Temporal Simultaneity Predicts Annotation Quality in Sentiment Corpora

Anonymous ACL submission

Abstract

Annotation quality is difficult to sustain when campaigns span weeks or months with small annotator pools. We present a Setswana sentiment dataset of 3,565 tweets annotated by three native-speaker annotators across eight batches and examine why inter-annotator agreement (IAA) declines over time. Despite an aggregate Randolph's free-marginal Kappa of $\kappa = 0.76$, "excellent," per batch κ falls by more than 32 points across the annotation task. Through six targeted analyses, we find that (i) label confusion concentrates on the negative/neutral boundary, (ii) two annotators show run-length drift consistent with *autopilot* labeling, and (iii) the dominant predictor of κ is *temporal simultaneity*: tweets labeled within one minute achieve $\kappa = 0.98$, while those labeled more than a day apart reach only $\kappa = 0.65$. Annotation speed and tweet-level linguistic features show no meaningful association with κ. We benchmark three open multilingual encoders and proprietary models (GPT-5 and Gemini) on three-class sentiment classification; fine-tuning yields gains of 29 to 43 macro-F1 points over pretrained baselines, with GPT-5 few-shot leading overall (62.2 macro-F1). We release the dataset, per-annotation timestamps, and analysis code to support reproducible quality auditing for future African language NLP resources.

1 Introduction

The rapid growth of African-language NLP has produced a wave of labeled resources covering tasks from named-entity recognition to sentiment and emotion analysis (Muhammad et al., 2023a; Adelani et al., 2022). These datasets are often created with small volunteer or student annotator pools working over extended periods. These conditions are known to degrade annotation quality in high-resource settings (Snow et al., 2008; Søgaard et al., 2014) and have rarely been studied for low-resource African languages.

A central challenge in low-resource dataset development is the trade-off between scale and quality. A small, carefully annotated corpus is generally more useful for training and evaluating downstream models than a larger, noisier one (Northcutt et al., 2021; Aroyo and Welty, 2015). Understanding the causes of quality erosion is therefore not a matter of methodological interest alone; it should directly inform annotation campaign design and the resulting dataset curation.

We present a new Setswana sentiment dataset and a systematic analysis of quality decline during the annotation campaign. The dataset comprises 3,565 tweets annotated by three native speakers across seven batches. Our annotation logs record the precise timestamp and other metadata for every annotation event, enabling analyses that go beyond the standard practice of reporting aggregate inter-annotator agreement (IAA). We investigate six complementary hypotheses about what causes inter-annotator disagreement:

RQ1 Does IAA decline monotonically across batches?

RQ2 Which label pairs are most frequently confused, and does the confusion pattern shift over time?

RQ3 As the campaign progresses, do annotators exhibit *autopilot* behavior: increasingly long runs of the same label?

RQ4 Is faster annotation (shorter inter-annotation gap) associated with lower IAA at the tweet or batch level?

RQ5 Does the time elapsed between the first and last annotator's timestamp for the same tweet (***temporal simultaneity***) predict IAA?

RQ6 Do tweet-level linguistic features (e.g., token count) explain variation in IAA?

Contributions

- A manually annotated Setswana Twitter sen-

timent dataset of 3,565 tweets with full per-annotation timestamps (released publicly).

- A characterization of κ decline over a multi-batch campaign, identifying temporal simultaneity as the dominant structural predictor.
- Null results for annotation speed and tweet-level linguistic complexity as predictors of κ, narrowing the search space for future mitigation strategies.
- Benchmarks on the dataset using various encoder and decoder models with concrete recommendations for annotation campaign design that require no additional budget.

2 Related Work

2.1 African-language sentiment and classification resources

Muhammad et al. (2023a) introduced AfriSenti, a multilingual Twitter sentiment benchmark covering 14 African languages, providing baselines for few-shot and fine-tuned models. A shared task based on this data was later organized at SemEval-2023 (Muhammad et al., 2023b). Earlier efforts include NaijaSenti (Muhammad et al., 2022), covering Hausa, Igbo, Nigerian Pidgin, and Yorùbá; the Yorùbá Sentiment Corpus for movie reviews (Shode et al., 2022); and SAfriSenti (Mabokela and Schlippe, 2022), a South African sentiment corpus covering Setswana and related languages. Myoya et al. (2024) analyzed commuter sentiment across multilingual public transport contexts in Kenya, Tanzania, and South Africa.

Beyond sentence- or document-level polarity, recent work has extended sentiment to aspect-level and dimensional representations. Lee et al. (2026) released DimABSA, a multilingual multidomain dataset for aspect-based sentiment analysis that replaces categorical labels with real-valued affective dimensions, enabling more nuanced opinion modeling. Becker et al. (2026) similarly introduced DimStance, which models stance along dimensional valence and arousal axes across multiple languages.

More recent work has expanded from coarse sentiment polarity to include multi-label emotion classification. Muhammad et al. (2025) released BRIGHTER, a human-annotated emotion recognition collection spanning 28 languages with emphasis on under-represented African languages; Belay et al. (2025) introduced Ethio-Emo, a multi-label emotion dataset for four Ethiopian languages; Adelani et al. (2023) released MasakhaNEWS, a topic classification benchmark across 16 African languages.

Across these efforts, IAA is typically reported as a single aggregate figure, without analysis of annotation quality trends or behavioral patterns.

2.2 Annotation quality and IAA

Inter-annotator agreement is most commonly measured by Cohen's (Cohen, 1960) or Fleiss' (Fleiss, 1971) κ, or Krippendorff's α (Krippendorff, 2011). For equal-frequency label distributions, Randolph (2005) proposed a free-marginal multi-rater Kappa that does not penalize annotators for collectively preferring one category, a property important for naturally skewed corpora such as sentiment datasets. We adopt Randolph's κ throughout. By convention (Randolph, 2008), $\kappa < 0.40$ is "poor," $0.40 \leq \kappa < 0.75$ is "intermediate to good," and $\kappa \geq 0.75$ is "excellent." Klie et al. (2024) analyzes annotation quality practices in over 100 NLP dataset papers, and it is found that quality assurance is routinely under-reported.

2.3 Annotation fatigue and drift

Snow et al. (2008) and subsequent crowdsourcing research (Kittur et al., 2008; Kazai et al., 2012) document that annotator performance degrades with session length and repetition. Geva et al. (2019) show that NLU datasets harbor annotator-specific biases that accumulate over time. In their work, Muhammad et al. (2022) reported a visible decline in inter-annotator agreement across batches for Hausa, Igbo, and Yorùbá in the NaijaSenti corpus but did not investigate the causes of the decline. In the African-language context, such longitudinal analyzes are almost entirely absent, which is the primary motivation for the approach we take here.

2.4 Disagreement as signal

A growing body of work argues that annotator disagreement should not be reflexively treated as noise to be resolved but as a meaningful signal encoding genuine label ambiguity, subjective interpretation, and demographic perspective (Fleisig et al., 2024; Plank, 2022; Uma et al., 2021; Xu and Jurgens, 2026). We subscribe to this view, while also recognizing that not all disagreement is alike: some reflects genuine ambiguity in the data, while some reflects annotation process artifacts such as fatigue or scheduling. Distinguishing between the two is a goal in this work.

