Heuristic-Augmented Denoising in Distant Supervision for Crypto Assets Named Entity Recognition Using IndoBERTweet-CRF

Jonathan Davin
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jonathan.davin@binus.ac.id

Shane Anthony
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
shane.anthony@binus.ac.id

Bryan Dale
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
bryan.dale@binus.ac.id

Clark Sompie
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
clark.sompie@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

Rifqi Charisma
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
rifqi.charisma@binus.ac.id

***Abstract*—Indonesia's rapidly expanding digital asset market has fostered an active cryptocurrency community on Twitter (X), but extracting entities from informal social media text is difficult because of the community's complex linguistic registers, high out-of-vocabulary rates, and the scarcity of domain-specific annotated datasets for Indonesian. This study proposes a hybrid Named Entity Recognition (NER) pipeline built on an IndoBERTweet-CRF architecture. Distant Supervision labels a Silver Standard of 25,726 tweets using the CoinGecko API as a Local Knowledge Base, with a Heuristic-Augmented Denoising mechanism to mitigate label noise from lexically ambiguous terms, after which the base model is fine-tuned on a Gold Standard of 1,461 manually annotated tweets whose token-level Fleiss' κ is 0.9654. Four comparative baselines and four ablated variants are evaluated under an identical 5-Fold Cross-Validation protocol, with differences assessed by paired bootstrap resampling and approximate randomization under Holm correction. The proposed model reaches a mean Precision of 83.28%, a Recall of 87.65%, and an F1-Score of 85.41%. Ablation attributes 26.36 points of that score to hybrid fine-tuning, 5.19 to distant supervision, and 1.25 to heuristic denoising, while the Conditional Random Field layer is indistinguishable from a softmax head under a single-type, predominantly single-token schema. Residual errors are dominated by asset–platform metonymy and by a recall gap between entities inside the knowledge base (0.9550) and outside it (0.5442). Code: https://github.com/SauHin/crypto-ner-indobertweet-crf.**

Keywords—Named Entity Recognition (NER), Crypto assets, Distant supervision, IndoBERTweet-CRF, Indonesian Twitter, Ablation study.

I. INTRODUCTION

The digital asset market in Indonesia has expanded rapidly: the Otoritas Jasa Keuangan (OJK) recorded a 46.94% increase in registered users over nine months, from 13.3 million in February 2025 to 19.5 million in November 2025 [1]. Much of that growth developed alongside the crypto community on Twitter (X), which has become a primary source of investment information because mainstream financial media and institutional analysts rarely cover digital asset projects [2], [3], and where sentiment and discussion activity measurably influence the price movements and trading volumes of certain cryptocurrencies [4].

That dependence carries financial risk. Influencer tweets have been associated with short-term price volatility averaging 1.83% in daily returns, followed by corrections as steep as 6.53% over longer horizons [2], and the Federal Trade Commission reported more than $1 billion in consumer losses to cryptocurrency-related fraud between January 2021 and March 2022, nearly half of it beginning with an advertisement, post or message on social media [5]. Monitoring thousands of posts continuously requires automation, and Named Entity Recognition (NER) is its necessary first step: before sentiment or risk analysis can take place, the system must identify which assets are actually being discussed.

That identification task is far from straightforward. Crypto communities operate with at least 52 distinct linguistic registers [3]: domain-specific abbreviations (FUD, NFT, DeFi), ordinary words repurposed with specialized meanings ("miner," "whales," "block"), and insider vocabulary with no equivalent outside the community ("altcoin," "gamefi"). Surface-level patterns cannot determine whether such a term names an entity or is used in its everyday sense, and the informal, abbreviated nature of social media text compounds this with heavy Out-of-Vocabulary exposure [6], particularly for numerical values [7], [8].

Building a domain-specific annotated dataset is itself a problem. Manual annotation at the required scale has consistently slowed NLP development for low-resource languages such as Indonesian [9], while distant supervision automates labeling at the cost of noisy labels [10], arising both from entities missed by the labeling process and from entities assigned to the wrong type [11]. Financial entity extraction for Indonesian has been addressed by IPerFEX, covering 15 personal finance entity types [7], but not for the crypto domain, and reviews of available datasets show Indonesian NER resources concentrated in the general, healthcare and political domains, omitting cryptocurrency entirely [12], [13].

The present work addresses this directly. Distant Supervision is adopted, using the CoinGecko API as an external knowledge base to annotate Indonesian-language tweets automatically, with a Heuristic-Augmented Denoising mechanism to reduce label noise, after which IndoBERTweet is fine-tuned on the resulting dataset [14]. Alongside the pipeline, this study contributes an annotated Indonesian crypto NER corpus with reported inter-annotator agreement, a controlled comparison against four alternative architectures under one protocol, a component-wise ablation, and an error

analysis decomposed by knowledge base membership, surface form, span length and sampling stratum.

