bot-risk. Exceptional one-off windfalls thus do not dominate long-run grades. We map S to {A/B/C/D} using transparent thresholds and disclose the evidence behind each category. When desired, the final grades can also be summarized in the report as an optional budget-allocation view that proportionally maps grades to shares under simple caps; we do not analyze this view further in the paper.

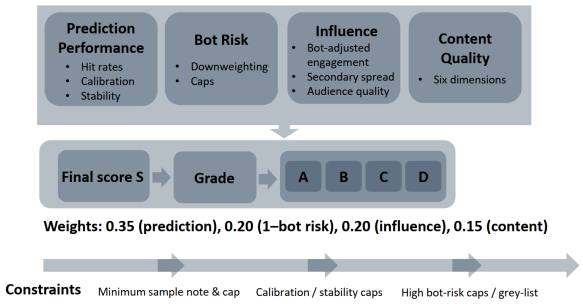

Figure 4: Scoring framework: four category scores $\{Q_{\text{pred}}, Q_{\text{bot}}, Q_{\text{infl}}, Q_{\text{cont}}\}$ combined by fixed weights into S (Eq. 5); light constraints promote stability and accountability; outputs include grade (A/B/C/D) and evidence.

4 Implementation & Results

This section describes an implementation of **KOLIntel** as a modular, reproducible pipeline and reports qualitative results from a small-scale run. We intentionally avoid prescriptive automation or strong performance claims; instead, we emphasize transparent inputs, consistent processing, and explainable outputs.

4.1 System Implementation

The pipeline is organized as decoupled stages that can run end-to-end or as stand-alone modules:

- **Data Ingestion.** Batched collection of posts, account metadata, and engagement from Twitter (X) APIs; optional retrieval of comment threads for bot/astroturf screening.
- **Normalization and Storage.** Canonicalization of timestamps, assets, and identifiers; local storage (e.g., JSON/SQLite) to support deterministic re-runs and audit trails.
- **Evaluation Engine.** (i) verifiable-prediction extraction and truth checking; (ii) multi-signal bot/astroturf screening; (iii) content-quality and influence computation under bot-adjusted engagement.
- **Scoring & Grading.** Category scores are combined with fixed weights (Eq. 5); light, stability-oriented constraints prevent sparse or high-risk cases from dominating grades.
- **Reporting.** An interactive HTML dashboard renders a ranked overview, account-level pages, and structured evidence snippets (prediction verdicts, bot-signal notes, content-quality rationales). Bot risk is presented as a screening signal *used for downweighting/caps*, not a definitive judgment.

4.2 Experimental Setup (Inputs and Procedure)

We provide a repository anonymized with code, configurations, and minimal data to reproduce our pipeline. the link is given in Rreference(Anonymous, 2025).

We conducted a small-scale run to exercise the full pipeline:

- **Scope.** A curated cohort of **100 crypto-focused accounts** discovered by social-graph expansion from seeds and **manually vetted** for authenticity; posts restricted to a bounded time range.
- **Data.** Post texts with timestamps and engagement, author metadata, and (when available) comment threads; for price/event verification, only publicly auditable sources were used. *Before rendering the report, all personally identifying account fields (display name, handle, profile image, bio text, numeric IDs, and external links) were replaced with randomized placeholders; underlying metrics, timelines, and verdicts were not modified.*
- **Procedure.** (1) ingest and normalize; (2) run the evaluation engine (prediction verdicts, bot signals, content quality, influence); (3) aggregate category scores and apply grading; (4) generate the HTML report.

The run validates processing logic and report structure rather than generalizable performance figures; on this set, the *KOLAgent*-assisted prediction parsing achieved **95%** agreement with manual labels.