3 The Setswana Sentiment Dataset

3.1 Setswana Language

Setswana (ISO 639-3: `tsn`) is a Bantu language spoken by approximately 8 million people, predominantly in South Africa and Botswana, where it both holds official status. Based on the South African National Census of 2022, it is the fifth most-spoken home language in South Africa (Statistics South Africa, 2025). Like other Bantu languages, Setswana is agglutinative with extensive noun-class morphology, making tokenization and lexical-semantic analysis non-trivial. Despite its speaker base, Setswana remains severely under-resourced in NLP: pre-trained language models and labeled corpora are scarce, and the language is absent from most mainstream benchmarks.

3.2 Data Collection

Collection method Tweets were collected from the public Twitter API (v2) using keyword and location filters targeting Setswana content between 2021 and 2022. A broad query was applied to maximize recall, with subsequent filtering applied.

Language identification All collected tweets were passed through the AfroLID language identification model (Adebara et al., 2022), which assigns confidence scores to the top-3 candidate languages. We retained tweets for which Setswana (`tsn`) appeared among the top-2 predictions. In the final annotated sample, 70.3% of tweets have Setswana as the top-1 prediction (mean LID score 0.831) and 29.7% have Setswana as the second-best prediction (mean competing language score 0.756, mean Setswana residual score 0.134). The competing language is predominantly Northern Sotho/Sepedi (`nso`), which shares significant lexical overlap with Setswana.

Anonymization All usernames, mentions, URLs, and other PII information were replaced with placeholder tokens before annotation to protect user privacy.

3.3 Annotation

Label schema Each tweet was assigned one of five labels: **Positive**, **Negative**, **Neutral**, **Mixed** (tweet expresses both positive and negative sentiment), or **Indeterminate** (tweet is unintelligible, in an unknown language, or cannot be confidently categorized).

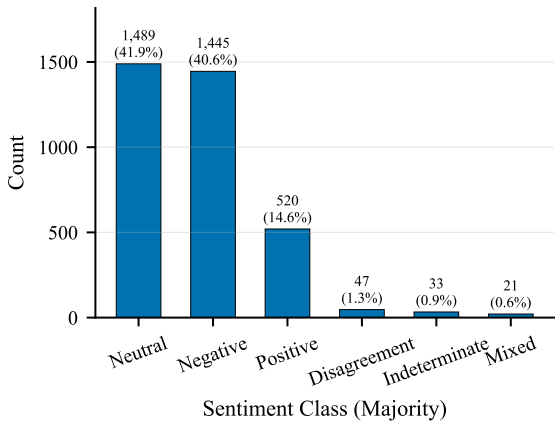


Figure 1: Sentiment label distribution after adjudication.

Annotators. Three undergraduate students of African languages annotated the dataset. All three are native speakers of Setswana. Before beginning the proper annotation task, the annotators completed a two-round training phase on 65 tweets, with adjudication and discussion of disagreements between rounds. The annotators are referred to as Ann. A, Ann. B, and Ann. C throughout.

Annotation tool and batching Annotations were collected through the LightTag annotation tool (Perry, 2021), now part of the Primer.ai NLP platform. The platform recorded the precise UTC timestamp of each label submission and other metadata such as annotator identity, per-example view logs (recording which annotators were assigned each item), a review and adjudication slot, and a free-text comment field (unused in our campaign). Production annotation was divided into seven batches of 500 tweets each (Batches 1–7), completed sequentially. Each annotator labeled all tweets in a batch independently and asynchronously, without access to the labels assigned by the other two annotators.

3.4 Dataset Statistics

The full corpus contains 3,565 annotated tweets. For tweets with unanimous three-way agreement, the label is taken directly; for 2-of-3 majority agreement, the majority label is retained; tweets with full three-way label conflict are excluded from downstream classification. The final released dataset contains 520 Positive, 1,445 Negative, and 1,489 Neutral tweets. Mixed, Indeterminate, and fully-conflicted labels together account for less than 3% of all examples and are not included in the classification split. Figure 1 shows the overall label distribution.

4 Annotation Quality Analysis

We conduct six targeted analyzes using the full annotation logs, including per-annotator timestamps for every label submission. All analyzes are restricted to tweets for which all three annotators provided a label (3,555 of 3,565 tweets; 10 edge-case tweets at batch boundaries are excluded).

4.1 RQ1: κ Over Batches

Figure 2 shows both the per-batch κ and the cumulative κ as each batch is added. Per-batch κ starts high ($\kappa = 92.17$ for Batch 1) and declines to $\kappa = 60.09$ by Batch 7, a fall of more than 32 points. The sharpest single drop occurs between Batch 4 ($\kappa = 82.13$) and Batch 5 ($\kappa = 64.29$), a 17.8-point decline in a single batch. The cumulative κ ends at 75.66, which is technically "excellent" by Randolph's thresholds but masks the severity of the deterioration in the final three batches. We investigate this decline from five angles in the sections that follow.

4.2 RQ2: Label Confusion

To understand which distinctions annotators find hardest, we construct a pairwise confusion matrix, counting, for each annotator pair on each tweet, how often two annotators assigned different labels. Figure 3 shows this matrix aggregated over all batches and split into early (Training–Batch 3) and late (Batch 4–7) periods.

The dominant confusion is overwhelmingly **Negative vs. Neutral** (1,448 pairwise cases), an order of magnitude larger than the next most frequent pair, Positive vs. Neutral (420 cases). This pattern is linguistically interpretable: Setswana political and social commentary frequently uses indirect or ironic phrasing where the boundary between a statement of fact (Neutral) and a negatively-valenced judgement (Negative) is genuinely ambiguous. This ambiguity is a property of the language and content domain, not solely of annotator quality. This motivates retaining these instances rather than discarding them.

Comparing the early and late heatmaps, the Negative/Neutral cell darkens in later batches while confusions involving Mixed and Indeterminate remain negligible throughout. This is consistent with annotators applying less careful distinctions as the campaign progresses, rather than simply encountering harder tweets.

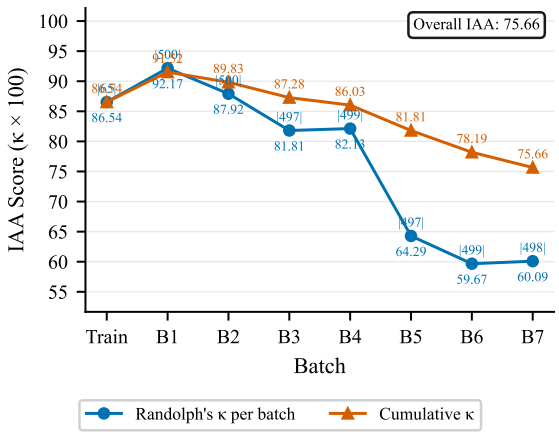


Figure 2: Per-batch Randolph's κ (blue) and cumulative κ (orange) across the training batch and Batches 1–7. Batch sizes are shown as $|n|$ above each point. κ declines monotonically from Batch 2 onward; the steepest drop is between Batch 4 and Batch 5.