II. Related Works

A. NER for Indonesian and for Social Media Text

Named Entity Recognition research for Indonesian has shifted from static feature extraction toward Transformer-based contextual representations. Budi and Suryono [12] documented that earlier systems built on algorithms such as Support Vector Machines faced a chronic shortage of standardized datasets and could not capture sentence-level context, whereas IndoBERT [8], pre-trained on the Indo4B corpus, handled Indonesian morphology through sub-word representations and outperformed both mBERT and FastText on the NERP dataset; Khairunnisa et al. [13] then fed IndoBERT representations into a BiLSTM-CRF stack on a formal news corpus.

Social media text resists these pipelines: posts are short, grammatically irregular and dense with OOV terms [6]. Yanti et al. [15] responded with aggressive preprocessing, removing mentions, URLs, hashtags and all numerical content, which worked for location entities but is a serious liability in financial contexts where those figures are often the entities themselves. Koto et al. [14] instead developed IndoBERTweet through domain-adaptive pretraining with a domain-specific vocabulary initialization, outperforming both mBERT and standard IndoBERT on an informal NER benchmark. Alfatah [16] reports the same direction for sentiment classification on Indonesian Twitter data, a task whose scores are not comparable with entity-level NER figures.

B. NER in the Financial and Crypto Asset Domains

Financial entity extraction is harder than general-domain NER for two compounding reasons: numerical values are semantically ambiguous without context, and community vocabulary remains largely incomprehensible to models trained on general corpora. Affandi [3] catalogued dozens of registers used in Indonesian crypto communities, including whale, miner and bearish, whose meanings differ sharply from their everyday definitions. Dave and Chowanda [7] tackled Indonesian financial entity extraction with an IndoBERT-BiGRU-CRF architecture, surfacing an extreme OOV problem specific to numerical tokens and finding a CRF layer necessary to model inter-token dependencies, while Wang [17] evaluated FinBERT on English financial documents. What remains missing is a dataset or model bringing the registers of Indonesian crypto communities into a Transformer-based NER pipeline.

C. Distant Supervision for Dataset Scarcity

Distant supervision sidesteps manual annotation by aligning raw text to entries in an external knowledge base; Wei et al. [18] report a reduction in dataset construction time from this alignment. The cost is label quality, since automatic alignment introduces noise wherever terms are lexically ambiguous. Hedderich et al. [9] addressed this through ANEA, using rule-based filtering as a post-hoc denoising step that lowered false positive rates on low-resource benchmarks, while algorithmic alternatives pursue the same goal through loss design: Meng et al. [10] down-weight likely-noisy tokens via Generalized Cross-Entropy in RoSTER, and Wang et al. [11] separate boundary detection from type classification in DesERT. Both report state-of-the-art performance on English corpora, but for locally-scoped social media datasets with scarce training data, domain-specific heuristic filtering is more stable and carries lower risk of overfitting.

III. Methodology

To train a NER model capable of identifying cryptocurrency asset entities, this study adopts a hybrid methodology combining Distant Supervision with manual annotation. The complete workflow is illustrated in Fig. 1.

A. Dataset

The dataset consists of primary data extracted from Twitter (X) by web scraping with advanced search queries targeting Indonesian-language tweets about cryptocurrency. Search parameters combined crypto entity keywords, covering both specific coins (Bitcoin, ETH, Solana) and general community terms (altcoin, memecoin, gamefi), with crypto transaction terms drawn from Indonesian trader slang (serok, aggressive accumulation; cuan, profit; nyangkut, stuck in a losing position). Collection was bounded to a one-year window from April 1, 2025 to April 1, 2026 by two disjunctive queries, released verbatim alongside the dataset, which yielded a raw corpus of 60,283 tweets.

B. Data Pre-processing

Raw social media text is noisy and structurally inconsistent, so a four-stage pre-processing pipeline was applied before