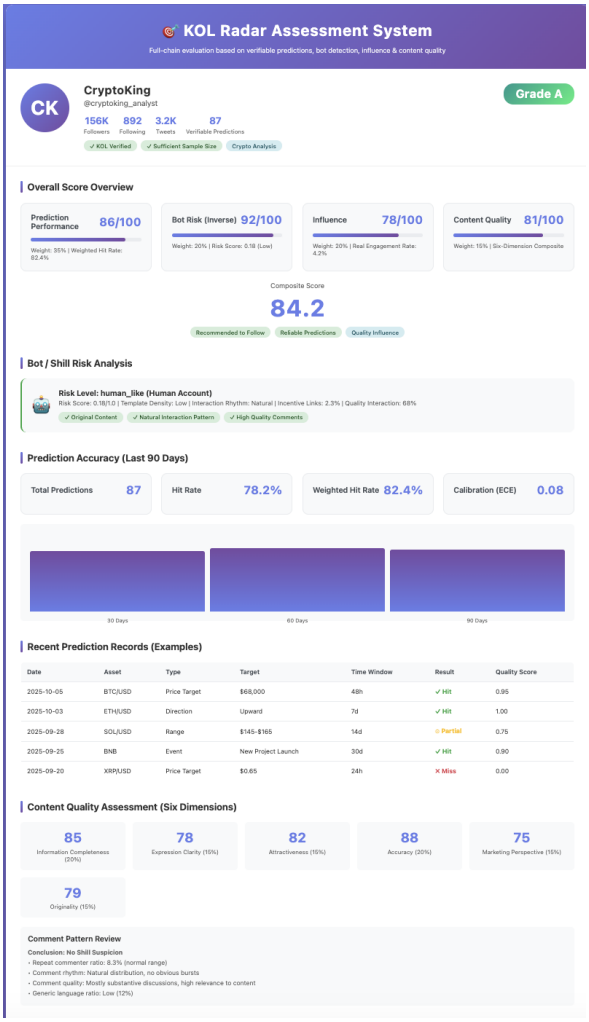

Figure 5: HTML report snapshot with *real underlying results* and *randomized account identifiers*. Panels summarize the four categories—Prediction Performance, Bot Risk (screening/downweighting), Influence, and Content Quality—plus a composite grade. Personal identifiers are replaced; metrics and verdicts are unchanged.

4.3 Qualitative Observations (No Strong Claims)

The dashboard in Figure 5 illustrates several consistent, explainable patterns:

- **Verifiable predictions.** Posts that specify asset/target/time can be adjudicated at checkpoints (HIT/MISS/PARTIAL/DATA-VOID), yielding post-level scores and author-level summaries (rolling/weighted hit rates and calibration).

- **Bot/astroturf screening.** Template/clone text, timing regularity/bursts, and homogeneous interactions surface as practical indicators for *downweighting* or caps; institutional/official accounts are labeled `not_applicable` to avoid mixing propagation regimes.

- **Influence and content quality.** Influence under bot-adjusted engagement (with secondary spread and audience-quality proxies) complements six-dimension content scoring, enabling evidence-backed A/B/C/D grades.

These observations show how auditable inputs translate into interpretable outputs; they are not positioned as aggregate accuracy or superiority claims.

4.4 Efficiency and Reproducibility

All stages operate in batch mode and are configuration-driven (thresholds, weights, and time windows). Local storage and deterministic preprocessing enable exact re-runs of the same inputs to reproduce identical outputs. While end-to-end throughput depends on external API latency and data volume, the modular design allows parallelization across accounts or time slices when needed. We do not report speed or precision/recall figures here, as the goal is to document process fidelity and interpretability rather than to claim benchmark performance.

4.5 Deliverables

The pipeline produces:

- **Interactive HTML report:** ranked overview, account pages, and structured evidence snippets, suitable for review and audit; bot-risk displays are explicitly noted as screening signals used for downweighting/caps.

- **Structured artifacts:** normalized inputs, intermediate features (e.g., prediction verdicts, bot indicators), and final scores/grades for downstream analysis.

5 Related Work

Research on social-media influencer assessment spans influencer marketing analytics, bot/astroturf detection, crypto-market dynamics, and LLM-assisted content evaluation.

Influencer marketing and content analysis. Computational studies analyze how to identify influencers, characterize strategies, and evaluate sponsored content. Gui et al. (2024) survey core themes including influencer identification, engagement modeling, disclosure and fairness. Multimodal pipelines have been explored for commercial-content detection (Sánchez Villegas et al., 2023), while audience authenticity and inauthentic promotions have been studied via multi-relational or automated signals (Kim et al., 2021a,b). NLP-based quality assessment for social media has also been shown to be effective (Fang and Wang, 2022). Our pipeline aligns with this line of work by coupling account discovery with structured, auditable post-level analysis.

Bot/astroturf detection and crypto-market manipulation. Bot detection spans feature- and graph-based methods and highlights susceptibility to domain shift (Orabi et al., 2020). Generative/adversarial approaches have been proposed for robust detection (Shukla et al., 2023), and decentralized/permissionless settings introduce additional challenges (Lee et al., 2021). In crypto, manipulation and coordinated campaigns have been documented (Liu et al., 2021; Feng et al., 2021; Kamps and Kleinberg, 2018), motivating authenticity checks beyond surface metrics. We incorporate multi-signal screening (template/clone text, timing regularity/bursts, homogeneous interactions, incentive links, structural imbalance) to produce an interpretable risk score used for downweighting.