4.3 RQ3: Autopilot Behavior

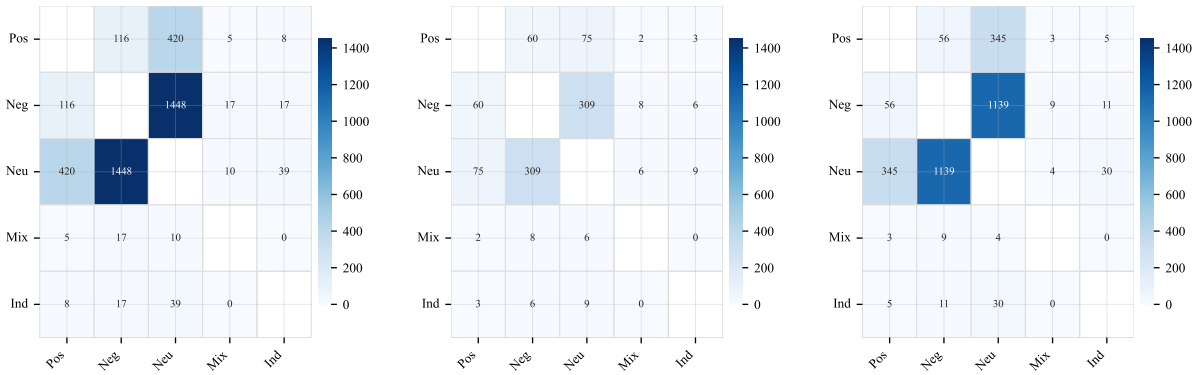
A common symptom of annotation fatigue is ***autopilot*** behavior, where the annotator repeatedly applies the same label without deliberation, producing long uninterrupted runs of identical labels in timestamp order. We calculated this as the "**`mean same-label run length`**" per annotator per batch, an uninterrupted sequence of identical consecutive labels in annotation submission order.

Figure 4 shows that mean run length increases for all three annotators across batches, starting near 1.4–1.5 labels per run in the training batch and rising to 1.7–1.9 by Batch 6 to 7. Mean run length increases consistently across batches for Ann. B and Ann. C, reaching its highest values in Batches 6 and 7, while Ann. A shows a weaker and less consistent trend. The upward pattern for two of the three annotators indicates that the labelling pattern, not just the labels themselves, changes over the annotation process.

Longer runs of length ≥ 5 provide a conservative indicator of non-deliberate labelling (Figure 10 in Appendix B). The mean count of such runs per annotator grows from Batch 1 to Batch 7, with the increase concentrated in Ann. B and Ann. C from Batch 5 onward. These are the same batches where κ falls most steeply.

4.4 RQ4: Annotation Speed

We compute the inter-annotation gap for each annotator as the elapsed time in seconds between consecutive label submissions in timestamp order. Figure 5 compares the gap distribution across three



(a) All batches (b) Early (Train–B3) (c) Late (B4–B7)

Figure 3: Pairwise annotator label confusion heatmaps. Each cell (i, j) counts annotator pairs where one chose label i and the other chose label j; diagonal cells are masked. All three heatmaps share the same colour scale. The Negative/Neutral cell (1,448 cases) dominates across all periods and intensifies in the late batches.

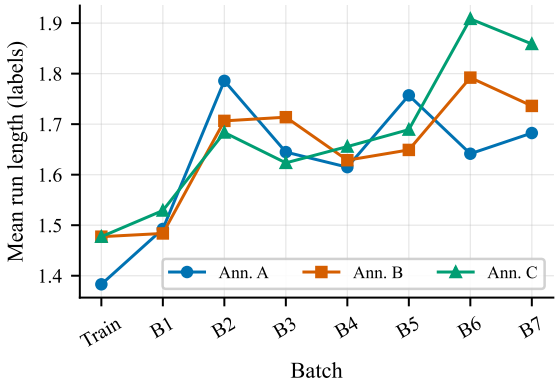


Figure 4: Mean same-label run length per annotator per batch. All three annotators trend upward; the increase is most pronounced and consistent for Ann. B and Ann. C, particularly from Batch 5 onward.

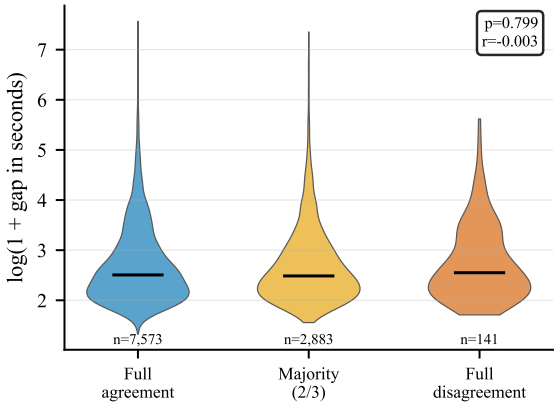


Figure 5: Distribution of log-transformed inter-annotation gaps (seconds) by tweet-level agreement category. Medians (horizontal bars) are virtually identical across all three groups, indicating no relationship between annotation speed and agreement.

agreement categories at the tweet level. The gap distributions of fully-agreed, majority-agreed, and fully-disagreed tweets are nearly identical, with near-overlapping medians and shapes (Figure 5).

At the batch level (Figure 6), median annotation speed does increase over the annotation process. Annotators label tweets progressively faster, with median gap falling from approximately 20s in the training batch to under 9s by Batch 6. Critically, this acceleration is decoupled from κ: the largest speed-up occurs between Batch 1 and Batch 2, while κ remains stable in that period; conversely, the sharpest κ drop (Batch 4→5) occurs after speed has already plateaued. Annotation speed and annotation quality therefore evolve independently.

4.5 RQ5: Temporal Simultaneity

Because annotators worked asynchronously, the same tweet might be reviewed by one annotator on one day and another days or weeks later. We define the ***time span*** for a tweet as the elapsed time between the first and last annotator's timestamp. Figure 7 shows the κ values after grouping tweets into four bins: $<$ 1 min, 1–60 min, 1 hr–1 day, and $>$ 1 day. The gradient is very visible:

- $<$ 1 min ($n = 901$): $\kappa = 0.98$
- 1–60 min ($n = 120$): $\kappa = 0.80$
- 1 hr–1 day ($n = 386$): $\kappa = 0.80$
- $>$ 1 day ($n = 2{,}148$): $\kappa = 0.65$

Tweets labeled within a single minute achieve near-ceiling agreement; tweets labeled more than a day apart lose 33 points of κ relative to the synchronous group. The majority of tweets in the dataset (60.3%) fall in the $>$ 1 day bin, which means the annotation's asynchronous structure structurally suppresses the overall κ. At the batch level (Figure 8), there is a strong negative association between mean inter-annotator time span

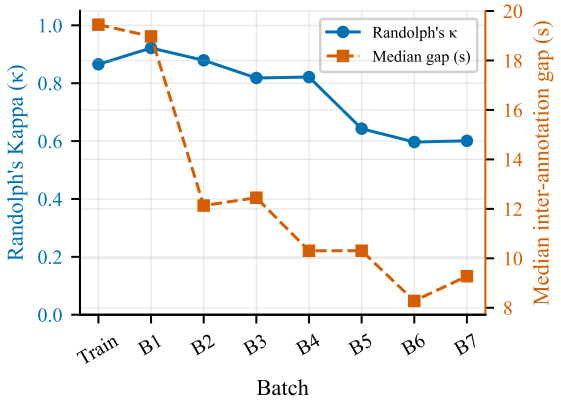


Figure 6: Randolph's κ (left axis, blue) and median inter-annotation gap in seconds (right axis, orange dashed) per batch. The two series decouple: speed drops sharply at Batch 2 while κ remains stable; κ falls steeply at Batch 5 while speed has already plateaued.