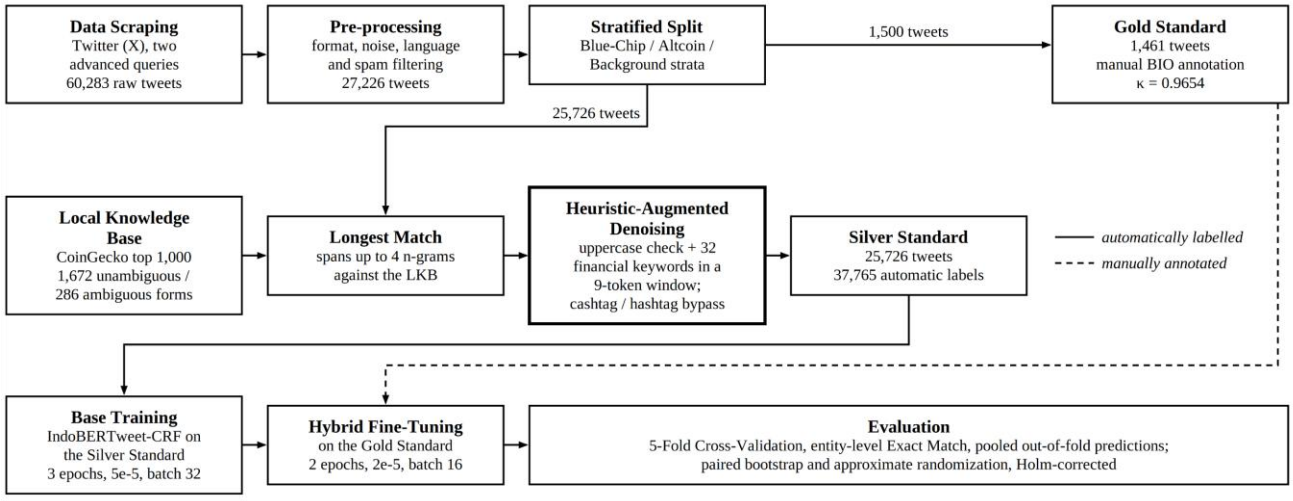

Fig. 1. Proposed Methodology

annotation. (1) IndoBERTweet format standardization replaced mentions and URLs with @USER and HTTPURL, removed redundant whitespace and stripped irrelevant symbols while retaining letter casing, numerals, dollar signs, hashtags and basic punctuation. Casing is preserved not as an encoder feature, so BTC and btc resolve to identical token identifiers, but because the denoising heuristic in Section III-F inspects the raw surface form before tokenization. (2) Global noise filtering applied a regular expression blacklist discarding tweets that referenced non-crypto contexts such as e-commerce, electricity tokens, gaming vouchers, K-Pop and cosmetics vocabulary, and giveaway content. (3) Hybrid language filtering retained a tweet if its metadata carried an Indonesian language tag or if it contained common Indonesian function words (yg, udah, aja, kalo, sih), since Twitter's native detection does not reliably separate Indonesian from Malay. (4) Spam filtering discarded tweets below 15 characters, or carrying five or more mentions or six or more hashtags.

C. Building the Local Knowledge Base

Distant Supervision relies on a comprehensive reference dictionary. A Local Knowledge Base (LKB) was built by querying the CoinGecko API for the top 1,000 assets by market capitalization, extracting the official name, ticker symbol and coin identifier for each. Refinement was required because many cryptocurrency names overlap with common Indonesian or English words, and produced three tiers: a Blacklist of single-character entries and names too generic to be reliable markers (gpt, game, ai, koin), which were removed; an Ambiguous tier of coins whose names carry an everyday secondary meaning (Ini, Dia, Gas, Meme), subject to stricter matching constraints; and an Unambiguous tier of names that exclusively denote an asset (Bitcoin, Ethereum, Cardano), matched directly, to which aliases such as betece for Bitcoin were added manually. Matching operates on surface forms: every key is expanded into its identifier, ticker symbol, official name, parenthesis-stripped name and aliases, after which the Unambiguous tier resolves to 1,672 distinct surface forms and the Ambiguous tier to 286.

D. Data Splitting and Tokenization

The pre-processed dataset was split into a Silver Standard for automated annotation via Distant Supervision and a Gold Standard for manual annotation. A sample of 1,500 tweets was drawn for the Gold Standard by Stratified Sampling across three strata: Blue-Chip (450 samples), covering the most widely held coins (BTC, ETH, SOL, BNB, XRP); Altcoin (750 samples), covering mid- to low-capitalization assets, allocated the largest share because vocabulary coverage at this tier drives generalization; and Background (300 samples), tweets containing financial terminology but no specific coin, included for robustness against false positives. All remaining tweets formed the Silver Standard, and both subsets were word-tokenized in preparation for BIO sequence annotation.

E. Manual Annotation and Inter-Annotator Agreement

Gold Standard samples were annotated manually by three annotators using the BIO scheme, with B-CRYPTO marking the first token of a coin entity, I-CRYPTO continuation tokens and O all non-entity tokens. Guidelines required contextual interpretation over surface-level name matching and restricted positive labels to the traded asset itself; a name referring to a company, a network or a platform was labeled O, and tweets written entirely in a foreign language were excluded during annotation.

Primary annotation was divided evenly among the three annotators, with discussion reserved for deadlock, so each tweet carries a single annotator's judgment. A stratified overlap sample of 150 tweets (45 Blue-Chip, 75 Altcoin, 30 Background) was independently re-annotated by all three, blind to the labels already assigned. Token-level Fleiss' κ on this sample is 0.9654 and pairwise entity-level F1-Scores are 0.9858, 0.9638 and 0.9489. These figures characterize the overlap sample rather than the primary single-pass annotation, the appropriate interpretation when annotation is partitioned among coders rather than duplicated [19].