LLM-assisted evaluation and domain context. LLMs have been applied to characterize social-media content and extract structured signals (Jin et al., 2022). KOL-oriented language agents further explore tone, expertise, and task-specific prompts (Anonymous, 2024). Given the interplay between discourse and market movement in

crypto (Huang et al., 2021), we emphasize verifiable, time-anchored prediction checking and evidence-backed scoring. Interpretable scoring schemes can be preferable for decision support in finance-related contexts (Cretarola and Figà Talamanca, 2022).

6 Conclusion

We presented **KOLIntel**, a transparent pipeline that assesses crypto-focused accounts on Twitter (X) using a single, auditable methodology. The approach couples four core stages end-to-end: (i) KOL discovery via social-graph expansion with hop-wise quality filters; (ii) verifiable-prediction extraction and truth checking that adjudicates claims at predefined checkpoints; (iii) multi-signal bot/astroturf screening that informs downweighting through interpretable indicators; and (iv) influence and six-dimensional content-quality measurement under bot-adjusted engagement, followed by fixed-weight scoring and grading. Throughout, evidence is made explicit—structured prediction verdicts, bot-signal summaries, and reasoned content scores—so conclusions can be independently verified under a consistent set of rules. The reporting layer renders these results as ranked overviews and account pages with traceable rationale; when desired, final grades can be mapped to proportional budget shares using simple, auditable rules. By emphasizing verifiable evidence, unified evaluation criteria, and interpretable presentation, KOLIntel supports accountable decision-making for influencer screening and monitoring without relying on opaque heuristics or undeclared assumptions. Future work can broaden coverage, refine signal quality, and study longitudinal robustness, but the core contribution is a reproducible, end-to-end path from public social data to explainable, reviewable assessments.

Limitations

Our methodology aims for auditability and comparability, but it entails several limitations. First, coverage and timeliness depend on public APIs and the availability/consistency of historical price or event sources; outages or cross-source conflicts can yield `data-void` verdicts that are excluded from denominators. Second, requiring structured elements (at least two among asset, target, and time window) improves accountability but reduces recall by excluding ambiguous or rhetorical claims; languages with rich morphology or colloquial markers may challenge extraction despite normalization and translation. Third, bot/astroturf indicators are screening heuristics for downweighting rather than definitive judgments; subtle coordination may be missed, and news-driven bursts can occasionally be flagged as suspicious. Fourth, influence is computed under bot-adjusted engagement and credible secondary spread, yet audience-quality proxies and topical filters can inherit platform-specific biases. Fifth, short evaluation windows at ground-truth checkpoints are stabilized with modest robustness settings (e.g., rolling windows and mild caps), but extreme market regimes or exogenous shocks may still distort local outcomes; long-horizon interpretations should remain cautious. Sixth, LLM-assisted parsing and content scoring can exhibit model/version drift and threshold sensitivity; while configuration files improve reproducibility, absolute scores may shift under alternative model choices or hyperparameters. Finally, the approach is tailored to crypto discourse and verification sources; transferring it to other domains with different ground truths or interaction norms will require re-specifying data sources, thresholds, and calibration procedures.

Ethics Statement

This study analyzes publicly accessible Twitter (X) posts and engagement metadata and does not access private data or circumvent platform controls. Data collection follows the platform's Terms of Service and applicable rate limits. We do not infer sensitive attributes (e.g., race, religion, health) and do not attempt de-anonymization.

To reduce privacy risk in artifacts, any showcased screenshots (e.g., Figure 5) use *de-identified real results*: all metrics and verdicts are unaltered, while personal identifiers (display name, handle, profile image) are replaced with randomized placeholders. We do not release raw, identifiable datasets. An anonymized system-code archive is provided for peer review to exercise the components described in the paper; broader availability or redistribution is out of scope for this submission and depends on governance and applicable platform policies.

The bot/astroturf component is explicitly a *screening* aid for downweighting or caps, not an accusation mechanism. Outputs are continuous risk scores accompanied by concrete, reviewable evidence windows (e.g., timing regu-

larities, duplicate content) and conservative labels (e.g., `unclear`, `not_applicable` for institutional/official accounts). The reporting workflow surfaces rationale for human review and supports correction or removal procedures to mitigate potential reputational harm.