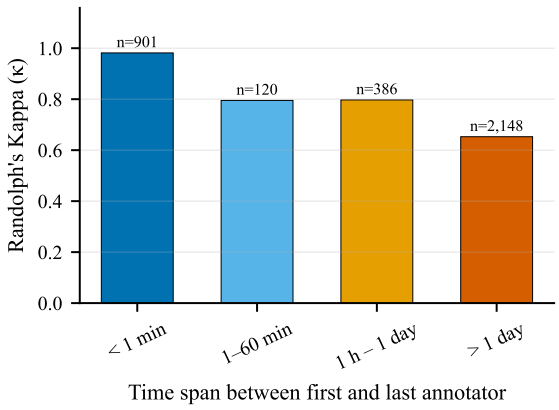


Figure 7: Randolph's κ by time-span bin (elapsed time between the first and last annotator's timestamp for the same tweet). Synchronous labelling (< 1 min, $n = 901$) yields near-perfect $\kappa = 0.98$; asynchronous labelling (> 1 day, $n = 2{,}148$) yields $\kappa = 0.65$.

and κ. As visible in the scatter, later batches were annotated with substantially greater time spread: Batch 7 has a mean span of over 20 days, compared to less than one day for Batches 1–2, and κ tracks this spread closely.

Cross-referencing annotation dates with the university academic calendar provides a plausible structural explanation for this pattern. Batches 1–3 (14–28 March) fell during normal lectures; Batch 5 (10–18 April) coincided with the end of Quarter 1 on 17 April, when assignment deadlines and tests are concentrated; and Batch 7 (24 April–24 May) extended into end-of-semester preparation, with two annotators completing their work on 24 May, the day before the university's cooling-off period for non-academic activities began (25 May). Academic pressure thus provides an external account of why annotators spread their work across progressively more days in later batches, increasing inter-annotator time spans and depressing κ.

4.6 RQ6: Tweet Complexity

We examine two tweet-level proxies for annotation difficulty: language identification (LID) rank and token count.

LID rank. Tweets where Setswana is the top-1 LID prediction ("lang1", $n = 2{,}497$, mean score $= 0.831$) and those where another language outscores Setswana ("lang2", $n = 1{,}058$, mean competing score $= 0.756$, mean Setswana score $= 0.134$) yield essentially identical κ: 0.757 and 0.757 respectively (Figure 9a). The agreement rates for the two groups are virtually identical, indicating no association between LID rank and annotator agreement. Per-batch trends are parallel for both groups (Figure 11 in Appendix C), confirming that LID ambiguity does not differentially affect agreement.

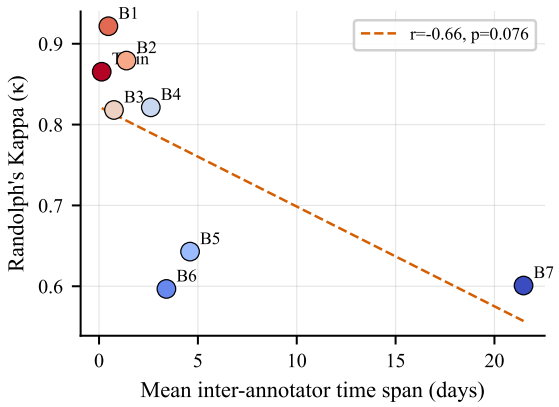
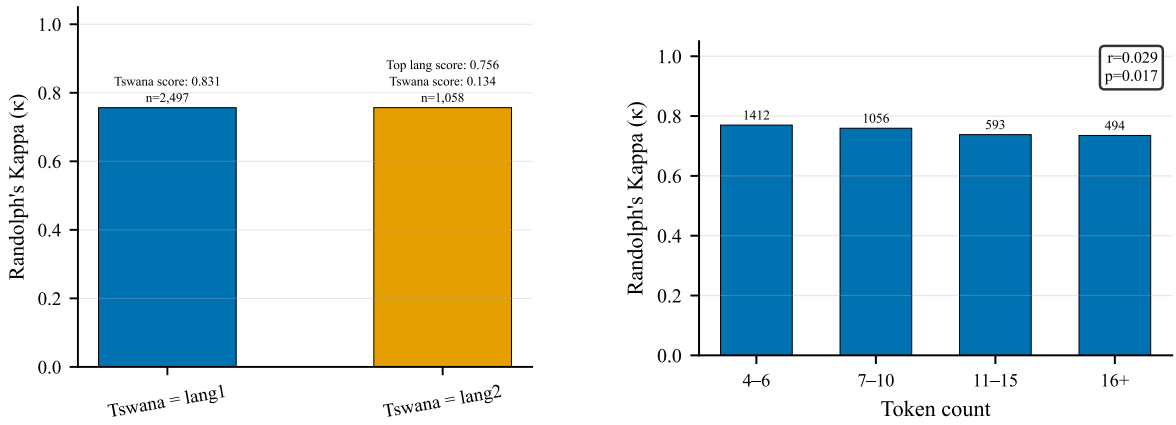


Figure 8: Batch-level mean inter-annotator time span (days) vs. Randolph's κ. The trend is consistent: batches with longer mean time spans have lower κ; Batch 7 is a notable outlier at over 20 days mean span.

Token count. Binning tweets by token count into four ranges (4–6, 7–10, 11–15, 16+) yields κ values of approximately 0.77, 0.76, 0.74, and 0.73, respectively, a near-flat profile with no consistent trend (Figure 9b). The difference across bins is negligible and tweet length accounts for virtually none of the observed variation in agreement.

Together, the LID and token-count null results establish that κ is not primarily driven by tweet difficulty. This shifts explanatory responsibility toward annotator-level factors such as fatigue, drift,



(a) Top-1 (lang1) vs. second-best (lang2) LID predictions.

(b) Agreement by token-count bin.

Figure 9: Annotation agreement (Randolph's κ) is stable across conditions: (a) no meaningful difference between top-1 vs. second-best LID predictions ($\kappa \approx 0.757$ for both groups); (b) minimal variation across token-count bins (κ ranges 0.73–0.77 with no consistent trend).

and scheduling. These are addressed by RQ3 and RQ5.

5 Discussion

5.1 Disagreement Is Not the Enemy

Before discussing individual findings, we emphasize a framing point. A growing body of work demonstrates that annotator disagreement encodes genuine ambiguity, diversity in perspectives, and domain-specific nuance that majority-vote aggregation actively discards (Plank, 2022; Fleisig et al., 2024; Xu and Jurgens, 2026). We take this seriously. The dominant confusion in our dataset, between Negative and Neutral (1,448 pairwise cases, RQ2), reflects a genuine property of Setswana political discourse, where indirect evaluative language straddles the boundary between factual statement and sentiment expression. This disagreement is linguistically meaningful, and we do not treat it as an error to be eliminated.