F. Distant Supervision Annotation for the Silver Standard

Silver Standard samples were annotated automatically by matching each token against the LKB with a Longest Match algorithm scanning spans up to 4 n-grams [9], combined with a Heuristic-Augmented Denoising mechanism to reduce the label noise introduced by lexical ambiguity. Tokens or spans matching the Unambiguous LKB were assigned B-CRYPTO or I-CRYPTO directly. Tokens matching the Ambiguous LKB (for example the string GAS) were matched at the unigram level only and had to satisfy two further checks: the token must appear in full uppercase in the raw text, and at least one of 32 financial keywords must occur within four tokens on either side, a nine-token window including the candidate; if either check failed the token was assigned O. Tokens carrying a cashtag or hashtag prefix are treated as explicit entity markers and matched against both tiers without denoising, the only rejection being a purely numeric pattern such as $100. The tiers were constructed additively rather than as a partition, so 227 of the 286 Ambiguous surface forms are resolved by the first rule before the denoising constraints are reached; Section IV-E reports a controlled variant with mutually exclusive tiers.

G. Hybrid Model Architecture

The encoder backbone is IndoBERTweet, pre-trained on 409 million Indonesian tweet tokens [14]. A Conditional Random Field (CRF) layer is stacked on the encoder output to enforce structurally valid label sequences, preventing I-CRYPTO without a preceding B-CRYPTO [20]. Sequences are truncated at 128 word-piece tokens and words beyond that limit are padded with O, so any gold entity lost to truncation counts as a false negative. Training proceeds in two stages: base training runs the full model on the Silver Standard for a fixed budget of three epochs at a learning rate of 5e-5 with a batch size of 32, and hybrid fine-tuning then continues on the Gold Standard for a fixed budget of two epochs with the learning rate reduced to 2e-5 and the batch size to 16, slowing parameter updates to limit catastrophic forgetting of the representations acquired in the first stage. No checkpoint selection is performed at either stage: no evaluation set is passed to the trainer and the epoch budget is fixed in advance, so no fold reserved for testing can influence the model later scored on it.

H. Baselines and Ablation Design

Nine systems were trained and scored under a single protocol: four comparative baselines, four ablated variants and the proposed system. The baselines are a gazetteer applying the labeling function of Section III-F directly to the Gold Standard with no training, a BiLSTM-CRF with randomly initialized embeddings trained for 20 epochs at a learning rate of 1e-3, and two Transformer encoders, multilingual BERT and IndoBERT, sharing the architecture and fine-tuning budget of the proposed system but trained on the Gold Standard only.

The ablated variants each remove one component. IndoBERTweet-CRF trained on the Gold Standard alone removes distant supervision (−DS) and doubles as an encoder-matched baseline; the base checkpoint evaluated without fine-tuning removes the second training stage (−FT); a base model pre-trained on a Silver Standard built with denoising disabled, yielding 53,169 entities instead of 37,765, removes denoising (−HD); and a softmax head replaces the CRF (−CRF). Because the base checkpoint stores CRF transition parameters with no counterpart in a softmax model, the last variant is initialized from the plain encoder and the CRF contribution is measured against IndoBERTweet-CRF (gold only), which differs from it only in the output head.

I. Evaluation Protocol

Evaluation is conducted on the Gold Standard using 5-Fold Cross-Validation to minimize sampling bias. The fold assignment is generated once with a fixed random seed and reused by all nine systems, and predictions from each held-out fold are pooled into a single out-of-fold set that is the basis for every reported score and for the error analysis.

Performance is assessed at the entity level under the Exact Match criterion, which awards no partial credit: a prediction counts as correct only when both the boundary span and the entity type are correct [20], [21]. Precision, Recall and their harmonic mean, the F1-Score, are the primary measures [21], [22]. Because differences between closely matched systems can arise by chance, every comparison against the proposed system is tested for statistical significance [23], using both a paired bootstrap with 100,000 resamples and an approximate randomization test with 100,000 trials on the same predictions. The word significant is used in this paper only for differences clearing the 0.05 threshold under both tests after Holm correction within their family. 95% confidence intervals for F1 come from bootstrap resampling, and the proposed system and the −HD variant were repeated with three random seeds to separate the denoising effect from run-to-run variance.

IV. Result and Discussion

A. Dataset and Corpus Statistics

The raw collection of 60,283 tweets was reduced to 27,226 by four sequential filters: de-duplication removed 14,076 duplicates or verbatim reposts, global noise filtering 8,411, hybrid language filtering 9,961 and spam filtering 609. From this pool 1,500 tweets were drawn by Stratified Sampling to form the Gold Standard and the remaining 25,726 formed the Silver Standard; 39 tweets written entirely in a foreign language were excluded during annotation, leaving a final Gold Standard of 1,461. The dataset and LKB are available at DOI 10.17632/rtw4d5c6hj.