Prediction evaluation is restricted to verifiable claims with objective ground truth; ambiguous or rhetorical posts are excluded. Cross-lingual and time-zone normalization may introduce minor noise; such limitations are disclosed. Influence scoring uses bot-adjusted engagement and audience-quality proxies that may inherit platform biases; results are presented as decision support for auditing and research rather than definitive judgments about individuals. This work is not investment advice; when the optional budget-allocation view is enabled, it is a rules-based visualization driven by evidence-backed grades under simple caps.

Intended Use & Scope. KOLIntel is intended for research and auditing on public social-media data. It must not be used to target or penalize individuals, and it should not be deployed to infer or profile any sensitive attributes. Users must comply with platform Terms of Service and honor removal/opt-out requests for showcased examples.

AI-Writing Disclosure. We used generative AI tools solely for language assistance (bilingual translation, grammar/style edits, and sentence-level paraphrasing). No tool was used to generate technical content, methods, analyses, or results. All AI-suggested text was reviewed and edited by the authors, and we checked for potential plagiarism.

References

Anonymous. 2024. Mindecho: Role-playing language agents for key opinion leaders. *arXiv preprint*. Paper under review.

Anonymous. 2025. Anonymized repository (supplementary materials). https://anonymous.4open.science/r/crptoagent-6BA8.

Alessandra Cretarola and Andrea Figà Talamanca. 2022. Social media sentiment and bitcoin market dynamics: A high-frequency perspective. *Digital Finance*, 4:89–116.

Xing Fang and Tianfu Wang. 2022. Using natural language processing to identify effective influencers. *International Journal of Market Research*, 64(5):611–629.

Wenqi Feng, Yunyao Xie, Xiang Li, Xuan Zhang, Bo Wang, and Qiang Xu. 2021. Detecting cryptocurrency pump-and-dump schemes on social media: A cross-platform approach. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 304–312.

Haoyang Gui, Thales Bertaglia, Catalina Goanta, and Gerasimos Spanakis. 2024. Computational studies in influencer marketing: A systematic literature review. *arXiv preprint*.

Xin Huang, Wenbin Zhang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang. 2021. Lstm based sentiment analysis for cryptocurrency prediction. *arXiv preprint arXiv:2103.14804*.

Mali Jin, Daniel Preotiuc-Pietro, A Seza Dogruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959.

John Kamps and Bennett Kleinberg. 2018. To the moon: Defining and detecting cryptocurrency pump-and-dumps.

Seungbae Kim, Xiusi Chen, Jyun-Yu Jiang, Jinyoung Han, and Wei Wang. 2021a. Evaluating audience loyalty and authenticity in influencer marketing via multi-task multi-relational learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 278–289.

Seungbae Kim, Jyun-Yu Jiang, and Wei Wang. 2021b. Discovering undisclosed paid partnership on social media via aspect-attentive sponsored post learning. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 319–327.

Seungyeop Lee, Seongwook Jeong, Youngseok Choi, and Hyoungshick Kim. 2021. Sybil detection in decentralized social networks using local rule-based

trust propagation. In *Proceedings of the IEEE International Conference on Blockchain (Blockchain 2021)*, pages 158–165. IEEE.

Yuchen Liu, Yujie Li, Haoran Liu, Xiaowei Hu, and Raymond Y.K. Lau. 2021. Understanding cryptocurrency pump-and-dump campaigns: Bounty hunters and fake crowds. In *Proceedings of the Web Conference 2021 (WWW)*, pages 3125–3136.

Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of bots in social media: A systematic review. *Information Processing & Management*, 57(4):102250.

Python Software Foundation. 2025. *Python Language Reference*. Version 3.x; Accessed: 2025-10-06.

Danae Sánchez Villegas, Catalina Goanta, and Nikolaos Aletras. 2023. A multimodal analysis of influencer content on twitter. *arXiv preprint arXiv:2309.03064*.

Anant Shukla, Martin Jureček, and Mark Stamp. 2023. Social media bot detection using dropout-gan. *arXiv preprint arXiv:2311.05079*.

Twitter, Inc. 2025. Twitter (x) developer platform documentation. https://developer.x.com/en/docs. Accessed: 2025-10-06.

A Appendix

This appendix summarizes operational definitions, decision rules, and system details for **KOLIntel**, a deterministic, end-to-end pipeline for transparent assessment of crypto-focused accounts on Twitter (X).