What we *do* target are two qualitatively different patterns that emerge in later batches: autopilot labelling (RQ3) and scheduling-driven temporal asynchrony (RQ5). These are process artefacts that inflate disagreement beyond whatever the data itself warrants, and they are detectable and correctable without discarding the inherent ambiguity that makes the task interesting. The key empirical observation is that the Negative/Neutral confusion is *stable* across batches (it was always the dominant pair), whereas the κ decline in Batches 5–7 is driven by a broad-based increase in confusion across *all* label pairs, consistent with process degradation rather than growing data ambiguity.

RQ	Factor	Finding
1	Batch trend	κ declines 92→60; sharpest at B4→B5
2	Label confusion	Neg/Neu dominates (1,448 cases); intensifies in late batches
3	Autopilot	Run length increases significantly for Ann. B & C
4	Speed	Not a predictor; gap distributions are virtually identical across groups
5	Simultaneity	Strongest predictor: $\Delta\kappa = 0.33$ from <1 min to >1 day
6	LID / tokens	No meaningful association; κ stable across both variables

Table 1: Summary of findings across the six research questions.

5.2 Interpretation

The temporal simultaneity finding (RQ5) is both the largest in magnitude and the most actionable. When all three annotators label a tweet within minutes, they are likely in a shared annotation context, with similar energy levels, similar interpretation of boundary cases, and no intervening exposure to other sessions that might subtly shift their decision criteria. The 33-point κ gap between synchronous (< 1 min) and asynchronous (> 1 day) tweets is far larger than any other factor we examined, and the batch-level pattern confirms the same gradient at a coarser granularity.

The run-length results (RQ3) provide a complementary, behaviorally grounded account. By Batches 5–7, Ann. B and Ann. C show statistically significant increases in mean run length and in the count of long streaks (≥ 5 labels), suggesting

that the decision process has shifted from deliberate per-tweet judgment to more automatic pattern-matching. The sharpest κ decline (Batch 4→5) coincides with the period where run-length growth is most pronounced. This points to a common underlying mechanism: reduced annotator engagement, not harder tweets.

The null result for annotation speed (RQ4) corrects a common intuition. The widely-held belief that "faster ⇒ worse" is not supported at the tweet level: gap distributions for agreed and conflicted tweets are virtually identical. The aggregate annotation-level co-occurrence of faster labelling and lower κ in later batches therefore reflects a shared cause (reduced engagement) rather than a direct causal chain.

The null results for LID ambiguity and token count (RQ6) are practically important. They establish that filtering tweets by linguistic complexity would not meaningfully improve κ, and that the process-driven quality issues are annotator-level rather than content-level. This means the same tweets, re-annotated under better scheduling conditions, would likely achieve substantially higher κ because the disagreement in the data is not the problem.

To improve annotation quality in future campaigns, we recommend implementing interventions based on our findings, as presented in Appendix F.

6 Classification Benchmark Results

Given the observed annotation quality patterns, we conduct a preliminary evaluation of the dataset's utility for sentiment classification. This serves as a benchmark for future work. The detailed experimental setup is presented in the Appendix E.

Results Table 2 reveals three key findings. First, pre-trained encoders without task-specific data perform near or below random (8.7–20.6 macro-F1), confirming that multilingual pre-training alone is insufficient for Setswana sentiment prediction, directly motivating the dataset. Second, fine-tuning on our data yields large, consistent gains across all three encoders (+29–43 macro-F1 points), with AfroXLMR-base achieving the best fine-tuned result (53.6), consistent with its broader African-language pre-training (Alabi et al., 2022). Third, GPT-5 in few-shot mode outperforms all fine-tuned encoders (62.2), suggesting that proprietary LLMs can leverage limited in-context examples effectively even for low-resource languages, though at significantly higher inference cost. Gemini 3, by contrast, shows little benefit from few-shot prompting (57.1 zero-shot vs. 57.2 few-shot), suggesting that in-context examples do not consistently help across frontier models for this low-resource language.

Model	Setting	Macro-F1	Accuracy
mBERT	pre-trained (probe)	20.6	24.2
AfriBERTa	pre-trained (probe)	8.7	15.0
AfroXLMR-base	pre-trained (probe)	10.6	16.1
mBERT	fine-tuned	50.6±0.9	55.5
AfriBERTa	fine-tuned	49.0±2.1	53.8
AfroXLMR-base	fine-tuned	**53.6**±1.2	56.7
GPT-5	zero-shot	56.8	57.6
Gemini	zero-shot	**57.1**	57.4
GPT-5	few-shot	**62.2**	62.5
Gemini	few-shot	57.2	57.4

Table 2: Three-class sentiment classification results on our test set (macro-F1 and accuracy, %). Pre-trained models are evaluated via MLM probing with no task-specific data. Fine-tuned models report mean ± std over 5 random seeds. Bold indicates best within each group.

7 Conclusion

We have presented a Setswana Twitter sentiment dataset and a systematic, timestamp-driven analysis of annotation quality across an eight-batch annotation process. The dominant predictor of inter-annotator agreement is ***temporal simultaneity***: tweets labeled within one minute achieve $\kappa = 0.98$; those labeled more than a day apart reach only $\kappa = 0.65$. Run-length drift in two annotators provides a complementary behavioral account of quality decline.

For low-resource NLP annotation tasks that rely on small pools working independently, precisely the conditions that maximize inter-annotator time spans, batch-scoped annotation windows, and run-length monitoring are low-cost interventions that could substantially improve dataset quality. We release the full dataset, timestamps, and analysis code to support reproducible quality auditing in future African-language NLP.

Limitations

Several limitations of this work should be noted. **Language scope:** The dataset covers Setswana as used on South African Twitter and may not generalize to other dialects or registers of the language. **Annotator pool:** Three annotators from a single institution participated in the campaign; a

more diverse pool might yield different agreement patterns. **Label schema:** The four-class schema (Positive, Negative, Neutral, Mixed) collapses nuanced sentiment; binary or fine-grained schemes could be explored in future work. **Platform bias:** Twitter data over-represents urban, younger, and digitally active speakers, which may not reflect the broader Setswana-speaking population. **Temporal coverage:** Tweets were collected during a specific period; sentiment distributions may shift with news cycles or social events. **LLM evaluation:** LLM results reflect a single evaluation run with fixed prompts; different prompt formulations may yield different results. The gap between GPT-5 and Gemini 3 few-shot performance may also reflect differences in instruction-following rather than underlying language competence.

Ethics Statement

Data source: All tweets were collected via the official Twitter/X Academic API under its terms of service for academic research. We do not release tweet IDs alongside raw text; instead, we release cleaned text only, in line with common practice for low-resource NLP datasets where data scarcity makes ID-only release impractical for reproducibility. **Annotator welfare:** Annotators were university students who participated voluntarily and were compensated at the official university hourly rate. No sensitive personal information was collected beyond the annotation labels. **Ethical clearance:** This project received ethical clearance from the University Research Ethics Committee before data collection and annotation. **Potential harms:** Sentiment classifiers trained on this data could be misused for surveillance or opinion mining of Setswana-speaking individuals. We release the dataset for research purposes and encourage users to consider these risks before deployment. **Broader impact:** Setswana is an under-resourced language; we hope this dataset lowers the barrier to NLP research for the language and contributes to more equitable representation in multilingual AI.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, and 46 others. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, and 26 others. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36:15–24.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Google. 2025. A new era of intelligence with Gemini 3.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2583–2586, Maui, Hawaii, USA. ACM.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456, Florence, Italy. ACM.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Klaus Krippendorff. 2011. *Computing Krippendorff's Alpha-Reliability*. University of Pennsylvania, Philadelphia, PA.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Ronny Mabokela and Tim Schlippe. 2022. A sentiment corpus for South African under-resourced languages in a multilingual context. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 70–77, Marseille, France. European Language Resources Association.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023a. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

Rozina L. Myoya, Vukosi Marivate, and Idris Abdulmumin. 2024. Analysing public transport user sentiment on low resource multilingual data. *Preprint*, arXiv:2412.06951.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. In *Journal of Artificial Intelligence Research*, volume 70, pages 1373–1411.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k[free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*, Joensuu, Finland.