TABLE I. Silver and Gold Standard Composition

Property	Silver Standard	Gold Standard
Sentences	25,726	1,461
Tokens	850,987	50,242
Entities	37,765	1,182
Entities per sentence	1.468	0.809
Sentences without entity	7,536	847
Unique surface forms	1,123	298

The Silver Standard carries 1.468 entities per sentence against 0.809 in the Gold Standard, a direct consequence of distant supervision labeling every knowledge base match while human annotators reject names referring to a platform or a company rather than a traded asset. Entity density also explains why the Gold Standard is the harder target: 847 of its 1,461 sentences contain no entity, so over-prediction is penalized heavily. The label distribution is skewed and almost entirely single-token: B-CRYPTO accounts for 1,182 tokens (2.35%), I-CRYPTO for 15 (0.03%) and O for 49,045 (97.62%), with 1,167 of the 1,182 gold entities (98.7%) spanning a single token.

Heuristic-Augmented Denoising suppresses 15,404 of the 53,169 labels that naive longest matching would assign, a reduction of 28.97%, retaining 438 across the Ambiguous tier. The most frequently suppressed forms are ini (9,115), amp (2,033) and dia (1,375). The forms amp and gt (391) are undecoded HTML entities surviving from scraping rather than lexical collisions, so part of the effect is data hygiene rather than disambiguation. The form bnb is the opposite case: 154 of its 864 occurrences are retained as genuine references to a top-capitalization asset that a static blacklist would have deleted outright. Retention reaches 17.8% for bnb and 63.3% for sei while falling below 1% for function words such as ini and dia, and that asymmetry is the practical argument for a context-sensitive rule over a blacklist, since 130 of the 1,182 Gold Standard entities (11.0%) carry a surface form drawn from the Ambiguous tier.

TABLE II. Comparative Evaluation and Ablation on Pooled Out-of-Fold Predictions

System	Training data	P	R	F1	95% CI	ΔF1	p (Holm)
Gazetteer longest-match	LKB only	0.4691	0.8029	0.5922	[0.564, 0.619]	−0.2619	< 0.001
BiLSTM-CRF	gold	0.7887	0.6252	0.6975	[0.667, 0.726]	−0.1566	< 0.001
mBERT-CRF (gold only)	gold	0.8227	0.8325	0.8276	[0.805, 0.849]	−0.0265	0.028
IndoBERT-CRF (gold only)	gold	0.8561	0.8553	0.8557	[0.835, 0.875]	+0.0016	0.852
IndoBERTweet-CRF (gold only) [−DS]	gold	0.8454	0.7631	0.8021	[0.777, 0.826]	−0.0519	< 0.001
Base model, no fine-tuning [−FT]	silver	0.4672	0.8020	0.5905	[0.562, 0.617]	−0.2636	< 0.001
Proposed w/o denoising [−HD]	silver + gold	0.8191	0.8655	0.8416	[0.821, 0.861]	−0.0125	0.052
IndoBERTweet-softmax (gold only) [−CRF]	gold	0.8209	0.7910	0.8057	[0.782, 0.828]	−0.0484	< 0.001
Proposed (IndoBERTweet-CRF, DS+HD+FT)	**silver + gold**	**0.8328**	**0.8765**	**0.8541**	**[0.834, 0.873]**	–	–

All nine systems share identical folds and an identical evaluation protocol. Rows 2–5 and row 8 are restricted to Gold Standard supervision by design, which is what allows row 5 to serve as the −DS ablation; row 1 involves no training at all. ΔF1 is measured against the proposed system; p-values combine a paired bootstrap and an approximate randomization test, Holm-corrected across the eight comparisons in this family.

B. Comparative Evaluation and Ablation

Table II reports all nine systems on the pooled out-of-fold predictions, with per-fold standard deviations ranging from ±0.0106 to ±0.0461. Three of the four comparative baselines fall below the proposed pipeline: the gazetteer by 0.2619, BiLSTM-CRF by 0.1566 and mBERT-CRF by 0.0265, all significant under both tests after Holm correction ($p \leq 0.028$). The fourth, IndoBERT-CRF trained on the Gold Standard alone, is not distinguishable from it (+0.0016, $p = 0.852$), and that is worth stating plainly. As a raw encoder IndoBERTweet is the weaker of the two: under an identical protocol on identical data it reaches 0.8021 against

IndoBERT's 0.8557. Distant supervision and denoising close that gap. The claim the evidence supports is that the pipeline improves IndoBERTweet by 0.0519, not that the proposed architecture surpasses IndoBERT.

The Training data column makes the asymmetry explicit: five of the nine systems are restricted to Gold Standard supervision and a sixth uses no training data at all, which is what allows IndoBERTweet-CRF (gold only) to serve as the −DS ablation. The 25,726 additional Silver Standard sentences carry automatically generated labels rather than manual ones, so the comparison is between identical human annotation with and without automatic augmentation. Volume alone does not explain the outcome: the gazetteer draws on the same knowledge base and reaches only 0.5922.