A.1 A. System Artifact (Scope)

Scope. The pipeline comprises four stages: (1) KOL discovery and identification via social-graph expansion from curated seeds with hop-wise quality filters; (2) *verifiable-prediction* extraction and truth checking, retaining only claims with sufficient structure (asset, target type/value or direction/range/event, and a time window) and adjudicating them at unified *ground-truth checkpoints*; (3) bot/astroturf risk screening using multiple independent signals treated as *screening heuristics* for *downweighting* rather than definitive judgments; (4) influence and *six-dimensional* content-quality assessment under *bot-adjusted engagement* and secondary spread, followed by fixed-weight scoring and final letter grades (A/B/C/D). No external pretrained models are required; all thresholds and rules are configuration-driven.

A.2 B. Data Governance, Privacy, and Intended Use

Public data only. We process publicly available posts and profile metadata for aggregate analysis. No redistribution of raw posts, user IDs/handles, or other identifying fields. **PII/content risk.** Outputs are aggregate statistics and redacted/paraphrased snippets; figures avoid content reasonably linkable to specific individuals. **Intended use.** Research on transparent, auditable influence measurement; not for law-enforcement, surveillance, or individual-level credit assignment. **Compliance.** Collection and analysis comply with platform Terms of Service and rate limits; no bypass of access controls.

A.3 C. Schema for Verifiable-Prediction Extraction

We retain only posts that specify at least two of {`asset`, `target`, `deadline/window`}. Normalized fields:

- `asset` (string): canonical ID (e.g., `BTC`, `ETH`).
- `target_type` (enum): {`price`, `direction`, `range`, `event`}.
- `target_value` (number/string; conditional on type).
- `window` (ISO8601 instant or [start,end] interval; UTC unless noted).
- `stance` (enum): {bullish, bearish, neutral}.
- `evidence_link` (string): permalink or internal ID.

Ground-truth checkpoints. Evaluation checkpoints are chosen from {6h, 24h, 3d, 7d, 14d, 30d} based on textual cues and author style, ensuring consistent adjudication windows.

A.4 D. Adjudication Rules: Hit / Miss / Partial / data-void

Price/direction/range/event claims are adjudicated against the selected checkpoint set:

- **HIT**: target reached/touched before the relevant deadline/checkpoint (or correct direction conclusively realized).
- **MISS**: deadline elapsed without touch (or opposite direction conclusively realized for direction claims).

- **PARTIAL**: 95%–100% proximity to target or within a stated tolerance band; directionally correct with small shortfall.
- **DATA-VOID**: conflicting or missing ground truth (e.g., exchange outage, inconsistent sources); excluded from denominator-level accuracy.

A.5 E. Multi-Signal Bot/Astroturf Risk Screens

Independent, interpretable indicators used as *screening heuristics* for *downweighting* or *caps*:

- **Template/clone text**: near-duplicate n-gram patterns across accounts.
- **Temporal regularity**: abnormal posting periodicity or 24/7 patterns.
- **Homogeneous engagement**: tightly coupled like/retweet clusters.
- **Incentive links**: persistent referral/airdrop campaigns.
- **Structural imbalance**: anomalous follower–following ratios.
- **Freshness mismatch**: sudden bursts after long dormancy.
- **Handle churn**: frequent changes to handle/bio/profile picture.

High Q_{bot} may trigger a score *cap* or a *grey-listed* tag; `not_applicable` excludes institutional/official accounts from penalties/bonuses.

A.6 F. Scoring and Decision Rules

Components. (i) *Influence* (reach/engagement calibrated by graph position, including secondary spread); (ii) *Content Quality* under a *six-dimensional* rubric—completeness, clarity, attractiveness, accuracy, marketing effectiveness, originality—combined with the rate of verifiable-prediction production and accuracy; (iii) *Risk adjustment* from bot/astroturf screens. **Hard constraints and caps.**

- *Minimum sample size*: verifiable-prediction count below a threshold (e.g., 20) ⇒ visible *low-sample* note and a grade cap.
- *Calibration/Stability*: poor calibration or excessive volatility over rolling windows ⇒ gentle grade cap.
- *High bot risk*: cap or grey-listing as in Section A.5; `not_applicable` for institutional/officials.

Final letter grades (A/B/C/D) map to score intervals after caps under fixed, transparent thresholds.

A.7 G. Transparency: Aggregate Run-Time Reporting

We report aggregate statistics: number of discovered accounts, fraction flagged by each screen, number of evaluated verifiable predictions, and the adjudication distribution (`Hit/Miss/Partial/data-void`). No raw identifiers are released.