Justus J. Randolph. 2008. Online kappa calculator. Computer software.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. YOSM: A new Yorùbá Sentiment Corpus for Movie Reviews. *AfricaNLP 2022 @ICLR*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 17 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Statistics South Africa. 2025. Cultural dynamics in south africa. Technical Report 03-01-84, Statistics South Africa, Pretoria, South Africa.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in NLP. *Preprint*, arXiv:2601.09065.

A Annotation Guidelines Summary

Annotators were provided with a written guideline document covering: (i) definitions of each sentiment category with six worked examples per category; (ii) decision rules for the Negative/Neutral boundary (factual statements about negative events are Neutral; explicit evaluative language is Negative); (iii) rules for code-switched content (label according to the dominant-language sentiment); and (iv) examples of Indeterminate tweets including non-Setswana content and unintelligible text.

B Long Run Counts (RQ3)

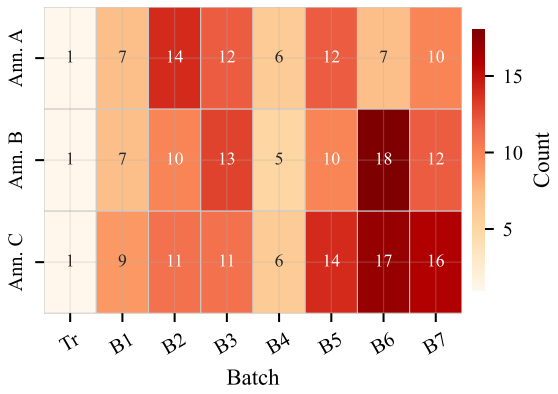


Figure 10: Count of runs of length ≥ 5 per annotator per batch. Darker shading indicates more long streaks of the same label. The increase is concentrated in Ann. B and Ann. C from Batch 5 onward, the same batches where κ falls most steeply.

C LID Null Result (RQ6)

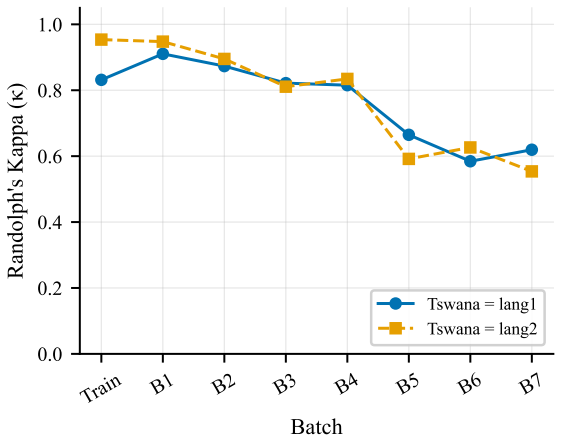


Figure 11: Per-batch κ for the lang1 and lang2 LID groups. The two lines track each other closely across all batches, confirming that LID ambiguity does not differentially drive quality decline.

D Cumulative Annotation Progress per Batch

Figure 12 plots cumulative annotations against wall-clock time, a distance-time graph where slope encodes speed and flat segments are rest periods. Per-session speed is labeled on each active segment; red arrows mark extreme idle gaps (above the 75th percentile for that panel, minimum 4 h). Later batches show increasing fragmentation into more and longer rest periods, consistent with annotators spreading work across more sessions and the consequent depression of κ documented in RQ5.

E Experiment Setup

We evaluate the dataset on three-class sentiment classification (Positive, Negative, Neutral) using the majority-vote labels from Section 3.4, with an 80/10/10 train/dev/test split stratified by label.

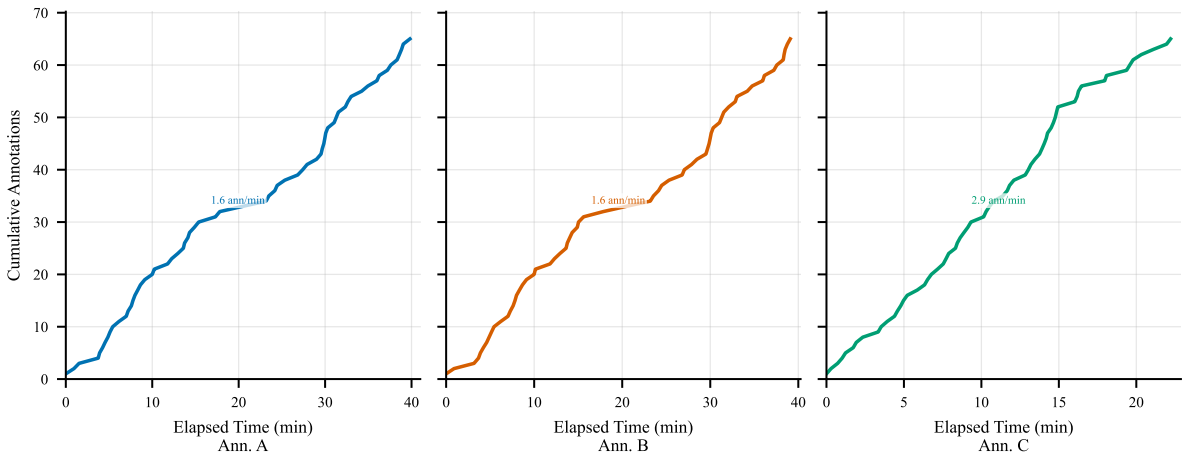
Three multilingual pre-trained language models are evaluated: mBERT (Devlin et al., 2019) as a general multilingual baseline; AfriBERTa (Ogueji et al., 2021), and AfroXLMR-base (Alabi et al., 2022), both adapted to African languages. Each model is first zero-shot evaluated via masked language modeling (MLM): the sentiment word in a fixed template is replaced with `[MASK]` and the highest-probability candidate determines the predicted label.

Fine-tuning follows the procedure of Shode et al. (2022) adapted to our three-class setup. We use AdamW with a learning rate of 5×10^{-5}, a linear schedule with 6% warmup steps, batch size 32, and maximum sequence length 128. Models are trained for up to 20 epochs; training is stopped early if dev macro-F1 does not improve for 3 consecutive epochs, and the best checkpoint by dev macro-F1 is used for final test evaluation. Gradients are clipped at a norm of 1.0. Each model is trained with five random seeds; we report macro-F1 and accuracy averaged over seeds. We additionally evaluate GPT-5 (Singh et al., 2025) and Gemini 3 Flash Preview (Google, 2025) in zero-shot and six-shot (2 examples per class) settings.