TABLE III. COMPONENT CONTRIBUTIONS WITH PAIRED CONTROLS

Removed component	Control	Contribution	p raw	p (Holm)
Hybrid fine-tuning [−FT]	Proposed	+0.2636	< 0.001	< 0.001
Distant supervision [−DS]	Proposed	+0.0519	< 0.001	< 0.001
Heuristic denoising [−HD]	Proposed	+0.0125	0.026	0.052
CRF layer [−CRF]	IndoBERTweet -CRF (gold)	−0.0036	0.553	0.553

Table III isolates each component against its own control. Hybrid fine-tuning dominates at 0.2636, followed by distant supervision at 0.0519. Heuristic denoising contributes 0.0125; raw p-values from both tests sit near 0.025, but the value rises to 0.052 after Holm correction and therefore does not meet the threshold adopted here. Repeating the proposed system and the −HD variant with three random seeds gives mean F1-Scores of 0.8568 and 0.8454 (standard deviations 0.0035 and 0.0033); the difference is positive in all three seeds, averages 0.0113, and the lowest score with denoising, 0.8541, exceeds the highest without it, 0.8476. Denoising therefore has a consistent effect, small relative to the other components and not established at the 0.05 level by a single-run resampling test.

The CRF layer produces no measurable benefit on this schema, changing F1 by −0.0036 with p = 0.553 against a control identical except for the output head. Zero BIO transition violations were recorded across 50,242 tokens, but this holds for all nine systems, including the softmax variant, the BiLSTM baseline and the untrained gazetteer. With I-CRYPTO covering 0.03% of tokens and 98.7% of entities spanning a single token, there are almost no transitions for a CRF to model, so the absence of invalid sequences is a property of the annotation schema rather than evidence of the layer's contribution.

C. Error Analysis

The proposed system recovers 1,036 of the 1,182 Gold Standard entities under exact match. Its 354 errors comprise 203 false positives, 142 false negatives and 9 boundary errors in which a span was detected but delimited incorrectly. Because the schema contains a single entity type, a breakdown by entity category is not available; Table IV instead decomposes recall along four dimensions the schema does support. Stratum membership was not persisted with the annotated corpus and was re-derived post hoc from LKB tier membership, giving 387 Blue-Chip, 754 Altcoin, 313 Background and 7 unclassified tweets, so the stratum rows describe that re-derivation rather than the original sampling assignment.

TABLE IV. RECALL BREAKDOWN ON OUT-OF-FOLD PREDICTIONS

Dimension	Category	Recall
LKB membership	In LKB (n = 956)	0.9550
	Not in LKB (n = 226)	0.5442
Span length	1 token (n = 1,167)	0.8817
	2 tokens (n = 15)	0.4667
Surface form	Cashtag or hashtag	0.9311
	Uppercase ticker	0.8731
	Lowercase ticker	0.8594
	Full coin name	0.8482
	Multi-token name	0.4667
Stratum	Blue-Chip	0.9084
	Background	0.8116
	Altcoin	0.7832

The strongest signal is knowledge base membership: entities whose surface form appears in the LKB are recovered at 0.9550 against 0.5442 for those outside it, so distant supervision transfers the contents of the knowledge base into the model effectively but does not generalize far beyond it. The ceiling this implies is concrete: the LKB covers 956 of the 1,182 Gold Standard entities (80.9%), and a naive gazetteer with denoising disabled reaches a recall of 0.8080, exhausting what lexicon matching alone can achieve, so the 123 out-of-vocabulary entities the full pipeline recovers are what no gazetteer variant can produce.

Span length is the second discriminator, 0.8817 for single-token entities against 0.4667 for two-token entities, though the latter rests on 15 instances, and multi-token names trail every other surface-form category. Subword fragmentation explains this: entity surface forms fragment into 2.45 word pieces on average under IndoBERTweet against 1.20 for the corpus as a whole.

Manual inspection of a random sample of 50 of the 203 false positives identifies asset–platform metonymy as the largest single category: tokens such as Solana, Ripple and Axie are correctly labeled O when the tweet discusses the network, the company or the game rather than the traded asset, and the model has no discourse-level signal for that distinction. Cross-domain lexical ambiguity accounts for a comparable share; in one sampled tweet about football, eth abbreviates a manager's name but was predicted B-CRYPTO, because across the Silver Standard eth appears almost exclusively in cryptocurrency contexts. A third group consists of hashtagged community terms such as #degen and #memecoin and a fourth of non-crypto tickers such as $COIN and $AAPL; both follow from the cashtag rule in Section III-F, since 32 of the 203 false positives (15.8%) carry a $ or # prefix and bypassed denoising entirely, which is why #memecoin survives while the bare form memecoin was suppressed 313 times.

The same inspection found 9 of the 50 sampled false positives (18%) to be asset mentions not annotated in the Gold Standard, with a further 10% doubtful, so reported precision is a conservative lower bound. The Gold Standard was left un-adjudicated, since revising it after observing model predictions would invalidate every score in Table II.