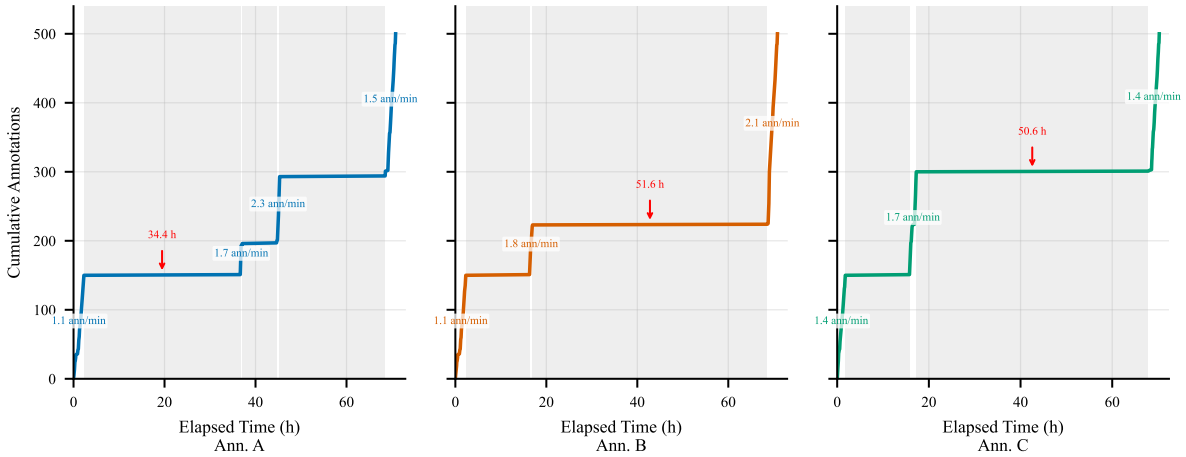
F Recommendations

To improve annotation quality in future campaigns, we recommend the following interventions based on our findings:

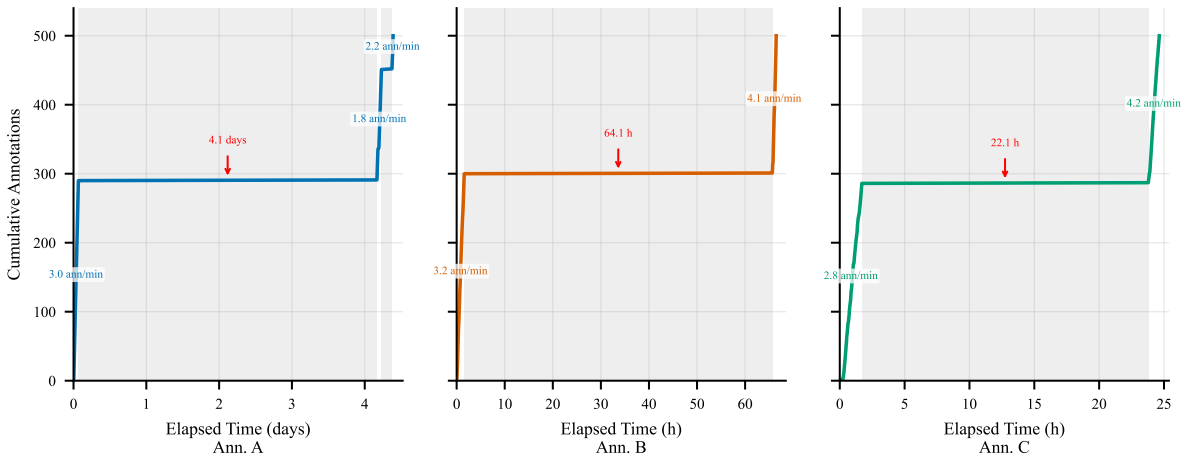
1. **Maximize temporal simultaneity.** Requiring all annotators to complete each batch within a short fixed window (e.g., 48 hours) would reduce mean inter-annotator time spans and is the single highest-leverage intervention our results support. Where feasible, shared online annotation sessions, where all annotators label the same tweets in real time, could bring κ closer to the ≥ 0.98 we observe for the $<$ 1 min group.
2. **Monitor run lengths between batches.** Mean run length is detectable before a batch is finalized and requires only the annotation log. A simple threshold alert (e.g., mean run length >1.8) could trigger a calibration discussion or rest break before quality degrades further.
3. **Include periodic calibration items.** Re-annotating a small fixed set from the training batch at the start of each new proper annotation batch provides an early-warning signal of label drift and allows re-calibration before it compounds.
4. **Report per-batch κ as standard practice.** Aggregate κ over a multi-batch campaign can be misleadingly high even when the final batches are severely degraded. We recommend per-batch κ as a mandatory element of annotation quality reporting in African-language NLP dataset papers.



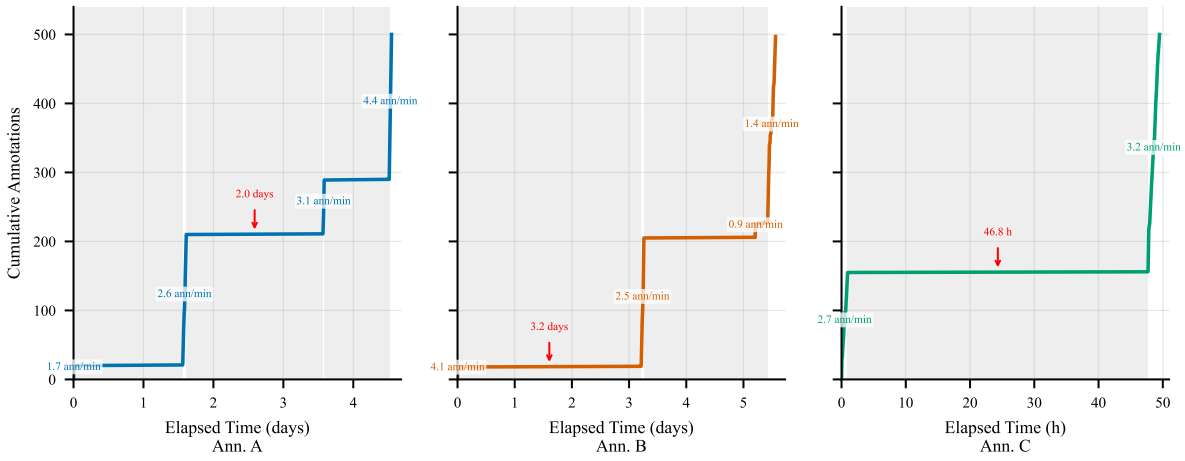
(a) Training batch



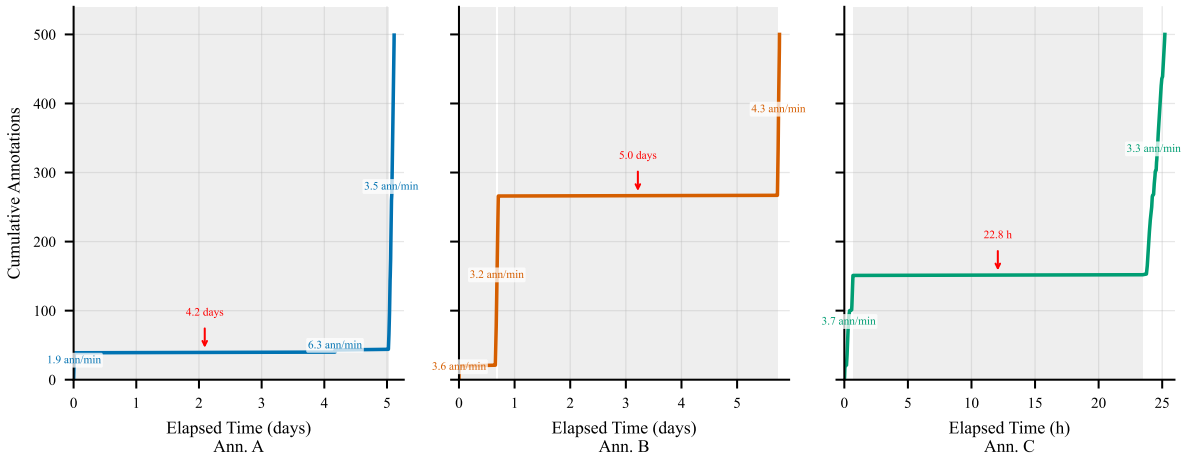
(b) Batch 1



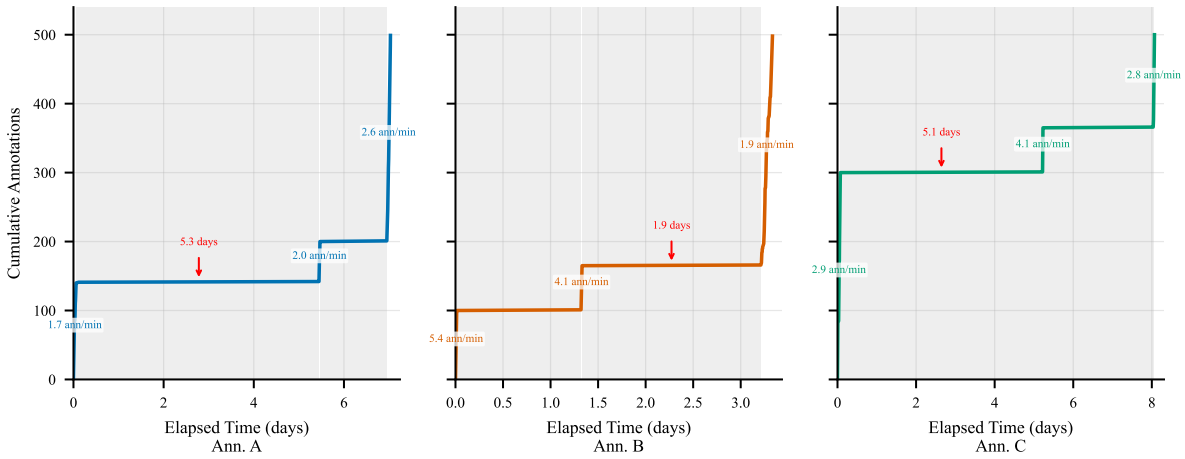
(c) Batch 2



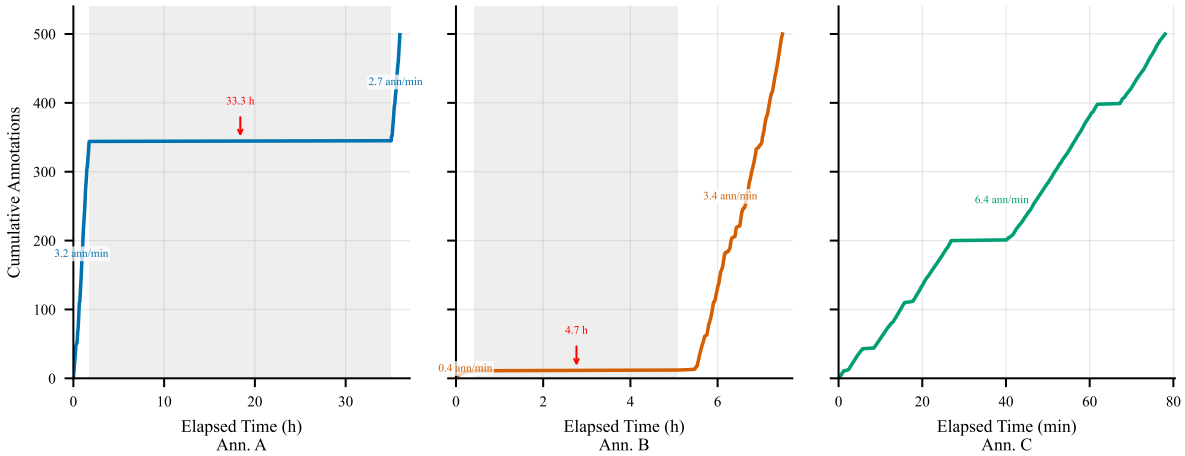
(d) Batch 3



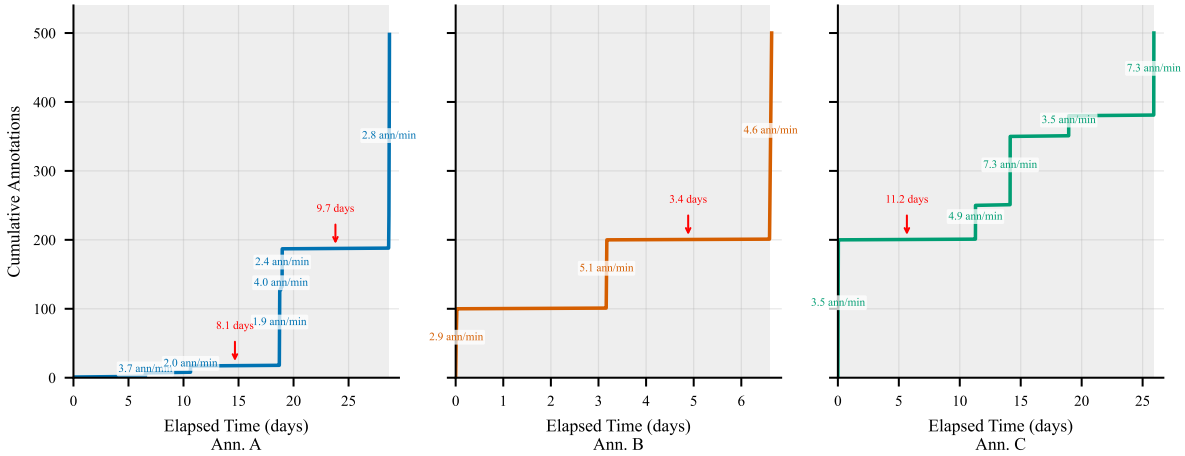
(e) Batch 4



(f) Batch 5



(g) Batch 6



(h) Batch 7

Figure 12: Cumulative annotation progress per annotator per batch. X-axis: elapsed time (min / h / days); y-axis: cumulative tweets annotated. Grey shading = rest period ($\geq$ 1 h idle); red arrows = extreme idle gaps (duration at arrow base); coloured labels = per-session speed (ann/min).