D. Comparison with Prior Work

Direct numerical comparison with published Indonesian NER results is not meaningful, because the corpora, entity types and register differ in every case, but the figures give

context. Khairunnisa et al. [13] reach 94.90% on a formal news corpus with clean orthography, Wilie et al. [8] report 79.25% on the general-domain NERP dataset, and Wang [17] reports 91.3% for FinBERT on English financial documents. The closest comparison by register is Koto et al. [14] at 85.4% on an informal Indonesian NER benchmark; the 0.8541 obtained here is of the same order on a narrower domain with a single entity type, a heavy out-of-vocabulary load and no manually annotated training data beyond 1,461 tweets. Methodologically the closest precedent is ANEA [9]; the results here agree with it in direction and quantify how much post-hoc filtering contributes once a manually annotated fine-tuning stage is present: 0.0125 against 0.2636 for fine-tuning itself.

E. Limitations

Asset–platform metonymy is not addressable by any knowledge base, because the ambiguity is discourse-level rather than lexical. The recall gap between in-knowledge-base entities (0.9550) and out-of-vocabulary entities (0.5442) is the principal constraint on generalization: distant supervision propagates the contents of the LKB into the model but does not teach it what a coin name looks like in general, so coverage places a practical ceiling on recall for newly listed assets.

The tiers were constructed additively, so 227 of the 286 Ambiguous surface forms are shadowed by the Unambiguous tier and never reach the denoising constraints. A controlled variant with mutually exclusive tiers was therefore evaluated: removing the 122 single-token surface forms shared between the tiers raises suppression from 28.97% to 56.94% and lifts the training-free gazetteer from 0.5922 to 0.7852, yet the effect on the final system is negligible, F1 moving to 0.8445 ($p = 0.16$) with the measured contribution of denoising falling to 0.0029 ($p = 0.67$). Silver label quality is therefore decisive before human supervision and largely erased by it, since 1,461 annotated tweets absorb a difference of nearly 15,000 labels. The contribution in Table III holds at the 28.97% operating point and should not be assumed invariant to it.

Tokens prefixed with $ or # bypass denoising by design, on the assumption that an author using a cashtag intends an asset reference; that assumption fails for equity tickers and hashtagged jargon and accounts for 15.8% of all false positives. Finally, primary annotation was single-pass, so the agreement figures in Section III-E describe a 150-tweet overlap sample rather than the corpus as a whole, and part of the measured denoising effect removes undecoded HTML entities rather than linguistic ambiguity.

V. Conclusions

This study implemented Named Entity Recognition for cryptocurrency assets on noisy, informal Indonesian-language text from Twitter (X), using an IndoBERTweet-CRF architecture trained on two complementary label sources: a Distant Supervision Silver Standard and a manually annotated Gold Standard. Under 5-Fold Cross-Validation the proposed method achieved a mean Precision of 83.28%, a Recall of 87.65% and an F1-Score of 85.41%, indistinguishable from an IndoBERT-CRF model trained on the Gold Standard alone ($p = 0.852$) and ahead of mBERT-CRF, BiLSTM-CRF and the gazetteer baseline at $p \leq 0.028$ under both tests after Holm correction.

Controlled ablation locates the contributions unevenly. Hybrid fine-tuning accounts for 0.2636 of the final score and distant supervision for 0.0519, both at $p < 0.001$ after Holm correction. Heuristic denoising contributes 0.0125, consistent in direction across three random seeds but above the 0.05 threshold after correction, and its magnitude depends on how aggressively the Ambiguous tier is applied. The CRF layer is not distinguishable from a softmax head: with 98.7% of entities spanning a single token, the absence of invalid label transitions reflects the annotation schema rather than the layer. Distant Supervision supplied 37,765 entity labels across 25,726 sentences without manual annotation and confined human annotation to 1,461 tweets; annotation effort was not timed, so this is reported as a difference in label counts and not as a measured saving in time or cost.

Two limitations dominate the residual error: asset–platform metonymy, where a coin name refers to the network or the company behind it, and a recall drop from 0.9550 to 0.5442 for entities absent from the knowledge base. Dynamic updates to the Local Knowledge Base would capture newly listed low-capitalization tokens and raise the coverage ceiling that currently bounds recall, though this requires a premium CoinGecko API subscription. Replacing the rule-based labeling mechanism with generative Large Language Models as pre-annotation agents is a further direction for suppressing false positives at the source, particularly for metonymy cases that no lexicon can resolve. The parity observed against IndoBERT also suggests that the same pipeline applied to a general-domain encoder with a larger pre-training corpus may yield a stronger system than the one reported here.

Author Contributorship

Jonathan Davin: conceptualization, methodology, software, formal analysis, investigation, resources, data curation, writing — original draft, visualization, project administration. Shane Anthony and Bryan Dale: conceptualization, resources, data curation, writing — review and editing. Clark Sompie: conceptualization, visualization. Henry Lucky and Rifqi Charisma: supervision.

References

[1] Otoritas Jasa Keuangan (OJK), "Laporan Perkembangan Jumlah Konsumen Aset Kripto Resmi Periode Februari-November 2025," Jakarta, 2025.

[2] K. J. Merkley, J. Pacelli, and M. Piorkowski, "Crypto-influencers," Rev. Account. Stud., vol. 29, pp. 2254–2297, July 2024

[3] R. I. Affandi, "Register Analysis of Cryptocurrency Community on Social Media Twitter: Sociolinguistics Perspective," Thesis, Univ. Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia, Aug. 2023.

[4] E. Lansiaux, N. Tchagaspanian, and J. Forget, "Community Impact on a Cryptocurrency: Twitter Comparison Example Between Dogecoin and Litecoin," Front. Blockchain, vol. 5, p. 829865, Apr. 2022.

[5] E. Fletcher, "Reports show scammers cashing in on crypto craze," Federal Trade Commission Consumer Protection Data Spotlight, Jun. 2022. [Online]. Available: https://www.ftc.gov

[6] H. Jiang, Y. Hua, D. Beeferman, and D. Roy, "Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis," in Proc. 13th Lang. Resources and Evaluation Conf., Marseille, France, Jun. 2022, pp. 7199–7208.

[7] E. Dave and A. Chowanda, "IPerFEX-2023: Indonesian personal financial entity extraction using indoBERT-BiGRU-CRF model," J. Big Data, vol. 11, no. 1, p. 139, Sept. 2024.

[8] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in Proc. 1st Conf. Asia-Pacific Chapter ACL and 10th IJCNLP, Suzhou, China, Dec. 2020, pp. 843–857.

[9] M. A. Hedderich, L. Lange, and D. Klakow, "ANEA: Distant Supervision for Low-Resource Named Entity Recognition," in Proc. Workshop Practical Mach. Learn. Developing Countries (PML4DC), ICLR, Apr. 2021.

[10] Y. Meng et al., "Distantly-Supervised Named Entity Recognition with Noise-Robust Learning," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Nov. 2021.

[11] H. Wang, Y. Dong, R. Xiao, F. Huang, G. Chen, and J. Zhao, "Debiased and Denoised Entity Recognition from Distant Supervision," in Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS), Nov. 2023.

[12] I. Budi and R. R. Suryono, "Application of named entity recognition method for Indonesian datasets: a review," Bull. Electr. Eng. Inform., vol. 12, no. 2, pp. 969–978, Apr. 2023

[13] S. O. Khairunnisa, Z. Chen, and M. Komachi, "Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in the Indonesian Language," ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 22, no. 6, art. no. 184, Jun. 2023

[14] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Nov. 2021, pp. 10660–10668

[15] R. M. Yanti, I. Santoso, and L. H. Suadaa, "Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study : Power Failure in the Special Region of Yogyakarta)", Indonesian J. of Inf. Syst., vol. 4, no. 1, pp. 76–86, Aug. 2021.

[16] D. Alfatah, "Application Of Transformer Model For Sentiment Detection On Indonesian Twitter Data", J. Komput., vol. 2, no. 2, pp. 67–, Jun. 2024.

[17] Z. Wang, "Enhancing Financial Named Entity Recognition through Adaptive Few-Shot Learning: A Comparative Study of Pre-trained Language Models", JACS, vol. 4, no. 7, pp. 13–25, Jul. 2024.

.

[18] X. Wei, L. Salsabil, and J. Wu, "Theory entity extraction for social and behavioral sciences papers using distant supervision," in Proc. 22nd ACM Symp. Doc. Eng. (DocEng), Nov. 2022, art. no. 7, pp. 1–4.

[19] R. Artstein and M. Poesio, "Survey Article: Inter-Coder Agreement for Computational Linguistics," Comput. Linguist., vol. 34, no. 4, pp. 555–596, Dec. 2008.

[20] D. Ignasius, I. N. Dewi, M. B. C. Norman, and R. R. Sani, "Medical Named Entity Recognition from Indonesian Health-News using BiLSTM-CRF with Static and Contextual Embeddings", JAIC, vol. 9, no. 6, pp. 2974–2985, Dec. 2025.

[21] T. Chavan and S. Patil, "Named entity recognition (NER) for news articles," Int. J. Artif. Intell. Res. Develop. (IJAIRD), vol. 2, no. 1, pp. 103–112, Jan.–Jun. 2024.

[22] Warto, "Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application", Stat., optim. inf. comput., vol. 12, no. 4, pp. 907-942, Feb. 2024

[23] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," in Proc. 56th Annu. Meeting Assoc. Comput. Linguist. (ACL), Jul. 2018, pp. 1383–1392.